KELPA

Kansas English Language Proficiency Assessment

KELPA Technical Manual

March 2021

Table of Contents

Table of Tables

Table of Figures

Table of Appendices

I. Statewide System of Standards and Assessments

The Kansas English Language Proficiency Assessment (KELPA) is the summative assessment for K–12 English learners (ELs) in Kansas, administered each spring. As part of the federal elementary and secondary education legislation for ELs, the test was developed according to the 2018 Kansas Standards for English Learners: Grades K–12 (hereafter referred to as the 2018 Standards). Assessed grade levels and bands include kindergarten, 1, 2–3, 4–5, 6–8, and 9–12. The target student population for KELPA are students who are identified as ELs from grades K–12.

This chapter provides an overview of KELPA. It describes the program's background, purpose, uses of the assessment scores, and the intended population.

I.1 Overview of English Language Standards

The Kansas State Department of Education (KSDE) mission statement for English language proficiency states:

> Kansas instruction of English for speakers of other languages prepares English Learners for success in school and in society through development of English proficiency, with specific emphasis on literacy skills needed to access academic content. (See Kansas English learners' website.)

An important component of the development of the 2018 Standards is to create an assessment to monitor the growing proficiency of ELs relative to those standards. Since 2013, various versions of the KELPA assessments (i.e., the previous version of KELPA from 2013 to 2015 and KELPA2 from 2016 to 2019) have served the purpose of monitoring ELs' English proficiency, based on previous standards in the domains of reading, writing, speaking, and listening. The 2020 administration of KELPA is the latest iteration of English language proficiency assessment, aligned with the newest 2018 Standards, to monitor English language proficiency for ELs.

I.1.1 Standards Committee

The 2018 Standards were developed by the standards committee. The standards committee consisted of four members and two co-chairs, representing over 100 years of combined experience in K–12 teaching and education, with a variety of expertise in elementary and secondary levels. The committee members specialized in different areas, but all were experienced in EL services and had worked with ELs in academic content areas. Several were experienced with developing standards for academic content areas and served on committees for English language arts (ELA) and social studies standards. Some had served as KSDE assessment committee members. The co-chairs had experience as classroom teachers, and each had served in a variety of other roles, including state social studies specialist, assessment development assistant, instructional coach, assistant principal, and district administrator.

After the standards were developed, they were available for public review and input. Public hearings were held in Wichita and Topeka during summer of 2018. Recommendations from public hearings were used to update the Kansas Standards for English Learners document. KSDE presented the EL standards document to the board of education in August 2018 for official approval. The presentation included an introduction to the structure of the EL standards document as well as updates made to the standards document to reflect public comments. Appendix A contains the KSDE presentation to the Kansas Board of Education. On September 11, 2018, the Kansas State Board of Education unanimously adopted the Kansas Standards for English Learners (p. 2, Board minutes).

I.1.2 Overview of the Standards

The 2018 Standards, developed for grades K–8 and grade bands 9–10 and 11–12, illuminate the critical language, knowledge about language, and language skills that ELs need to be academically successful. The four domains of ELA—reading, writing, speaking, and listening—are the foundation for the 2018 Standards. The 2018 Standards are progressions of the specific grade-level ELA standards within the four domains of listening, speaking, reading, and writing. The English language acquisition and development addressed by the 2018 Standards were drawn directly from the 2017 Kansas Standards for English Language Arts. The 2018 Standards are used to support individual students in gaining a level of proficiency in both social English and academic English that allows them to succeed in reaching the grade-level academic standards as quickly as possible. They also informed the design and content of the newly developed KELPA.

To support instruction, the 2018 Standards include a rubric of performance descriptions for ELs (p. 10) that profiles the general stages of language acquisition, continua related to using language to communicate in social context, use language to construct meaning (i.e., reading and listening), use language to convey ideas (i.e., writing and speaking), and use language to engage in grade-level content. The rubrics of performance include six levels:

- Level 0: Beginning (starting point)
- Level 1: Emerging
- Level 2: Developing
- Level 3: Approaching
- Level 4: Proficient
- Level 5: Mastery

As stated in the 2018 Standards "the state's determination of 'proficiency' central to decisions regarding EL participation in Title III programming and measured through the Kansas English Language Proficiency Assessment is at 'proficient/level 4'" (p. 10). Performance at level 4 is considered "proficient" when students demonstrate a level of English proficiency in which they produce language that is comparable to that of non-EL peers and do not require additional support to access grade-level curriculum and academic content.

There are 21 standards across the four domains and grades in the 2018 Standards. Certain standards in reading and writing are not included for younger grades because of developmental appropriateness. For example, Reading Standard (R8), "Follow the logic of an argument based on the validity of the claim and evidence presented," is not part of the kindergarten standards. Moreover, the same standard codes are used across grades, but the text of the standards differs slightly to reflect the varying expectations of performance across grades. Among the reading standards, there are three standards coded as reading foundations that span kindergarten through 12 and reading–literature/reading–informational, which covers that rest of the reading standards. Table I-1 shows the number of standards by domain and grade.

Table I-1: Number of 2018 Standards by Domain and Grade

Grade	Listening / Speaking	Reading	Writing
K	8	8	2
1	8	9	2
2	8	10	2
3–12	8	10	4

Each standard describes "various touch points of proficiency along a continuum of performance so language acquisition can be understood" (2018 Standards, p. 8). To support instruction, the performance rubric (2018 Standards, p. 12) outlines five levels, from beginning to mastery, of what students know and can do relative to each standard.

To support acquisition of academic contents for ELs, the 2018 Standards also provide examples of domain-specific vocabularies in mathematics, science, social studies, and ELA. These sample vocabularies represent content demands required in Kansas K–12 schools for not only ELs but all students in Kansas.

As indicated in the 2018 Standards, it is common for students' social language to be more developed than their receptive language and for their receptive language to be more developed than their expressive language. The timeline for students to achieve proficiency is highly individualized and depends on several factors, including (2018 Standards, p. 9):

- school's program type
- age at which a student entered the program
- initial proficiency level
- native language literacy
- linguistic and cultural background
- life and educational experiences
- additional needs (e.g., health, disability)

I.2. Test Purposes and Uses

KELPA, aligned to the 2018 Standards, is a yearly summative assessment for students in grades K–12 who are identified as not proficient in English and who receive EL services as required by Title I of the Elementary and Secondary Education Act (ESEA).[1] As part of the ESEA Title I accountability requirement, KELPA results are used to determine the level of English language proficiency of ELs and to assess their progress in acquiring the skills of listening, speaking, reading, and writing in English.

KELPA measures the English language proficiency of ELs to determine who may benefit from receiving EL services and support that ensure students can have the language skills to meaningfully participate in educational programs and services. KELPA scores classify ELs' English proficiency into four performance levels (i.e., Level 1—Beginning, Level 2—Early Intermediate, Level 3—Intermediate, Level 4—Early Advanced) in each of the four domains and provide an indicator of progress toward overall proficiency

[1] Title I of the Elementary and Secondary Education Act of 1965 (20 U.S.C. 6301 et seq.): Improving the Academic Achievement of the Disadvantaged

(i.e., Level 1—Not Proficient, Level 2—Nearly Proficient, Level 3—Proficient). The proficiency levels determine whether ELs have reached the level of English proficiency that allows them to participate in a standard instructional program in the classroom without additional language support. ELs who demonstrate the English language skills required for engagement with grade-level, academic content instruction at a level comparable to non-ELs in all four domains (i.e., listening, speaking, reading, writing) are considered proficient in English language and may exit the EL program services.

Beyond understanding common English usage, ELs need to understand the language used for grade-level instruction in ELA, mathematics, science, social studies, and other content areas. The standards highlight and amplify the critical language, knowledge about language, and skills for using language that are necessary for ELs to be successful in school.

I.3. Intended Population

KSDE is committed to including all eligible ELs in KELPA. Students are identified as ELs when their home or native language is not English and their limitations in the English language may affect their ability to participate in their school's education program. As described, all students in grades K–12 who are identified as ELs must take KELPA, whether or not they receive English language services. For example, parents may waive their student out of EL services, but if the student is identified as an EL, he or she is still required to take KELPA. Detailed information about participation in English for speakers of other languages (ESOL) services and the KELPA program can be found in ESOL Program Guidance.

When applicable, a student's Individualized Education Program (IEP) is used to guide accommodations use for KELPA. Accommodations are set before testing using the Personal Needs Profile (PNP) in the online testing platform and are consistent with other content tests. ELs with significant cognitive disabilities also take KELPA. The PNP submitted by teachers determines the availability of test accommodations for individual students. A detailed summary of accommodations is in Section V. Inclusion of All Students in this technical manual.

A few exemptions for assessment include:

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

II. Assessment System Operations

This chapter provides details about KELPA, including test design and development, item development, test administration, monitoring test administration, and test security.

II.1 Test Design and Development

KELPA assessments, a part of the Kansas Assessment Program (KAP), are entirely computer based for students in grades 2 through 12. Students in grades K–1 take a mostly computer-based exam but also complete a small number of writing items with paper and pencil. KELPA was designed to be a fixed-form test with one operational form for each domain (i.e., listening, speaking, reading, and writing) and grade level or grade band. All reading and listening items are machine scored, all speaking items are educator scored, and the writing section is composed of both machine- and educator-scored items. The assessments are delivered, in any order of the four domains, through the online test-delivery platform, Kite®. The Kite system also delivers other computer-based assessments within the KAP.

The University of Kansas's Achievement & Assessment Institute (AAI) worked with the Kansas State Department of Education (KSDE) to determine the content to be assessed by the KELPA for each domain and grade or grade band. The development leading to the 2020 KELPA administration occurred over multiple years. Table II-1 outlines the content-development timeline for the KELPA.

Table II-1: Development Timeline for the KELPA

Milestone	Date
Adoption of the 2018 Standards	September 2018
Discussion of standards emphasis (for blueprints) with EL educators and KSDE	October 2018
KELPA passage and item development	2018 to 2020
KELPA item content external review	Summer 2019
KELPA item bias & sensitivity external review	Summer 2019
Operational field testing	February–March 2020
Standard setting	Fall 2020
State board approval of proficiency standards	Winter 2021

II.1.1 Test Blueprints

In October 2018, a panel of 14 Kansas educators of English learners (ELs) met with staff from KSDE and AAI to discuss the new standards and their impact on assessment. Among these 14 Kansas educators, four had experience in teaching K–12, one had no teaching experience, and others had teaching experience with at least three grade levels. Those discussions informed KELPA's draft test blueprints. An important feature of the KELPA test blueprint is the use of clusters. The notion of clusters is not explicitly included within the 2018 Standards; however, clusters of standards (strands) are included in the 2017 Kansas English Language Arts Standards from which the 2018 Standards were adapted. Clusters are small groups of similar standards that are helpful in organizing individual English-language-proficiency standards. Table II-2 shows the relationships between clusters and standards. Not all content clusters are tested in each grade or grade band. For example, "production of writing" is not tested in kindergarten or grade-1 writing.

Table II-2: Relationship Between Clusters and Standards

Domain	Grade or grade band	Cluster	Standard
Listening	K, 1, 2–3, 4–5, 6–8, 9–12	Comprehension & collaboration	SL.1, SL.2, SL.3
	K, 1, 4–5, 6–8, 9–12	Language in speaking & listening	SL.7, SL.8
	1, 2–3, 4–5, 6–8, 9–12	Presentation of knowledge & ideas	SL.4, SL.6
Reading	K, 2–3, 6–8, 9–12	Craft & structure	R.4, R.10
	1, 2–3, 4–5, 6–8, 9–12	Discourse comprehension	R.1, R.8, R.13
	K, 1, 2–3, 4–5, 6–8, 9–12	Language in reading	R.11, R.12
	K, 1, 2–3, 4–5, 6–8, 9–12	Reading foundations	RF.2, RF.3, RF.4
Speaking	K, 1, 2–3, 4–5, 6–8, 9–12	Comprehension & collaboration	SL.1, SL.2, SL.3
	K, 1, 2–3, 4–5, 6–8, 9–12	Presentation of knowledge & ideas	SL.4, SL.6
Writing	K, 1, 2–3, 4–5, 6–8, 9–12	Language in writing	W.10, W.11
	2–3, 4–5, 6–8, 9–12	Production of writing	W.4

KELPA test blueprints were developed and used to assemble the 2020 KELPA operational field test. The test blueprints specify ranges of score points required for each cluster by grade or grade band for each of the four domains. Table II-3, Table II-4, Table II-5, and Table II-6 present the score-point ranges by cluster for KELPA test blueprints. Proportions of score points by clusters differ across grades or grade bands to reflect the varying emphasis of clusters across grade or grade bands.

Table II-3: KELPA Test Blueprint for Listening by Cluster and Grade or Grade Band

Grade or grade band	Cluster	Description of cluster	Range of score points
K, 1, 2–3	Comprehension & collaboration	• Engage in civil discourse and express original ideas professionally, clearly, and persuasively in a variety of settings and with diverse partners who both agree and disagree with their point of view. • Synthesize information presented in diverse media and formats, assessing its relevance and accuracy according to purpose and audience. • Objectively assess the relevance, accuracy, and validity of a speaker's claim and supporting evidence.	15–21
K, 1, 2–3	Presentation of knowledge & ideas	• Prepare a variety of presentations, each with a clear line of reasoning, meaningful organization, appropriate style, including information and findings. • Effectively adapt speech to fit a variety of contexts and communication situations.	0–5
K,1, 2–3	Language in speaking & listening	• Accurately and effectively use Standard English grammar and usage when speaking. • Use a variety of context-appropriate words in a range of situations and engage in effective strategies to determining word meanings and adding new words to a personal vocabulary bank.	0–8
4–5, 6–8, 9–12	Comprehension & collaboration	• Engage in civil discourse and express original ideas professionally, clearly, and persuasively in a variety of settings and with diverse partners who both agree and disagree with their point of view. • Synthesize information presented in diverse media and formats, assessing its relevance and accuracy according to purpose and audience. • Objectively assess the relevance, accuracy, and validity of a speaker's claim and supporting evidence.	17–21
4–5, 6–8, 9–12	Presentation of knowledge & ideas	• Prepare a variety of presentations, each with a clear line of reasoning, meaningful organization, appropriate style, including information and findings. • Effectively adapt speech to fit a variety of contexts and communication situations.	2–6

Grade or grade band	Cluster	Description of cluster	Range of score points
4–5, 6–8, 9–12	Language in speaking & listening	• Accurately and effectively use standard English grammar and usage when speaking. • Use a variety of context-appropriate words in a range of situations and engage in effective strategies to determining word meanings and adding new words to a personal vocabulary bank.	2–6

Table II-4: KELPA Test Blueprint for Speaking by Cluster and Grade or Grade Band

Grade or grade band	Cluster	Description of cluster	Range of score points
All grades and grade bands	Comprehension & collaboration	• Engage in civil discourse and express original ideas professionally, clearly, and persuasively in a variety of settings and with diverse partners who both agree and disagree with their point of view. • Synthesize information presented in diverse media and formats, assessing its relevance and accuracy according to purpose and audience. • Objectively assess the relevance, accuracy, and validity of a speaker's claim and supporting evidence.	9–18
All grades and grade bands	Presentation of knowledge & ideas	• Prepare a variety of presentations, each with a clear line of reasoning, meaningful organization, appropriate style, including information and findings. • Effectively adapt speech to fit a variety of contexts and communication situations.	9–18
All grades and grade bands	Language in speaking & listening	• Accurately and effectively use standard English grammar and usage when speaking. • Use a variety of context-appropriate words in a range of situations and engage in effective strategies to determining word meanings and adding new words to a personal vocabulary bank.	Tested through inclusion in constructed -response scoring rubrics

Table II-5: KELPA Test Blueprint for Reading by Cluster and Grade or Grade Band

Grade or grade band	Cluster	Description of cluster	Range of score points
K	Reading foundations	• Demonstrate understanding of spoken words, syllables, and phonemes. • Know and apply grade-level phonics and word analysis skills in decoding words. • Read with sufficient accuracy and fluency to support comprehension.	9–13
K	Language in reading	• Understand vocabulary and word use in a variety of contexts by consistently building knowledge of new words, as well as employing strategies for determining meanings of unfamiliar words. • Understand word meanings and nuances in word meanings when reading.	4–7
K	Discourse comprehension	• Read closely through multiple interactions with a text in order to determine what the text says explicitly and to make logical inferences; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. • Interpret meaning from a variety of texts on their own.	0–1
K	Craft & structure	• Recognize the ways in which the author's word choice and use of figurative language deliberately influences meaning, tone, or mood within the context of the text.	0–1
1	Reading foundations	• Demonstrate understanding of spoken words, syllables, and phonemes. • Know and apply grade-level phonics and word analysis skills in decoding words. • Read with sufficient accuracy and fluency to support comprehension.	12–17
1	Language in reading	• Understand vocabulary and word use in a variety of contexts by consistently building knowledge of new words, as well as employing strategies for determining meanings of unfamiliar words. • Understand word meanings and nuances in word meanings when reading.	5–8

Grade or grade band	Cluster	Description of cluster	Range of score points
1	Discourse comprehension	• Read closely through multiple interactions with a text in order to determine what the text says explicitly and to make logical inferences; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. • Follow the logic of an argument based on the validity of the claim and evidence presented. • Interpret meaning from a variety of texts on their own.	3–7
1	Craft & structure	• Recognize the ways in which the author's word choice and use of figurative language deliberately influences meaning, tone, or mood within the context of the text.	0–2
2–3	Reading foundations	• Demonstrate understanding of spoken words, syllables, and phonemes. • Know and apply grade-level phonics and word analysis skills in decoding words. • Read with sufficient accuracy and fluency to support comprehension.	6–11
2–3	Language in reading	• Understand vocabulary and word use in a variety of contexts by consistently building knowledge of new words, as well as employing strategies for determining meanings of unfamiliar words. • Understand word meanings and nuances in word meanings when reading.	5–8
2–3	Discourse comprehension	• Read closely through multiple interactions with a text in order to determine what the text says explicitly and to make logical inferences; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. • Follow the logic of an argument based on the validity of the claim and evidence presented. • Interpret meaning from a variety of texts on their own.	6–10
2–3	Craft & structure	• Recognize the ways in which the author's word choice and use of figurative language deliberately influences meaning, tone, or mood within the context of the text.	0–3

Grade or grade band	Cluster	Description of cluster	Range of score points
		• Apply their knowledge of language and how it works to a variety of contexts and situations.	
4–5	Reading foundations	• Demonstrate understanding of spoken words, syllables, and phonemes. • Know and apply grade-level phonics and word analysis skills in decoding words. • Read with sufficient accuracy and fluency to support comprehension.	4–6
4–5	Language in reading	• Understand vocabulary and word use in a variety of contexts by consistently building knowledge of new words, as well as employing strategies for determining meanings of unfamiliar words. • Understand word meanings and nuances in word meanings when reading.	6–9
4–5	Discourse comprehension	• Read closely through multiple interactions with a text in order to determine what the text says explicitly and to make logical inferences; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. • Follow the logic of an argument based on the validity of the claim and evidence presented. • Interpret meaning from a variety of texts on their own.	7–10
4–5	Craft & structure	• Recognize the ways in which the author's word choice and use of figurative language deliberately influences meaning, tone, or mood within the context of the text. • Apply their knowledge of language and how it works to a variety of contexts and situations.	0–3
6–8, 9–12	Reading foundations	• Demonstrate understanding of spoken words, syllables, and phonemes. • Know and apply grade-level phonics and word analysis skills in decoding words. • Read with sufficient accuracy and fluency to support comprehension.	2–3
6–8, 9–12	Language in Reading	• Understand vocabulary and word use in a variety of contexts by consistently building knowledge of new words, as well as employing strategies for determining meanings of unfamiliar words.	6–10

Grade or grade band	Cluster	Description of cluster	Range of score points
		• Understand word meanings and nuances in word meanings when reading.	
6–8, 9–12	Discourse comprehension	• Read closely through multiple interactions with a text in order to determine what the text says explicitly and to make logical inferences; cite specific textual evidence when writing or speaking to support conclusions drawn from the text. • Follow the logic of an argument based on the validity of the claim and evidence presented. • Interpret meaning from a variety of texts on their own.	10–13
6–8, 9–12	Craft & structure	• Recognize the ways in which the author's word choice and use of figurative language deliberately influences meaning, tone, or mood within the context of the text. • Apply their knowledge of language and how it works to a variety of contexts and situations.	0–3

Table II-6: KELPA Test Blueprint for Writing by Cluster and Grade or Grade Band

Grade or grade band	Cluster	Description of cluster	Range of score points
K, 1	Language in writing	• Accurately and effectively use standard English grammar and usage when writing. • Accurately and effectively use the mechanics of standard English for the purpose of productive communication.	12–21
2–3	Language in writing	• Accurately and effectively use standard English grammar and usage when writing. • Accurately and effectively use the mechanics of standard English for the purpose of productive communication.	7–11
2–3	Production of writing	• Create texts appropriate for specific purposes, audiences, and tasks.	4–7
4–5	Language in writing	• Accurately and effectively use standard English grammar and usage when writing. • Accurately and effectively use the mechanics of standard English for the purpose of productive communication.	8–13
4–5	Production of writing	• Create texts appropriate for specific purposes, audiences, and tasks.	6–9
6–8, 9–12	Language in writing	• Accurately and effectively use standard English grammar and usage when writing. • Accurately and effectively use the mechanics of standard English for the purpose of productive communication.	8–15
6–8, 9–12	Production of writing	• Create texts appropriate for specific purposes, audiences, and tasks.	6–9

II.1.2 Test Design

KELPA is administered for six grades or grade bands: kindergarten, grade 1, grades 2–3, grades 4–5, grades 6–8, and grades 9–12. It includes domain-specific tests in listening, speaking, reading and writing and can be administered in any order. All domain tests are untimed. Table II-7 shows item counts for the 2020 KELPA administration for all grades by domain.

Table II-7: Item Counts for the KELPA Operational Field Test by Domain and Grade

Domain	Grade	No. of items
Listening	K, 1–12	25
Reading	K	20
	1–12	25
Speaking	K, 1–12	10
Writing	K, 1	15
	2–5	19
	6–12	20

All reading and listening items are designed to be computer scored, and all speaking items are constructed-response (CR) items with student responses recorded in Kite and scored by local educators. Writing includes both CR items scored by local educators and computer-scored items. All CR items (in speaking and writing) are scored on a 0–3 scale using a provided scoring rubric (see Section IV.3.1.2 Educator Scoring for descriptions of educator scoring). Section II.3 Test Administration and Scoring provides more information about CR item scoring. Students in kindergarten and grade 1 are asked to first complete the computer-administered writing items and then to respond to the last few items in a paper test booklet. Test administrators read a script pertaining to the items for students to respond to in their paper booklets. Responses from kindergarten and grade-1 writing are then scanned into PDF files.

Some items, including items in the listening domain, require the student to listen to audio clips in Kite. Optional audio clips are made available for students in all domains for directions (e.g., background information, instructions to look at an image, etc.), stems, and prompts to support students' understanding of these three components, which helps isolate student proficiency more precisely in the assessed domain. Audio for item components can be listened to repeatedly, as needed. Domain tests are described in the following subsections.

II.1.2.1 Reading

The reading test focuses on the literacy skills necessary for academic success. Stimuli represent a range of genres, including informational texts and literary texts. Specifications for stimulus length and linguistic complexity for the 2020 test were determined in collaboration with KSDE (see Section II.2.1 Passage Development). In kindergarten and grade 1, the reading test includes only discrete items. In all other grade bands, the test includes both discrete and set-based items. According to grade level or band and the standard tested, students may be asked to:

- select a missing letter for a word depicted by an image, which is related to reading foundation standards
- match a single word with the image it represents, which is related to reading foundation standards
- read a story and order pictures representing each activity, which is related to craft and structure, general reading, and language in reading standards, depending on the type of question asked and specific grade or grade band

- read a poster or flyer with graphic support and answer questions about it, which is related to craft and structure, general reading, and language in reading standards, depending on the type of question asked and specific grade or grade band
- read an informational or literary text and answer questions about the content of the text and/or the author's craft, which is related to craft and structure, general reading, and language in reading standards, depending on the type of question asked and specific grade or grade band

II.1.2.2 Listening

Although the 2018 Standards combine speaking and listening standards, the constructs are tested separately, and separate scores are provided for each domain. The listening test, including monologic and dialogic stimuli, employs both discrete items and item clusters built around a single stimulus, tapping those standards focusing on aural receptive skills. Grade-level reading literacy is neither required nor assumed for this section of the test. Responses generally require students in kindergarten and grade 1 to select images. Text-based response options used in grade 2 and above are written to a below-grade reading level to mitigate introducing potential construct-irrelevant variance. In addition, students may play all audio repeatedly. Images are frequently used to set the scene for the stimuli. Students may be asked to:

- follow oral directions to arrange images
- listen to a narrative and order images representing key activities
- select an option that answers a question or represents the stimuli heard
- listen to an academic discussion or presentation and answer questions about the content

II.1.2.3 Writing

The writing test presents both selected-response (SR) and CR items. The SR items often have a one- or two-sentence prompt or stem, followed by a one- to three-sentence stimulus to evaluate. The SR items focus on applying knowledge of language use including grammar, vocabulary, and mechanics while CR items require the production of written text. SR tasks include but are not limited to:

- choosing correct grammatical options or punctuation
- selecting words to be capitalized in a text
- arranging words to form a sentence

CR items in kindergarten and grade 1 require written responses that are letters, words, or short sentences. In grade bands 2–3 and 4–5, responses to one or two sentences of background information or a prompt are required; these responses can range from a single word to a paragraph. CR items in grade bands 6–8 and 9–12 generally require students to draft multisentence to paragraph-length responses to prompts. These prompts usually comprise three sentences, including background information or lead-in and prompt.

II.1.2.4 Speaking

All items in the speaking test require verbal responses. Student responses typically range from making a short statement or finishing a story to retelling a narrative or making a presentation. Similar to the listening test, grade-level literacy is not required or assumed for the speaking test. Students can utilize supporting graphics and can play audio files for prompts and lead-in or background information as many times as desired.

II.1.3 Test Construction

When building the 2020 operational field-test forms, KELPA content-development staff constructed the test forms according to the blueprints for each domain test and according to the following guidelines:

- Items from a wide range of estimated item difficulties were chosen. Although the items included on the test forms did not yet have statistical properties, AAI content-development staff with background in English language assessment development and/or English language education used their own expert judgment and the corresponding performance level descriptors included in the 2018 Standards to evaluate each item's level of difficulty. KELPA content-development staff aimed to have the highest number of items near the midpoint of the performance levels to increase the measurement power around Levels 3 and 4, where the most important distinction for English proficiency lies.
- Discrete items in listening and reading domains were roughly grouped together by item type and appeared in test forms before passage-based items.
- Writing CR items were grouped together and appeared after machine-scored items. Machine-scored writing items were grouped into different levels of difficulty perceived by content-development staff; that is, easier items appeared in the test form before more-difficult items.
- Similar to writing items, speaking items were also grouped into different levels of difficulty.
- Reading and listening passage-based items were ordered according to established protocol (i.e., starting with the main idea and referencing where the concept appears in the text; for example, if there were two vocabulary questions, the item for the one appearing earlier in the text comes first).

II.2 Content Development

Content development entails various efforts to ensure item quality, including ongoing research into best practices for assessing EL proficiency, recruiting highly qualified item writers, developing and providing comprehensive and clear item-writer training materials, conducting item-writer training, and reviewing and revising items.

Item review is conducted in two phases: after items are created and again after items are field-tested. Before utilizing item on any assessment, AAI staff do content review and editing, with an eye for bias and sensitivity issues prior to the external reviews (i.e., content, bias and sensitivity) with Kansas EL educators and KSDE staff. AAI staff members use item-review feedback to revise test items as needed. Items are then prepared for operational field testing, according to the blueprint and following established guidelines for general presentation. After operational field testing, item and test data are analyzed; this data analysis guides decisions about the use of items on operational assessments. The next section describes typical procedures for different stages of item development.

II.2.1 Passage Development

For KELPA assessments, some reading and listening items have stimuli that are passages. This section describes the processes of passage writing, passage selection, and passage review for KELPA reading and listening items.

II.2.1.1 Passage Writing

With the exception of poetry, which was selected from the public domain, KELPA passages were either written by AAI staff or commissioned. Authors external to AAI included professional assessment freelancers and writers of children's literature.

All passage writers were provided information regarding the readability indices, topics of interest, universal design (UD) principles, and bias-and-sensitivity guidelines. AAI internal passage writers, familiar with passage-creation best practices, were given KELPA-specific guidance on topics of interest, readability indices, and standards to review and edit passages produced by non-AAI writers.

KELPA passages include literary passages, poems, and informational texts. For reading passages, writers for KELPA referred to both Lexile and Flesch–Kincaid grade expectations to guide the complexity of the passage being developed. Passages were assigned to specific grades or grade bands. Final grade or grade-band decisions were made in consultation with KSDE consultants. Table II-8 presents the length progression of reading passages by type and grade level.

Table II-8: Progression of Length of Reading Passages by Type and Grade or Grade Band

Grade or grade band	Word	Sentence	Multisentence	Paragraph	Multiparagraph	Max. word count
K	X	X				7
1	X	X				10
2–3		X	X			50
4–5			X	X	X	200
6–8			X	X	X	200
9–12				X	X	500

Listening passages were reviewed for appropriateness of content, vocabulary, length, and general ease of comprehension. Table II-9 presents the length progression of listening passages by type and grade level.

Table II-9: Progression of Length of Listening Passages by Type and Grade or Grade Band

Grade or grade band	Sentence	Multisentence	Multiparagraph	Max. length (in seconds)
K	X	X		30
1	X	X		30
2–3	X	X		45
4–5		X	X	60
6–8		X	X	60
9–12		X	X	90

II.2.1.2 Passage Selection and Revision

In addition to writing new passages for KELPA, existing passages were also identified for potential use. The AAI content-development staff selected passages that met both the Lexile and the Flesch–Kincaid grade expectations for complexity. KELPA content-development staff also evaluated passage vocabulary using word lists from EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies (Taylor et al., 1989) and Children's Writers Word Book (Magilner & Magilner, 2006). Use of the criteria for passage complexity and for vocabulary ensured that texts were appropriate for measurement of EL proficiency. AAI passage writers also revised selected existing passages as needed to make them more suitable for KELPA.

II.2.1.3 Passage Review

All passages for KELPA went through multiple rounds of AAI internal review including editorial, content, and bias-and-sensitivity reviews. Internal reviewers included content-development staff with backgrounds in English-language-proficiency assessment development and/or English language education, as well as editors with expertise in reviewing items and passages for multiple assessment programs. The passages were also reviewed by KSDE consultants.

Editors at AAI first conducted an editorial review. All editors received specific, in-house training in reviewing, editing, and revising assessment materials. This training included an awareness of culture, bias, and sensitivity issues. AAI editors have 26 cumulative years of editing experience, of which 23 years are specific to reviewing and editing assessment passages. Editors hold undergraduate and graduate degrees in English, education, and foreign languages.

AAI content experts with backgrounds in ELA and English for speakers of other languages reviewed passages from both content and bias-and-sensitivity perspectives. One internal content reviewer had about four years of item writing experience at the time, had participated in ELA internal passage reviews, and helped with post-external reviewer reconciliation. Another internal reviewer was a district assessment coordinator for a Kansas district for nine years and trained teachers on KELPA administration and scoring. This reviewer also served on the Kansas Assessment Advisory Committee. KSDE consultants performed the final review.

All passages underwent a review and revision process before selection for item development. They were reviewed for:

- editorial elements
    - plagiarism, testable content, formatting, clarity, adherence to style guide, correct mechanics, readability
    - potential fact, bias-and-sensitivity, or accessibility concerns
- content
    - grade appropriateness
    - fact-based information that is supported by reputable sources
- bias and sensitivity
    - suitability of text- and graphic-based passages for test takers, regardless of gender, ethnicity, or cultural origin

Passages were revised or rejected depending on the issues. Passages that needed significant changes were rejected rather than revised. After several rounds of internal review, KSDE consultants reviewed passages for content, level of difficulty, accessibility (as related to language clarity), and bias and sensitivity. AAI content-development staff addressed KSDE reviewers' comments and revised passages accordingly.

II.2.2 Item Writing

KELPA item writing started with an in-person item-writing event in January 2019. Items for the operational field test were developed in this January event. Item writers were recruited from across districts in Kansas and received training in item writing before and during the event.

II.2.2.1 Item Writers

Twenty-two Kansas EL educators and two KSDE staff members participated in the 2019 KELPA item-writing workshop. Geographically, 12 item writers were from northeastern Kansas, six from central Kansas, and four from southwestern Kansas. Although the Kansas English language proficiency standards were new at the time of the item-writing workshop, all participants had EL teaching experience and most educators had extensive experience with ELs across multiple grade levels. For example, one educator had been working with ELs across all 13 grade levels (i.e., K–12) when the item-writing workshop occurred, and two educators had previous experience working with ELs across all 13 grade levels. Three educators were not teaching when the workshop occurred but had extensive experience working with either elementary-level ELs or ELs in both elementary and secondary schools. One educator only had experience working with ELs in kindergarten. Table II-10 provides the number of educators with current and/or previous experience in each grade or grade band.

Table II-10: Item Writers' Experience Teaching English Learners (ELs) by Grade or Grade Band (N = 22)

Grade or grade band	Currently teaching or previously taught ELs (n)
K	19
1	19
2–3	17
4–5	19
6–8	12
9–12	11

Note. The total does not sum to 22 because teachers could have experience in multiple grade levels.

II.2.2.2 Item-Writing Training

Before the three-day item-writing workshop, item writers reviewed both the Item Writer Guidelines[2] for KELPA and the 2018 Standards. During the item-writing workshop, item writers received training from AAI's content-development staff on several topics, including:

- alignment to the 2018 Standards
- relationship between the 2018 Standards and the 2017 Kansas ELA Standards
- principles of UD and accessibility
- bias and sensitivity
- item types available for KELPA

To guide the item-writing process, item writers also were trained in the structure and format of items, including stems, prompts, and answer choices (i.e., keys and distractors). Key points of these guidelines, modified from Haladyna and Downing (1989), are summarized in Table II-11.

[2] This is considered a secure document for training purposes, so no link or appendix is provided.

Table II-11: Summary of Item-Writing Guidelines

Guideline category	Key points
General	• Write items that have clearly correct answer choice(s). • Ensure that items are clearly worded. • Proofread items for correct grammar, punctuation, and spelling.
Content	• Ensure items will elicit evidence of student performance with regard to the targeted skill(s) named within the standards. • Ensure that multiple-choice items measure a single concept. • Ensure that comprehension items focus on key ideas and salient details. • Use vocabulary that is at or below students' grade level. • Write items to a variety of difficulty levels.
Structure	• Write prompts and stems as directly as possible. • Write stems in the form of questions and prompts as complete sentences.
Stem construction	• Avoid negatives in stems. • Ensure that the central task of the item is made clear in the stem and that students do not need to read through the answer options to understand what the item is asking.
Answer-choice development	• Order answer choices following option guidelines. • Create independent answer choices that do not overlap. • Write answer choices that are roughly of the same length and parallel in structure. • Do not offer "all of the above," "none of the above," or "I don't know" as answer choices. • Avoid cluing between the stem and answer choices. • Avoid specific determiners such as "always" or "never." • Create plausible distractors which stem from reasonable misinterpretations of the stimulus.
Accessibility	• Consider the access needs of disability populations and the ways in which accommodations affect an item's intent. • Utilize the simplest sentence structures that are appropriate for the assessed standard. • Minimize the use of words with multiple meanings, unless that understanding is being assessed per the standards. • 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, if needed.

Guideline category	Key points
Bias and sensitivity	• Avoid the use of stereotypes. • Avoid gender bias. • Consider the regional and cultural nuances of words. • Avoid the use of unduly negative or demeaning materials. • Avoid the use of controversial materials. • Avoid the use of upsetting or offensive materials (e.g., material with violent or sexual content). • Avoid the use of religious references such as holidays. • Ensure that items are not related to socioeconomic status or family attributes.

The item training also covered content related to security of test materials, including the following:

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

II.2.2.3 Item-Writing Process

Educators were grouped roughly equally by grade level or grade band for their writing assignments. After reviewing standards for a particular domain and learning about item types available for that domain, educators wrote and reviewed items. They were given a goal for how many items to write to different standard clusters. The time the item writers spent on their assignments varied by domain. The variance was from about two hours up to six hours. (This pattern was true for all domains.) Several resources were available to support item writers: Item Writer Guidelines for KELPA, standards for their assigned grade level or grade band, Children's Writers Word Book, and a word-level list from EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies. For reading and listening, both discrete and set-based items were needed for most grades or grade bands. AAI content-development staff distributed passages to which item writers were to write the set-based items.

II.2.3 Item Review

II.2.3.1 Internal Review

After the item-writing workshop, content-development staff reviewed items for content, alignment to standards, item type and structure, needed graphics, and language accessibility. Items received formal style editing, and instructions about needed images were provided to the graphic artists. After images were prepared, the content-development lead either approved each item or requested revisions.

II.2.3.2 External Review

Staff from AAI content development and from KSDE recruited Kansas educators for two separate types of review: content review and bias-and-sensitivity review. KELPA external review occurred in summer 2019. Item reviewers were recruited from different districts in Kansas, and all educators had experience with ELs.

Seventeen Kansas EL educators participated in the external KELPA item review. Most item reviewers had extended experiences working with ELs of multiple grade levels. There were about equal numbers of educators from elementary and secondary levels: elementary (eight educators), middle (five educators), and high school (four educators). Among these 17 educators, seven were from districts in northeastern Kansas, three from north-central districts, two from central districts in, four from southern districts, and one from a western district. Nine of these 17 reviewers conducted the content review, and the remaining eight reviewers did the bias-and-sensitivity review of the items.

Content-development staff grouped educators by grade band: K–1, 2–3, 4–5, 6–8, and 9–12. Item reviews occurred in a secure, online reviewing system within Kite. All item reviewers completed two web-based sessions of item-review training led by AAI content-development staff: use of the online review system and specialized training for bias-and-sensitivity and content reviews. The training sessions included information about the KSDE–AAI partnership, test and item security, item-writing guidelines, and the item-review process. Item-review training also provided participants with practice items and contact information for content-development staff. After completing the training, reviewers evaluated items and provided feedback at their own pace over a 2-week period.

Content reviewers examined items for:

- alignment to the relevant standard
- grade-level appropriateness of content, context, and vocabulary
- a clear, complete statement or question
- grammatically correct text
- a correct key
- accuracy of content (i.e., the content does not misrepresent the stimulus nor the real world in any way that may introduce confusion for students)
- accurate, relevant graphics
- well-designed answer choices that do not require background knowledge

Content-development staff asked bias-and-sensitivity reviewers to identify barriers that could prevent students from demonstrating what they know and are able to do when those barriers are not related to the 2018 Standards. Possible concerns included:

- potential bias related to gender, race or ethnicity, socioeconomic factors, or other
- barriers related to uncommon language, unnecessary linguistic complexity or lack of clarity, assumed prior knowledge, cultural restrictions, accessibility, or other
- sensitivity concerns related to stereotype, religion, socioeconomic factors, status, specific topic, or other

According to their analysis of items, reviewers recommended that items be accepted, revised, or rejected, and gave specific reasons for their decisions (e.g., “item aligns better to this standard”).

Content-development staff examined the reviewers’ comments and revised items as needed. The primary reasons for revisions included alignment to different standards, wording changes for grade appropriateness, clarity and specificity (especially in instructions for speaking prompts), and graphics revisions. Two KSDE program consultants with expertise in working with ELs or teaching ELA consulted on conflicting reviewer comments related to grade or standard alignment or impact on or relationship to classroom practices.

II.2.3.3 Accessibility Review

Accessibility was addressed at various points in the item-development process. Before writing items, item writers had been trained in the principles of UD and accessibility (as related to language use) to enable the widest range of students to be able to access the items they wrote. External reviewers for bias and sensitivity were also asked to look for potential accessibility issues. Moreover, internal item reviewers identified potential issues during their own accessibility review. For information about accessibility features, refer to Section II.2.2 Item Writing.

II.2.3.4 Data Review

In 2020, KELPA was administered as an operational field test. After the operational field test, the psychometric and content teams reviewed item statistics of all the operational field-tested items. Only items with acceptable psychometric properties and without any issues reported were retained and used for operational scoring. Items with acceptable psychometric properties were items that had reasonable statistics; item statistical flagging criteria are included in Appendix B. Items with statistical flags were carefully reviewed by KELPA content-development staff for potential content issues. Flagged items were used only when they did not present any content concerns.

Table II-12 shows the numbers and percentages of retained items after the data review of item statistics for the 2020 operational field test; the retention percentages ranged from 53% to 100%. The lowest retention percentage (53%) was observed in kindergarten writing. Among the flagged items in kindergarten writing, three CR items were noted by the field as excessively difficult for the grade level; for another item, graphics clarity was an issue. The low retention percentage led to a shorter writing test than original desired. Table II-13 shows the number of items for 2020 operational items by grade or grade band. Additional kindergarten writing items will be field-tested during the 2021 administration and then added to the 2022 operational kindergarten writing test. Some items with statistical flags were sound from a content perspective and were retained for blueprint coverage after careful content review, but most items with statistical flags were not retained. The most frequently observed statistical flags included too-high or too-low item difficulty and distractors not performing as well as expected. Also, items reported from the field (i.e., five writing items, one speaking item, and one listening item) to have potential graphical or content issues were carefully reviewed by content and psychometric teams and were either retained if they were not considered flawed or rejected if otherwise.

Table II-12: Number and Percentage of Retained Items for the 2020 KELPA by Domain and Grade or Grade Band

Grade or	No. of retained items by domain			
grade band	Listening (%)	Speaking (%)	Reading (%)	Writing (%)
K	23 (92)	10 (100)	19 (95)	8 (53)
1	25 (100)	10 (100)	25 (100)	13 (87)
2–3	25 (100)	10 (100)	24 (96)	19 (100)
4–5	25 (100)	10 (100)	22 (88)	17 (89)
6–8	25 (100)	9 (90)	21 (84)	18 (90)
9–12	24 (96)	10 (100)	23 (92)	17 (85)

Table II-13: Number of Operational Items for the 2020 KELPA by Domain and Grade or Grade Band

Grade or grade band	Listening items		Speaking items		Reading items		Writing items		Total no. of items	Total score points
	N	Score points	N	Score points	N	Score points	N	Score points		
K	23 (0)	23	10 (10)	30	19 (0)	19	8 (2)	12	60	84
1	25 (0)	25	10 (10)	30	25 (0)	25	13 (4)	21	73	101
2–3	25 (0)	25	10 (10)	30	24 (0)	24	19 (4)	27	78	106
4–5	25 (0)	25	10 (10)	30	22 (0)	22	17 (4)	25	74	102
6–8	25 (0)	25	9 (9)	27	21 (0)	21	18 (3)	24	73	97
9–12	24 (0)	24	10 (10)	30	23 (0)	23	17 (3)	23	74	100

Note. Numbers in parentheses are number of educator-scored items.

II.2.4 Rubric Development

Development of rubrics for writing and speaking CR items began in July 2019. The rubrics went through several rounds of revision and were completed in December 2019. A norm-referenced process was used for rubric development in which a reasonable portion of students were expected to score at each score level 0 through 3. Rubrics within the same grade or grade band in each domain of speaking and writing were generic. That is, rubrics were not item specific within each grade or grade band by domain. Table II-14 presents the five phases of rubric development.

Table II-14: Activities of Rubric Development by Phase

Phase	Activities
1: Drafting rubrics	KELPA content-development staff read or listened to sample responses together, using their expert judgment to order them by proficiency and then describe the characteristics of responses in terms of language usage, content, communicative effectiveness, etc., as included in the 2018 Standards. The standards were then consulted to check appropriateness of the selected characteristics. This procedure was the foundation for the rubrics used in Phase 2.
2: Trialing rubrics	Four 4-hour sessions led by KELPA content-development staff (two sessions focused on speaking, two sessions on writing) were held, with two to three ELA test developers who had not participated in Phase 1 work. Participants were asked to apply the draft rubrics to sample student responses without training as a measure of transparency of descriptors, ease of application, and agreement of scores assigned. Rubrics were revised accordingly.
3: KSDE review	Draft rubrics were presented to KSDE for feedback. KSDE reviewed these rubrics primarily for content and grade appropriateness. Rubrics were revised accordingly.
4: Educator review	Educators who focused on writing and speaking at the item-writing event in October 2019 then reviewed the rubrics and suggested revisions for ease of use by raters in the field. Rubrics were revised accordingly.
5: Pilot testing	Three ELA content-development staff members practiced applying all of the rubrics to all of the calibration samples. Rubrics were revised and finalized before the 2020 operational field-test administration.

II.2.5 Development of Rater-Training Materials

Using the student responses and rubrics developed in Phase 1, initial rater-training materials were developed. These training materials were presented to KSDE for feedback, along with Phase 3 rubrics.

Educators who focused on writing at the item-writing event in October 2019 also reviewed the rater-training materials for that domain and suggested revisions for ease of use by raters in the field.

Student writing and speaking responses to exemplar CR items in each grade or grade band, which later were included in the 2020 KELPA practice tests, were collected from students in one Kansas school district in November 2019. Efforts were made to gather a variety of responses in terms of first language, proficiency level, and other demographic information. Responses to two CR writing items and one CR speaking item in each grade were collected from three students in kindergarten and seven students in grade 1. Responses to one CR speaking item in grade band 2–3 and one CR speaking item in grade band 4–5 were collected from 19 students in grade bands 2–3 and 4–5. Responses to one CR speaking item in grade band 6–8 and one CR speaking item in grade band 9–12 were collected from 15 students in grades 6–12. KELPA content-development staff applied the rubrics to the student responses and selected many student responses for use as exemplar responses in the rater-training materials and as samples for additional calibration practice. For all responses in the rater-training materials, KELPA content-development staff assigned scores and wrote explanations for those scores.

Three staff members in ELA content development self-trained using the exemplar responses and explanations of scores in the rater-training materials and practiced applying the rubrics to the additional practice samples in the materials. Rater-training materials and rubrics were revised according to their feedback and finalized for the 2020 operational field-test administration.

II.3 Test Administration and Scoring

A 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 2019–2020 KELPA Examiner's Manual (Examiner's Manual hereafter), which provides information for districts, schools, and teachers regarding standardized test administration. It also provides guidance and procedures related to administration of KELPA for the 2019–2020 test administration. The Examiner's Manual includes several key pieces of information:

- overview of KELPA
- test security and ethics
- accommodations
- preparation activities before test administration
- directions for test administration on testing day
- activities for after test administration
- overview of test scoring

KELPA is entirely computer based for students in grades 2 through 12 and is delivered through the Kite Student Portal (described in Section II.3.2 Test-Administration Procedures). Student Portal must be installed on students' computing devices; headsets with microphones must be used for KELPA. Students in kindergarten and grade 1 independently complete a mostly computer-based exam, along with a small number of paper-based writing items.

The technology-practice tests were available to help students and teachers become familiar with the assessment format and the procedures for answering different types of KELPA items before the test administration. Teachers and students were strongly encouraged to use the technology-practice tests before operational administration. Students should try out the headsets and microphones with the technology-practice tests. All educators who administer and score KELPA must complete training on test administration, reporting and documenting test types and accommodations, scoring, and test security and ethics procedures. They also must have Educator Portal accounts.

The 2020 KELPA testing window was open to students from February 3 through March 20. Educators were able to enter scores for CR items from February 3 through March 27. KELPA is administered by domain in no specific order. Each KELPA domain test is designed to take approximately one class period or approximately 45–60 minutes All KELPA domain tests are untimed, and students should be given enough time to complete the tests. Qualified educators may score speaking-domain assessment items while students are testing. That is, while students take the speaking test, qualified educators may sit beside them and score their responses. This scoring method is called simultaneous scoring. An alternate scoring method is deferred scoring: students record their responses for the speaking items, and qualified educators later listen to and score the recorded audio responses via Educator Portal. Additional information about scoring can be found in Section IV.3.1 Item Scoring.

II.3.1 Test-Administrator and Scorer Training

Kansas uses a train-the-trainer model in which district test coordinators receive training directly from KSDE and, in turn, train educators in their local school districts for test administration and scoring. In partnership with KSDE, AAI offers several types of training for coordinators, including test-security training, regional trainings for District and Building Test Coordinators, virtual training for District and Building Test Coordinators, and Kite-technology training webinars. Also in partnership with KSDE, AAI offers virtual training for KELPA administration, scoring, and use of Kite. Some training sessions are stand-alone and some are blended into the general KAP virtual training sessions.

A series of test-coordinator training sessions on KELPA administration was conducted. The training webinars were recorded and posted on the KELPA Training site, along with the training slides and the frequently asked questions and responses to these questions. Educators, test administrators, and other users could review the training content at any time. District coordinators are responsible for training educators in scoring CR items in speaking and writing as well as training test-administration staff on test security and ethics. An example rater training presentation delivered by Emporia school district is available online. It indicated that district staff work with local school staff to deliver rater training at individual schools taking advantage of sample student responses in speaking and writing included in the rater training material. District coordinators are required to keep records of trainings offered locally and assign qualified educators who completed the scorer training to score student responses in the Kite Educator portal. Records for scorer training and test-security training are subject to audit conducted by KSDE and the Kansas Assessment Advisory Council. Detailed information about test-security training and audits can be found at Kansas State Department of Education Test Security Guidelines. Table II-15 describes the training dates, topics, and resources and materials.

Table II-15: 2020 KELPA Test-Administrator Training Sessions

Date	Topic	Online resource and materials
January 15	All Things KELPA! Important Dates and Deadlines Test Coordinator Training: KELPA	Webinar Slides FAQs
January 30	KELPA Materials and Documents KELPA Technology Update Monitoring Testing and Scoring Getting Ready for KAP Summative Testing	Webinar Slides FAQs
February 12	KELPA Testing Updates Important Dates and Deadlines SC [Special Circumstances] Coding for KELPA Monitor Visits	Webinar Slides
February 26	KELPA Testing Updates KAP Summative Testing Updates Important Dates and Deadlines	Webinar Slides

II.3.2 Test-Administration Procedures

The Examiner’s Manual provides clear guidelines for student participation, as well as policies and procedures to ensure a standardized and secure test administration. These guidelines were developed and approved by KSDE. Educator responsibilities before, during, and after KELPA administration are described in detail. In addition, educators and test proctors follow detailed directions regarding security and ethics, including acceptable and unacceptable practices, described in the Examiner’s Manual. Administration of accommodations is discussed in Chapter V. Inclusion of All Students of this manual.

II.3.2.1 Before KELPA Administration

II.3.2.1.1 Student Preparation

Student preparation includes technology practice before KELPA and providing instructions, advice, and information regarding taking the test during KELPA. Before KELPA administration, technology-practice tests are available and are intended to familiarize students and teachers with the assessment format and the procedures for answering the different types of KELPA items. Technology-practice tests are provided for students within grade bands K–1, 2–5, and 6–12. Table II-16 summarizes item counts for the technology-practice tests. Each practice test covers all item types (including CR items) that appear in the corresponding summative grade or grade-band tests. Technology-practice tests are not secure and should be used liberally to help students understand how to listen to directions and assessment media, enter responses, access test-taking tools, and navigate through an assessment.

Table II-16: Number of Items in Technology-Practice Tests by Domain and Grade or Grade Band

Grade or grade band	Listening	Reading	Writing	Speaking
K–1	4	3	6	2
2–5	4	9	5	2
6–12	5	5	4	2

II.3.2.1.2 Test-Administrator Preparation

Before local testing, all educators who will administer and score KELPA writing and speaking domains must complete training on test administration, scoring, and test security and ethics procedures. Before KELPA administration, educators register students for KELPA, create accounts in Educator Portal, install Student Portal, complete initial Personal Needs Profile settings, and have the paper-based writing booklet[3] for kindergarten and grade 1 ready, if applicable. Other preparations include ensuring a quiet environment and room arrangement for the speaking-domain assessment, downloading KELPA Test Administration and Scoring Directions files from the Help tab in Educator Portal for all grades, room preparation, and obtaining the materials named in the checklist. Detailed information can be found in Section 4. Before KELPA of the Examiner’s Manual.

[3] Master copies of the paper-based booklets for kindergarten and grade 1 are in the KELPA Test Administration and Scoring Directions for Writing files and can be downloaded from the Help tab in Educator Portal.

II.3.2.2 During KELPA Administration

On assessment day, test administrators inspire confidence in and prepare students for testing by providing them with relevant information (e.g., reminding them they will hear and see the questions in English, that they must answer the questions in English, that they will have as much time as they need to answer the questions, etc.). Test administrators make sure students sit at a table or desk with enough space and encourage them to attempt all items to the best of their ability. Test administrators also direct students to use Student Portal correctly and to use the guidelines and scripts specific for each domain during the administration. Test administrators provide technology assistance, but the assistance must be limited to the technology directions only. They are also responsible for maintaining security: staying in the room with students to prevent their access to assessment materials, continuously circulating through the room to ensure no unauthorized use of electronic devices, etc. More information can be found in Section 5 of the Examiner’s Manual.

II.3.2.3 After KELPA Administration

After KELPA administration, test administrators monitor student testing status (see Section II.4 Monitoring Test Administration of this chapter for more information), reactivate student testing sessions if needed (e.g., session ended before the student was finished), and enter special circumstances codes for students who could not take or complete KELPA. At test completion, teachers must verify that all questions have been answered via the Review/End screen in Student Portal before a student exits a domain assessment. All materials, including scratch paper, need to be collected and destroyed. Detailed steps of these procedures are in Section 6 of the Examiner’s Manual.

Also after administration, educators score the items when deferred scoring was chosen; each item is scored individually after the student has recorded all responses in Student Portal.

II.4 Monitoring Test Administration

District and Building Test Coordinators can monitor student test progress via Educator Portal. The test coordinator can observe which students have finished testing, which ones still have sessions to finish, and which ones have incomplete tests. Depending on role and type of test, some users can use Educator Portal to monitor test sessions in real time. Real-time monitoring increases the load on Student Portal and local bandwidth. To keep real-time monitoring to a minimum, only users with the following roles can monitor student progress using Educator Portal:

- Building Test Coordinator
- building user
- District Test Coordinator
- district user

During the 2020 testing window, KSDE staff visited a sample of Kansas schools to monitor administration and test security.

Agile Technology Solutions (ATS), a center of AAI that oversees and manage the Kite system, hosted regular check-in calls from test administrators to monitor common issues and concerns during the testing window. Before the opening of the testing window, the Kite automated-enrollment process assigned tests to all rostered students. The process ran nightly throughout the window, assigning tests as additional students were added to the roster. Throughout the testing window, ATS ran a series of data queries at noon and midnight each day to identify testing irregularities. Specifically, a student or school that met at least one of three criteria was reported within the Kite dashboard for further review:

- A student completed a test in a short amount of time (i.e., under 5 minutes)[4].
- A test section started or ended outside of standard school hours.
- A student's test session was reactivated.

During the operational window, the AAI psychometric team monitored test delivery periodically to ensure quality administration. The monitoring process included computing basic frequency statistics to verify that counts appeared as expected by grade and domain, and gathering data to run scoring and key checks as an additional safeguard against incorrect scoring approximately two weeks after the testing window opened.

II.5 Test Security

Three important facets of test security need to be protected: the integrity and confidentiality of test materials, test-related data, and personally identifiable information. The protection should be present through the whole testing cycle from test development and administration to scoring and reporting. Both physical security and online-platform security requirements are needed to protect the security of test materials. Also, strict procedures are in place during administration and reporting to protect the security of test and student data.

Kite platform runs all production on Amazon Web Services (AWS) services in high availability mode with no single point of failure. This ensures that loss of any given server or even an entire availability zone (data center) will have minimal impact on Kite platform availability. Recovery times range from no downtime for loss of most servers to a few minutes for loss of an entire data center. Recovery for services running in high availability mode is automatic and fully managed by AWS. The Kite platform has a multi-layered design to prevent denial of service attacks and system intrusion, a service provided by AWS. AWS provides a central view of the security posture by consolidating findings from other security services and facilitates automated security checks to minimize disruptions in case of cyber-attacks. Since moving to AWS in 2017, no Kite platform outages have affected testing. The only outages have been during preapproved outage windows for updates, maintenance, and support.

KSDE has a detailed plan to ensure the security and confidentiality of state testing materials, which includes the following steps:

- Using a train-the-trainer model, all district test coordinators (DTCs) are trained yearly on test security and the components of test security.
- Each DTC verifies completion of training by signing an Agreement to Abide by Guidelines.
- Before local testing, DTCs train district- and building-level personnel involved in the administration of state assessments.
- Local personnel sign an agreement to abide by state ethical testing practices.
- DTCs provide the State Assessment Office with accurate testing schedules through Educator Portal.
- DTCs retain documentation related to test security.

[4] The time threshold was set with guidance from KSDE.

- To monitor test security, KSDE staff and members of the Kansas Assessment Advisory Council annually visit 5%–10% of Kansas schools during test administration.
- KSDE uses a monitoring checklist to evaluate testing sessions. Upon completion of monitor visits, all checklists are analyzed for discrepancies and potential security violations.

These steps, especially the monitor visits, effectively identify possible test irregularities. Identified testing discrepancies and potential security violations are reported to KSDE. Upon breach of security, appropriate consequences are initiated at the district level. Because each case is unique, possible steps vary and may include but are not limited to:

- no action because the breach was not severe enough to warrant any
- 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 conducts a follow-up monitor visit the next testing year to ensure changes to inappropriate practices have been made

For more details, refer to the 2020–2021 Kansas Assessment Fact Sheet: Test Security and Ethics and Kansas State Department of Education Test Security Guidelines.

II.5.1 Test Materials Security

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. ATS and the Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS, also part of AAI) are in secure wings that can be accessed only with a key. In general, work is done at one of our sites or using secure server systems via a secure virtual private network connection.

The electronic item bank, online administration system, and student responses are stored in Kite Suite, which is designed and maintained by ATS. Multiple portals are designed within the Kite Suite to serve the needs of item and test development (i.e., Content Builder), for educators to input and access test and student information (i.e., Educator Portal), and for online testing (i.e., Student Portal). All released items exist in a separate pool from items used for summative purposes, ensuring that no items are shared between secure and nonsecure pools. Only authorized users of Kite have access to items. Any external item-review panels of Kansas educators were required to sign nondisclosure agreements to ensure item and task confidentiality and security.

As to the paper-based writing booklets for kindergarten and grade 1, no copies of the booklet pages are allowed for use in an overhead projector or for any other purpose. The test booklets are kept secure and must be returned to the test administrator. All test materials, including test booklets, must be shredded after scores are entered in Kite.

II.5.2 Test-Related Data Security

For test administration, all Kite applications handle educator and administrative passwords using industry-standard encryption techniques; users must create strong passwords that they may change at any time in accordance with the password policy. All applications generate records that can be reviewed by system administrators to track access. Access to individual Kite applications is controlled according to the policies set forward for that application and the data the application maintains. All access policies and accounts are reviewed periodically to ensure that access to systems is limited to the appropriate populations. DTCs attend annual training provided by KSDE regarding test security and oversee test security for the entire district. They establish procedures that determine appropriate personnel access to Educator Portal and role assignments within the district, such as test administrators and test scorers. DTCs also use Educator Portal to remove or deactivate users who are no longer qualified by the end of September. If any breach of test security, loss of materials, or any other deviation are found within the school district, DTCs are responsible for reporting them to the KSDE assessment coordinator. For more information about DTC responsibilities, refer to Kansas State Department of Education Test Security Guidelines.

ATS monitors and provides data to KSDE that examines potential areas of test irregularities throughout testing. Data reported by ATS, using the Kite Dashboard, include

- DTC training log
- frequency of test reactivations
- click history
- allowable tests taken after school hours (e.g., students enrolled in a virtual schools test after the end of a typical school day).

For more information, refer to Kansas State Department of Education Test Security Guidelines. Also during administration, test security is promoted through required training for test administrators and the signing of the security agreement. Test administrators are expected to deliver assessments with integrity and maintain the security of assessments. State, district, and school users are expected to complete the security agreement within Educator Portal each year. By signing the security agreement, users agree not to store or save assessment materials to computers or personal storage devices, print assessment materials, or share personal passwords with others. Test administrators needed to follow test procedures outlined in the Examiner’s Manual, the KELPA Test Administration and Scoring Directions for each grade, and those in the training received regarding security and ethical practices for testing from their districts. They also needed to follow established district or building procedures for collecting and destroying testing materials (e.g., student notes, scratch paper, drawings) upon completion of each test session and the entire test.

II.5.3 Security of Personally Identifiable Information

In accordance with the Family Educational Rights and Privacy Act (FERPA), students’, teachers’, operators’, and administrators’ access to personal student data is limited to student records in which that person has a legitimate educational interest. All users are provided the minimum amount of necessary access. Throughout each school year, security levels, groups, and access are reviewed periodically to ensure continued compliance. DTCs inform staff that personally identifiable information (PII) should not be conveyed when testing issues are reported. The documentation for Kansas regarding allowable identifiers in an email specify that only the Student State ID number, but no other identifying details, should be provided in an email. Student PII cannot be sent via email or Live Chat when contacting the Kite Service Desk that provides support for Educator Portal and Student Portal. Building

Test Coordinators (BTCs) and educators and test proctors follow procedures established by the DTC for all aspects of testing and ensure the test security within the individual building site.

Operational access to all servers is controlled by keys that are provided only to system administrators who manage the production data center in the operations team. Access to the networking equipment and hardware consoles is limited to the data center itself; remote access to these devices is limited to the data-center-specific administration host. AAI staff working on Kansas programs are required to complete annual KSDE information technology security and data-privacy training to ensure compliance with FERPA.

Students' assessment data (e.g., the return files, score reports) are placed on a secure drive that only specified members of the ATS and KSDE team members can access. Descriptions of KELPA results in technical documentation are only reported at the aggregated level. There must be more than 10 students to provide grouping data.

II.5.4 Accommodations-Related Security

District- and building-level personnel (i.e., any staff member who administers a state assessment) must sign an agreement to abide by state ethical testing practices and receive training on ethics of testing, test security, and reporting and documenting accommodations. In the train-the-trainer model, DTCs train building-level personnel, and BTCs in turn assist DTCs with training and/or train building-level personnel in test-security procedures, ethics of testing, and reporting and documenting accommodations, before local testing begins. To ensure security related to accommodations, DTCs keep records of documentation for text-to-speech accommodations and any other accommodation that requires a deviation from the general administration of the assessment. They establish procedures for teachers working with students with disabilities to enter student accommodation information into the Personal Needs Profile (PNP) in Educator Portal. Section V.4.1 Selection of Accommodations provides more information about selecting and entering information in the PNP. Before the assessment, either DTCs or BTCs provide documentation for accommodations entered in the PNP. During the assessment, Kite audio (headsets), not human reader, is used for the text-to-speech accommodation.

III. Technical Quality—Validity

According to the Standards for Educational and Psychological Testing, validity refers to “the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests.” (American Psychological Association [APA] et al., 2014, p. 11). The Standards for Educational and Psychological Testing (APA et al., 2014) also describes the five sources of evidence that should be considered when evaluating test-score validity: evidence based on (a) test content, (b) response processes, (c) internal test structure, (d) relationships between test scores and other variables, and (e) consequences of testing. For example, when item response theory (IRT) is used to analyze test data, model assumptions, such as parameter invariance, should be evaluated. To support the proposed test-score interpretations and uses, a variety of validity evidence should be collected in an ongoing validation process. This chapter of the technical manual describes aspects of KELPA that support intended test-score interpretations and uses.

III.1 Validity Evidence Based on Test Content

Validity evidence based on test content is used to demonstrate that the content of the test is related to the specific content domains the test was intended to measure. Evidence of content validity for KELPA includes the alignment between KELPA items and the 2018 Standards, as well as the alignment between the test and test blueprint. The following evidence is used to evaluate the content validity of KELPA assessments:

- the development of the test blueprint and specifications
- the relationship between the blueprint and the 2018 Standards
- content reviews of KELPA items using a panel of content experts to see whether the items measure the intended construct or whether potential sources of construct-irrelevant variance exist
- fairness reviews of KELPA items to avoid bias-and-sensitivity issues related to specific subpopulations

Chapter II: Assessment System Operations of this technical manual presented content-based validity evidence related to development of the test blueprint, item and test development and content, fairness, and posttest-administration data review. It described the development of the test blueprints and the relationship of the blueprint with the 2018 Standards. All KELPA items were developed to align with the 2018 Standards, and item development followed well-established procedures. After items were developed, they underwent multiple rounds of internal and external review, such as editorial review, content review, bias-and-sensitivity review, and KSDE review. After the operational field-test administration, item statistics were computed to initiate a post-administration review. Items had to pass the data review from both content and psychometric perspectives before they contributed to operational scores. Moreover, Chapter II described tests that were administered according to standardized procedures with accommodations for students with disabilities. Specific efforts to ensure content validity are summarized.

- The construction of the blueprint is a collaborative process between Achievement & Assessment Institute (AAI), Kansas State Department of Education (KSDE), and educators in Kansas. Groups of 2018 Standards (clusters) are used in organizing the test blueprints to ensure that ranges and variety of standards measured in KELPA are appropriate.

- A summary of clusters by standards in Section II.1.1 Test Blueprints and proportion of items by clusters (i.e., Table II-3, Table II-4, Table II-5, and Table II-6 of this manual) show that each cluster has an adequate number of items to represent the knowledge and skills described in the 2018 Standards.
- Kansas English learner (EL) professionals (i.e., teachers and district coordinators) are selected and trained to ensure they write high-quality items.
- Item writers are trained using detailed item-writing guidelines (key points from several areas such as content, structure, accessibility, etc.; see Table II-11).
- Item writers participate in guided item writing.
- External content reviewers review each item to make sure all items align with the 2018 Standards; they also consider grade-level appropriateness, graphics, grammar and punctuation, language demand, and distractor reasonableness.
- External bias, fairness, and sensitivity reviewers review items for issues related to gender, race or ethnicity, socioeconomic factors, sensitivity concerns, and other factors.
- Before items are selected for operational use, several statistical analyses are conducted, including classical item analysis and distractor analysis. Content-development and psychometric staff from AAI again carefully review items' statistical characteristics.
- Administration of KELPA assessments is standardized and includes accommodations for students who need them. Furthermore, the tests are untimed to avoid any issues related to haste.

III.2 Validity Evidence Based on Response Processes

Response-process evidence pertains to the extent to which the cognitive skills and processes students use to answer an item match those targeted by the standards aligned to the item. For English language proficiency assessments, evidence can include the extent to which the linguistic processes students use to answer an item match those targeted by standards. The validity evidence based on response processes for KELPA includes educator review of the linguistic processes of items.

During item development, item writers considered the linguistic processes required by the standards at different performance levels and were asked to use language in items that elicited the same intended linguistic skills or knowledge. During the content-review stage, item writers also evaluated items to make sure they stimulated the intended linguistic process as indicated by the requirements of the standards.

In summer 2019, as part of the external review of items, a group of 17 Kansas EL educators evaluated whether the linguistic process required by the items aligned with the intended linguistic process inherent in the targeted standard. Among these 17 educators, seven were from districts in northeastern Kansas, three from north-central districts, two from central districts, four from southern districts, and one from a western district. Because a large percentage of the state population is located in northeastern Kansas, more teachers were from this area than other areas. The educators reviewed 99 kindergarten items, 114 items for grade 1, 115 items for grade band 2–3, 121 items for grade band 4–5, 145 items for grade band 6–8, and 119 items for grade band 9–12, across four domains. The reviewers appraised each item, considering several questions:

- Does the item use grade-appropriate vocabulary?
- Does the item allow students to use their EL knowledge and skills rather than rely on background knowledge outside of the content area?
- Does answering the item allow students to exhibit the linguistic behavior associated with the linked standard and performance expectation? If not, should it be revised, linked to another standard (or performance level), or rejected?

Most items required student response process that were consistent with the response processes anticipated by the test developers. For items in which educator feedback suggested a need for tighter alignment, edits were made to elicit the intended responses. Several steps were taken as a result of the educator feedback:

- updating item-standard alignment
- revising wording for grade appropriateness or clarity
- rejecting items that were based on students' understanding of grammatical terms rather than on applying knowledge

III.3 Validity Evidence Based on Internal Structure

According to the Standards for Educational and Psychological Testing (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). For KELPA, overall proficiency levels are derived from student performance on the domains (i.e., domain performance levels). Four separate unidimensional IRT models were used to fit the four domain tests, respectively, at each grade or grade band. Thus, unidimensionality is assumed for each domain test, that is, the items on the domain test all load on the domain construct. This assumption is examined first. Then the IRT model used to fit the data and the calibration process are described, followed by an examination of three IRT model assumptions: fit of item response function, local independence, and parameter invariance. Finally, items were examined using differential item functioning (DIF) on gender and ethnicity to help identify any internal consistency issues related to items performing systematically differently for student groups.

III.3.1 Dimensionality Study

For KELPA, confirmatory factor analysis (CFA) was conducted for each grade or grade-band domain test to evaluate whether a model with one dominant dimension fit the data reasonably well. CFA was carried out using tetrachoric or polychoric correlations for binary or ordinal item responses, respectively, and with robust weighted least-squares estimation using the lavaan R package (Rosseel, 2012). The one-factor CFA model was considered to fit when the comparative fit index (CFI) was .95 or greater and the standardized root mean square residual (SRMR) was .08 or smaller (Hu & Bentler, 1999). Table III-1 presents CFIs and SRMRs for each domain by grade or grade band. For all domains and all grades or grade bands, the CFIs were greater than .95, and SRMRs were smaller or equal to .08. Overall, each domain test can be considered unidimensional.

Table III-1: Summary of the Confirmatory Factor Analysis by Domain and Grade or Grade Band

Grade or grade band	Listening		Speaking		Reading		Writing	
	CFI	SRMR	CFI	SRMR	CFI	SRMR	CFI	SRMR
K	.97	.07	1.00	.03	.93	.06	.99	.06
1	.96	.07	1.00	.03	.99	.05	.98	.06
2–3	1.00	.03	.99	.05	.99	.05	.98	.05
4–5	1.00	.04	1.00	.03	.99	.04	.99	.04
6–8	.99	.04	1.00	.03	.99	.04	.99	.05
9–12	1.00	.03	1.00	.01	.99	.03	.97	.08

Note. CFI = comparative fit index; SRMR = standardized root mean square residual.

Correlations among domain raw scores are empirical evidence for evaluating the relationships among the four domain tests. To understand the relationships among the four domains, correlations and disattenuated correlations among domains' raw scores were calculated (see Table III-2). The disattenuated correlation is an estimate of the correlations among underlying domain true scores, which take into account the reliability of each domain score.

Table III-2: Correlations (C) and Disattenuated Correlations (DC) Among Domains by Grade or Grade Band

Grade or grade band	Listening vs. reading		Listening vs. speaking		Listening vs. writing		Reading vs. speaking		Reading vs. writing		Speaking vs. writing	
	C	DC	C	DC	C	DC	C	DC	C	DC	C	DC
K	.44	.57	.46	.52	.42	.52	.41	.50	.66	.88	.47	.57
1	.61	.70	.57	.65	.59	.72	.50	.55	.72	.85	.56	.65
2–3	.68	.77	.63	.70	.66	.76	.53	.58	.80	.91	.62	.69
4–5	.74	.85	.65	.72	.72	.84	.56	.65	.76	.92	.66	.75
6–8	.71	.84	.63	.69	.74	.85	.52	.58	.71	.83	.66	.73
9–12	.70	.80	.56	.61	.68	.81	.49	.53	.68	.81	.59	.67

Because unidimensionality is assumed for each domain test, specific relationships between domains are not assumed. However, medium to large correlations among domains are expected because it is difficult to completely isolate individual domains in listening, speaking, reading, and writing for English language proficiency. For example, the writing test still requires students to read the prompts even if the reading requirement for a writing prompt is designed to be below grade level. According to Cohen (1988), correlation coefficients around .10 are considered small correlations, around .30 are considered medium correlations, and larger than .50 are considered large correlations. The correlations among domains ranged from .41 to .80, which are medium to large correlations. The lowest correlation for KELPA was between reading and speaking for kindergarten (r = .41), and the highest correlation was between reading and writing domain scores for grade band 2–3 (r = .80). The disattenuated correlation among domains ranged from .50 to .92, which are considered large correlations. The lowest disattenuated correlation was between reading and speaking for kindergarten (r = .50), and the highest correlation was between reading and writing domain scores for grade band 4–5 (r = .92). Overall, these results are consistent with those of other English language proficiency assessments (e.g., Arizona English Language

Learner Assessment, Arizona Department of Education, 2017; ACCESS for ELLs® 2.0 English language proficiency test, Center for Applied Linguistics, 2018).

III.3.2 Item Response Theory and Model Assumptions

Validity inferences obtained from applying IRT models depend on the degree to which assumptions of the models are met and how well the models fit the data. In this section, the assumptions about the fit of item response function, local independence, and item-parameter invariance are evaluated.

III.3.2.1 Item Response Theory Calibration

IRT was used to analyze student responses to KELPA items and calibrate item parameters to create domain-specific scales. Only the items retained from operational field testing were included in calibration. For information about the number of items retained, please refer to Section II.2.3.4 Data Review. The following sub-sections introduce the IRT models, the sample used for calibration, the psychometric software, and the calibration procedures used for KELPA.

III.3.2.1.1 Item Response Theory Model

The two-parameter logistic (2PL) model is used to fit dichotomous items, and the graded-response model[5] is used to fit polytomous items. These two models allow for both item-difficulty and discrimination parameters to be estimated freely. The probability of student j achieving item score c = {0,1,. . . , $C - 1$} for item i = {0,1,. . . , I} is represented by $P_{ic}(\theta_j)$. The probability takes the same form for both the 2PL and graded-response models.

$$P_{ic}(\theta_j) = \frac{1}{1 + exp(-a_i\theta_j + b_{ic})}, \qquad \text{(III-1)}$$

where a'_j is the item-discrimination parameter, θ is the ability-level theta, and $b_{jc}$is the item-difficulty parameter for score category c. Note that the 2PL model is equivalent to a graded-response model when $C = 2$.

III.3.2.1.2 Sample

The student data file was cleaned before calibration. The estimation sample included all students who completed at least 50% of the items for a domain test. During calibration, missing data from omitted items were coded as incorrect. Table III-3 provides the number of students used for calibration by domain tests and grade or grade band.

[5] The graded-response model is the polytomous counterpart of the 2PL model.

Table III-3: Student Sample Size for Calibration by Domain and Grade or Grade Band

Grade or grade band	Listening	Speaking	Reading	Writing
K	4,537	4,520	4,536	4,496
1	4,606	4,597	4,601	4,584
2–3	8,772	8,746	8,769	8,762
4–5	6,980	6,963	6,973	6,980
6–8	8,092	8,067	8,093	8,093
9–12	10,773	10,741	10,780	10,792

III.3.2.1.3 Software

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

III.3.2.1.4 Calibration Procedures

Because the test design and the process of item-development are based on four separate domains and student performances are reported at the domain level, the four domain tests were calibrated separately (i.e., domain by domain) using unidimensional models. In other words, the item parameters for each domain (within each grade or grade band) were obtained separately in the four separate calibrations. The decision to implement separate unidimensional models for each domain test was informed by the Kansas Technical Advisory Committee and in collaboration with KSDE.

III.3.2.2 Model Fit of the Item Response Function

The Q1 chi-squared (χ^2) fit statistic (Bock, 1972; Yen, 1981) was used to evaluate the model fit for individual items. The mirt package in R computed this statistic during item calibration. The Q1 χ^2 fit statistic of one item followed the χ^2 distribution with degrees of freedom (df) equal to the number of possible total scores minus 1. The χ^2 tests are sensitive to sample size; therefore, the effect size was also used 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 0.01 and exhibited a medium to large effect size were flagged for model-fit issues. Table III-4 presents the number of items, the number of misfit items, and the percentage of misfit items by domain and grade or grade band. For each grade or grade band, the number of items flagged for model-fit issues was very small, and those flagged items will be monitored in future testing.

Table III-4: Item Response Theory Item Model-Fit Results by Domain and Grade

Grade or grade band	Listening			Speaking			Reading			Writing		
	No. of items	No. of MF items	% MF items	No. of items	No. of MF items	% MF items	No. of items	No. of MF items	% MF items	No. of items	No. of MF items	% MF items
K	23	0	0	10	0	0	19	2	11	8	1	13
1	25	0	0	10	0	0	25	0	0	13	0	0
2–3	25	0	0	10	0	0	24	0	0	19	0	0
4–5	25	0	0	10	0	0	22	0	0	17	0	0
6–8	25	0	0	9	0	0	21	0	0	18	0	0
9–12	24	0	0	10	0	0	23	0	0	17	0	0

Note. MF = misfit.

III.3.2.3 Local Independence

A foundational assumption of the IRT model is local independence, which posits that a student’s responses are independent from each other; that is, a response to one item is not affected by responses to other items. An example of violation of local independence is when answering an item correctly depends on correctly answering a previous item. That is, if a student answers the first item incorrectly, the probability of answering the second item correctly is then zero, regardless of how easy or difficult the second question is. A more subtle violation of local independence occurs when either the question itself or one of the answer choices provides information (i.e., cues) that changes the probability of correctly responding to another question in the same test.

The chi-squared (χ^2) based local dependence (LD) statistic (Chen & Thissen, 1997) was used to detect the item pairs with LD. The mirt package in R computes this statistic during item calibration. The χ^2 LD index of one item pair follows the χ^2 distribution with degrees of freedom (df) equal to 1. The χ^2 tests are sensitive to sample size; therefore, the effect size was also used to evaluate item fit. The effect size for χ^2 tests was calculated using Cramér’s V (Cramér, 1946). Across all domains and all grades or grade bands, only one pair of items was detected with a medium effect size LD. This pair of items was reviewed, and no cuing between these two items was found.

III.3.2.4 Parameter Invariance

For different groups of examinees, IRT models assume that item-parameter estimates are invariant up to a linear transformation. Pearson product-moment correlations and root mean square error are used to evaluate the relationship between the item parameters estimated from student groups that are expected to have the same ability distributions. The invariance assumption is met when the estimated item parameters for two samples are highly correlated.

Students were randomly divided into two groups, and their response data were calibrated separately. Correlations between the two sets of item-parameter estimates (i.e., a and b) obtained from random groups are presented in Table III-5. All correlations between the two sets of item parameters were high (i.e., at or near 1). Thus, the parameter-invariance assumption was met for all domains and all grades or grade bands.

Table III-5: Correlations (Corr.) of Item-Parameter Estimates for Random Samples

Grade or grade band	Listening		Speaking		Reading		Writing	
	Corr. a	Corr. b	Corr. a	Corr. b	Corr. a	Corr. b	Corr. a	Corr. b
K	.98	.99	.83	1.00	.96	.99	1.00	.99
1	.94	.99	.95	1.00	.98	1.00	.98	.99
2–3	.98	1.00	.99	1.00	.99	1.00	.99	1.00
4–5	.99	1.00	.89	1.00	.98	.99	.99	1.00
6–8	.98	.98	.86	1.00	.99	1.00	.99	1.00
9–12	.98	.99	.98	1.00	.99	1.00	1.00	.99

III.3.3 Differential Item Functioning

DIF analysis examines whether an item shows any statistical difference between two groups of students when the ability level is controlled. Logistic regression was used to detect items with uniform DIF. According to Jodoin and Gierl's (2001) DIF classification criteria, when the DIF test is significant, a moderate DIF has a Nagelkerke R^2 change between .035 and .07, and a large DIF has a Nagelkerke R^2 change greater than .07.

DIF was examined across gender (i.e., female vs. male) and ethnicity (i.e., Hispanic vs. non-Hispanic) student groups. All items retained for operational scoring were included in the analysis. No items were flagged for moderate or large DIF, gender-related DIF, or ethnicity-related DIF across grades or grade bands and all four domains[6]. The lack of items with DIF provided evidence for the effectiveness of bias-and-sensitivity-related training and guidance. Issues of bias have been addressed over time within AAI by providing item writers and content reviewers effective training on bias and sensitivity, as well as guidance for item writing. During item-writing training, item writing, and internal and external item reviews, AAI had emphasized the concept of developing unbiased items. From an overall program perspective, this effort resulted in a decrease in the number of items flagged for DIF over time.

III.4 Validity Evidence Based on Relations to Other Variables

According to the Standards for Educational and Psychological Testing (APA et al., 2014), "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" (p. 16). The external assessments used for collecting this piece of validity evidence are the Kansas Assessment Program (KAP) English language arts (ELA) and mathematics assessments, which are administered annually to students in grades 3–8 and 10. The correlation between KELPA domain scores and KAP ELA

[6] The analysis was repeated for selected grades by a second team member to ensure the results were accurate.

scores or KAP mathematics scores can provide validity evidence based on relationships to other variables. However, because of the COVID-19 pandemic, KAP was not administered in 2020. AAI plans to evaluate these relationships in subsequent years.

III.5 Validity Evidence Based on Consequences of Testing

The primary intended use of KELPA results is to identify ELs who are English proficient and can be exempt from English for speakers of other languages services. For ELs who still require those services, KELPA is intended to help monitor their progression toward English proficiency. Section IV.3 Scoring and Scaling includes the description of how items and domain tests are scored. After test scores are calculated, the domain performance levels are determined according to established cut scores. Chapter VI. Academic Achievement Standards and Reporting describes the standard-setting process used to set the cut scores and how overall proficiency levels are calculated according to the domain performance levels. Section IV.3.4.2 Test Results for All Students summarizes the percentage of students in each overall proficiency level. Moreover, a sample of a KELPA student score report is shown in Appendix D, and the interpretation of the student score report is described in Section VI.5.1 Student Reports. To help educators and parents interpret KELPA results and understand students' progress toward proficiency, AAI also provides the KELPA Educator Guide (see Appendix E) and the KELPA Parent Guide (see Appendix E).

Because 2020 was the first year the KELPA was administered, there was limited opportunity to collect data to evaluate consequential validity evidence. Because of COVID-19 school closures, the planned teacher survey associated with the 2020 administration was cancelled. In 2021, we plan to collect baseline data through a teacher survey, including questions about the utility of the test, importance of skills measured, and alignment of expectations of students to what is needed in the classroom.

IV. Technical Quality—Other

This chapter provides evidence related to the technical quality of KELPA, including reliability-related evidence, fairness and accessibility, and item statistics. This chapter also describes the item- and test-scoring procedures and test-results summary. Finally, a description of the quality-control steps taken to ensure the accuracy of test scores is provided.

IV.1 Reliability

Reliability is the degree of consistency of students' test scores across repeated measures. A reliable test means a student's test scores from multiple standard administrations under the same testing conditions are relatively stable. However, it is not feasible for a student to take the same test multiple times without any changes to the testing conditions. Therefore, reliability is typically estimated from student-response data rather than calculated directly. According to the Standards for Educational and Psychological Testing (American Psychological Association [APA] et al., 2014):

> The term reliability has been used in two ways in the measurement literature. First, the term has been used to refer to the reliability coefficients of classical test theory, defined as the correlation between scores on two equivalent forms of the test, presuming that taking one form has no effect on performance on the second form. Second, the term has been used in a more general sense, to refer to the consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported (e.g., in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various indices of classification consistency). (p. 33)

In this section, we calculated the reliabilities of KELPA in two ways: reliability coefficients from classical test theory (CTT) and IRT information functions as well as conditional standard error of measurement. For the CTT reliability coefficients, the student-group reliabilities also were calculated. Moreover, indices of classification consistency and accuracy of different domain performance levels are also provided.

IV.1.1 Test Reliability

Because KELPA uses only one fixed form for each domain test at each grade level or within each grade band, the coefficient alpha index of internal consistency (Cronbach, 1951) from CTT is calculated. The formula (i.e., Equation IV-1) for the coefficient alpha index is:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} \quad \sigma_i^2}{\sigma_x^2}\right], \tag{IV-1}$$

where k is the number of items on the test form, σ_i^2 is the variance of item i, and σ_x^2 is the total test variance. KELPA reliability coefficients by domain and grade or grade band can be found in Table IV-1. Reliabilities of the KELPA domain tests were adequate, with indices ranging from .81 to .97 across the majority of grade levels or bands and domains. The three exceptions were for the reading and writing tests in kindergarten (.73 and .75, respectively) and writing in grade band 9–12 (.79). Test length and test reliability are closely related, and shorter tests are usually less reliable. Not surprisingly, these domain tests also had the fewest score points within the domain. Table II-13 indicates the test lengths and total score points for all domain tests. On the other hand, speaking tests have very high reliabilities because these tests have more total score points compared to other domain tests. As the plan is to add more items to the kindergarten writing test (which will increase the test length), an increase in reliability estimates is expected.

Table IV-1: Coefficient Alpha by Domain and Grade or Grade Band

Grade or grade band	Listening	Speaking	Reading	Writing
K	.83	.91	.72	.75
1	.84	.91	.89	.81
2–3	.88	.91	.90	.84
4–5	.88	.93	.81	.82
6–8	.87	.95	.81	.85
9–12	.88	.97	.85	.79

IV.1.1.1 Student-Group Reliability

Reliability estimates were also calculated by student group[7] and are presented in Table IV-2. Results show that the student-group reliabilities were very similar within a domain and grade level or band. Also, the student-group reliabilities were similar to the overall reliabilities, with the majority of the estimates in the .80s to .90s; reading and writing in kindergarten had lower reliabilities, mostly in the .70 range. For the tests that demonstrated lower overall reliabilities (i.e., reading and writing in kindergarten and grade band 9–12), a slight decrease in estimates for students with disabilities (SWD) was noted when compared to students without disabilities. Again, as additional items are added to the kindergarten writing test, it is expected that these estimates will increase and student-group reliabilities will be reevaluated. The sample size of each student group can be found in Section IV.3.4.1 Test Enrollment Data.

[7] Economic disadvantaged (ED) 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: Coefficient Alpha for Student Groups by Domain and Grade or Grade Band

Domain and grade or grade band	Coefficient α							
	Female	Male	White	Non-White	Hispanic	Non-Hispanic	SWD	NSWD
Listening								
K	.83	.83	.83	.84	.83	.84	.84	.83
1	.84	.85	.84	.86	.84	.86	.84	.84
2–3	.87	.88	.87	.89	.88	.88	.87	.88
4–5	.87	.89	.88	.90	.88	.88	.87	.88
6–8	.87	.87	.86	.89	.87	.87	.83	.88
9–12	.87	.89	.88	.88	.88	.88	.84	.89
Speaking								
K	.91	.92	.91	.92	.91	.91	.92	.91
1	.92	.91	.91	.92	.91	.92	.91	.91
2–3	.92	.91	.91	.92	.91	.91	.90	.92
4–5	.94	.93	.93	.94	.93	.94	.91	.94
6–8	.95	.94	.94	.95	.95	.95	.94	.95
9–12	.97	.96	.97	.97	.97	.96	.96	.97
Reading								
K	.72	.71	.68	.77	.67	.79	.65	.72
1	.89	.90	.89	.90	.89	.90	.88	.89
2–3	.89	.90	.90	.90	.90	.89	.88	.89
4–5	.80	.82	.80	.83	.81	.82	.77	.80
6–8	.80	.82	.81	.82	.81	.83	.79	.81
9–12	.84	.86	.85	.86	.85	.87	.82	.85
Writing								
K	.75	.75	.73	.79	.72	.80	.71	.75
1	.79	.82	.80	.81	.80	.82	.83	.79
2–3	.84	.85	.84	.85	.84	.84	.84	.84
4–5	.82	.82	.81	.84	.82	.82	.78	.82
6–8	.85	.85	.84	.86	.84	.86	.81	.85
9–12	.79	.79	.80	.79	.79	.81	.71	.80

Note. SWD = students with disability; NSWD = students without disability.

IV.1.2 Test Information Function

As KELPA is scored using IRT (see Section III.3.2 Item Response Theory and Model Assumptions), a test information function (TIF) can be estimated for each individual level of theta across the whole performance continuum. The TIF is the sum of item information of all operational items on the test and is used to estimate the amount of information the test provides at each level of student ability; it is conceptually parallel to the reliability coefficient in CTT. Equation IV-2 shows the information function of an item, i:

$$I_i(\theta) = \sum_{c=1}^{m} \frac{a_i^2(P_{ic}(\theta)[1-P_{ic}(\theta)]-P_{i(c+1)}(\theta)[1-P_{i(c+1)}(\theta)])^2}{P_{ic}(\theta)-P_{i(c+1)}(\theta)}, \qquad \text{(IV-2)}$$

where $P_{ic}(\theta)$ is the probability of obtaining score c with ability level θ. The TIF at a given ability level indicates the amount of information that is provided by the test at that ability level. As the TIF increases, the accuracy of the corresponding theta estimates also increases. Figure *IV-1*, Figure *IV-2*, Figure IV-3, and Figure *IV-4* present the TIFs for theta values ranging from -3 to 3 in increments of 0.5 for each grade or grade band in the four domains.

Typical TIF values are large at the center of the theta distribution and gradually decrease toward the two ends of the scale, where thetas become very low or very high and result in a bell-shaped pattern. It is particularly important to inspect the TIFs to evaluate the extent to which the test provides sufficient information at the performance-level cuts, theta-level cuts included in Table IV-3. TIF values for KELPA, shown in Figure IV-1, Figure IV-2, Figure IV-3, and Figure IV-4, tend to have larger values at the lower end of the theta scale, which reflects slightly greater measurement precision at the lower end of the distribution in comparison to the higher end of the distribution. The TIF values for kindergarten writing and reading are relatively smaller than the TIF values of other grades or grade bands in the same domain. These results are consistent with the coefficient alpha results, and it is expected these TIF values will increase in the critical regions of the scale as additional items are added to the pool. Also, several KELPA theta-level cuts are at the lower end of the distribution, for example majority of Level 2 and Level 3 cuts and some of Level 4 cuts are lower than 0, the TIFs at the theta-level cuts are large, indicating the tests can provide sufficient information at the performance-level cuts.

Table IV-3: KELPA Theta-Level Cuts by Domain and Grade or Grade Band

Grade or	Listening			Speaking			Reading			Writing		
grade band	L2	L3	L4	L2	L3	L4	L2	L3	L4	L2	L3	L4
K	-1.23	-1.04	0.56	-0.51	0.19	0.87	-0.53	0.36	1.30	-0.94	-0.04	1.82
1	-1.63	-0.84	0.45	-1.41	-0.38	0.58	-0.96	-0.52	0.79	-1.63	-0.78	0.79
2	-2.25	-1.28	-0.54	-1.73	-0.81	0.24	-1.33	-0.68	-0.20	-1.60	-0.73	0.19
3	-1.99	-1.19	-0.42	-1.35	-0.73	0.27	-0.81	-0.40	0.46	-1.05	-0.41	1.01
4	-1.82	-1.52	-0.48	-1.76	-1.05	-0.07	-1.47	-0.69	-0.06	-1.60	-1.04	0.10
5	-1.67	-1.33	-0.36	-1.46	-0.86	0.15	-1.09	-0.27	0.32	-1.55	-0.57	0.54
6	-1.53	-1.33	-0.43	-1.56	-0.86	0.14	-1.66	-0.65	0.39	-1.78	-0.90	0.44
7	-1.50	-1.12	-0.31	-1.48	-0.78	0.20	-1.28	-0.29	0.66	-1.78	-0.48	0.69
8	-1.37	-1.04	-0.11	-1.41	-0.72	0.33	-1.16	-0.09	1.09	-1.74	-0.25	0.90
9–10	-1.39	-1.07	-0.51	-1.04	-0.56	0.04	-0.51	0.01	0.47	-1.27	-0.38	-0.06
11–12	-1.25	-0.94	-0.18	-0.97	-0.53	0.13	-0.36	0.32	0.74	-0.69	-0.15	0.41

Figure IV-1: Test Information Function for Listening

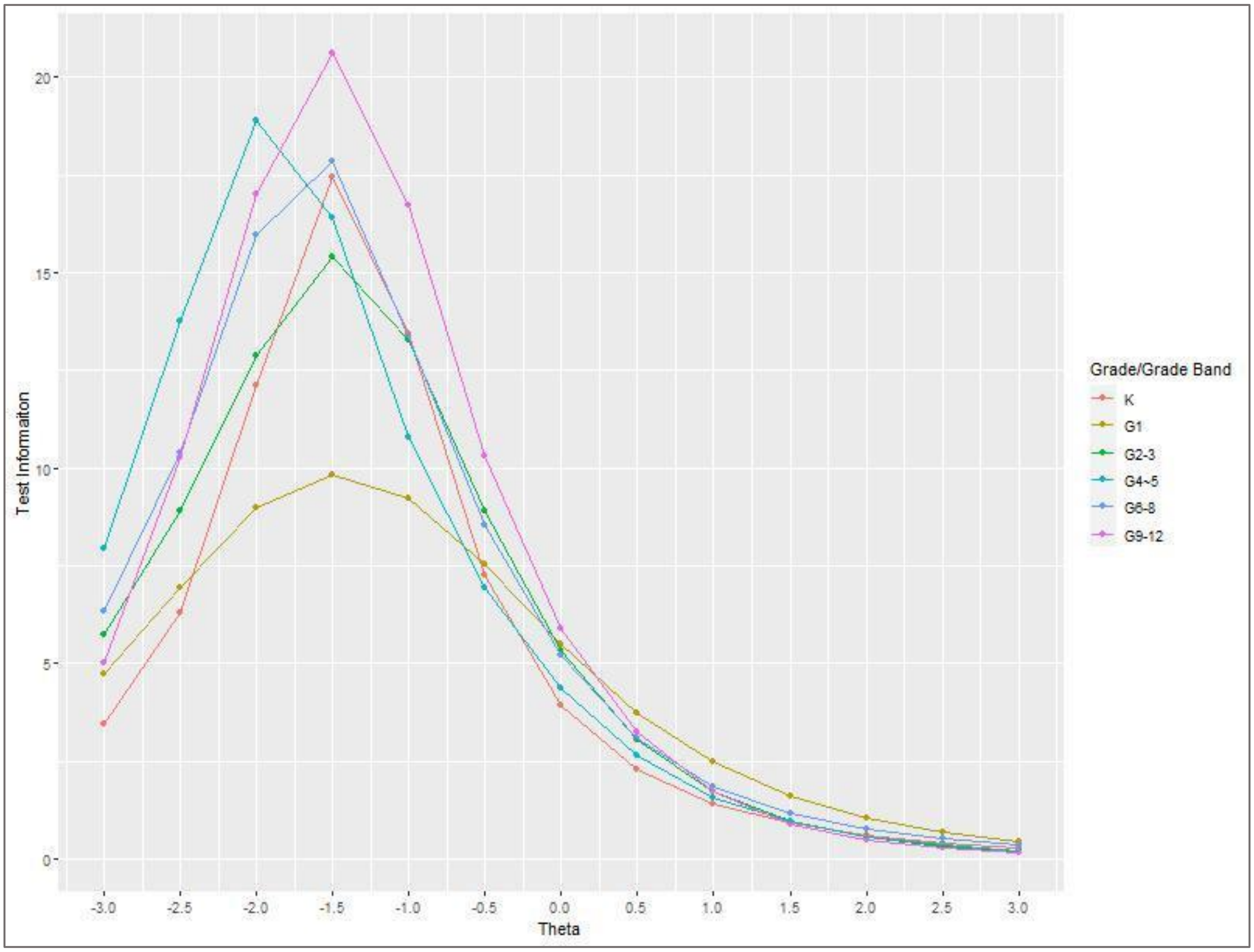


Figure IV-2: Test Information Function for Speaking

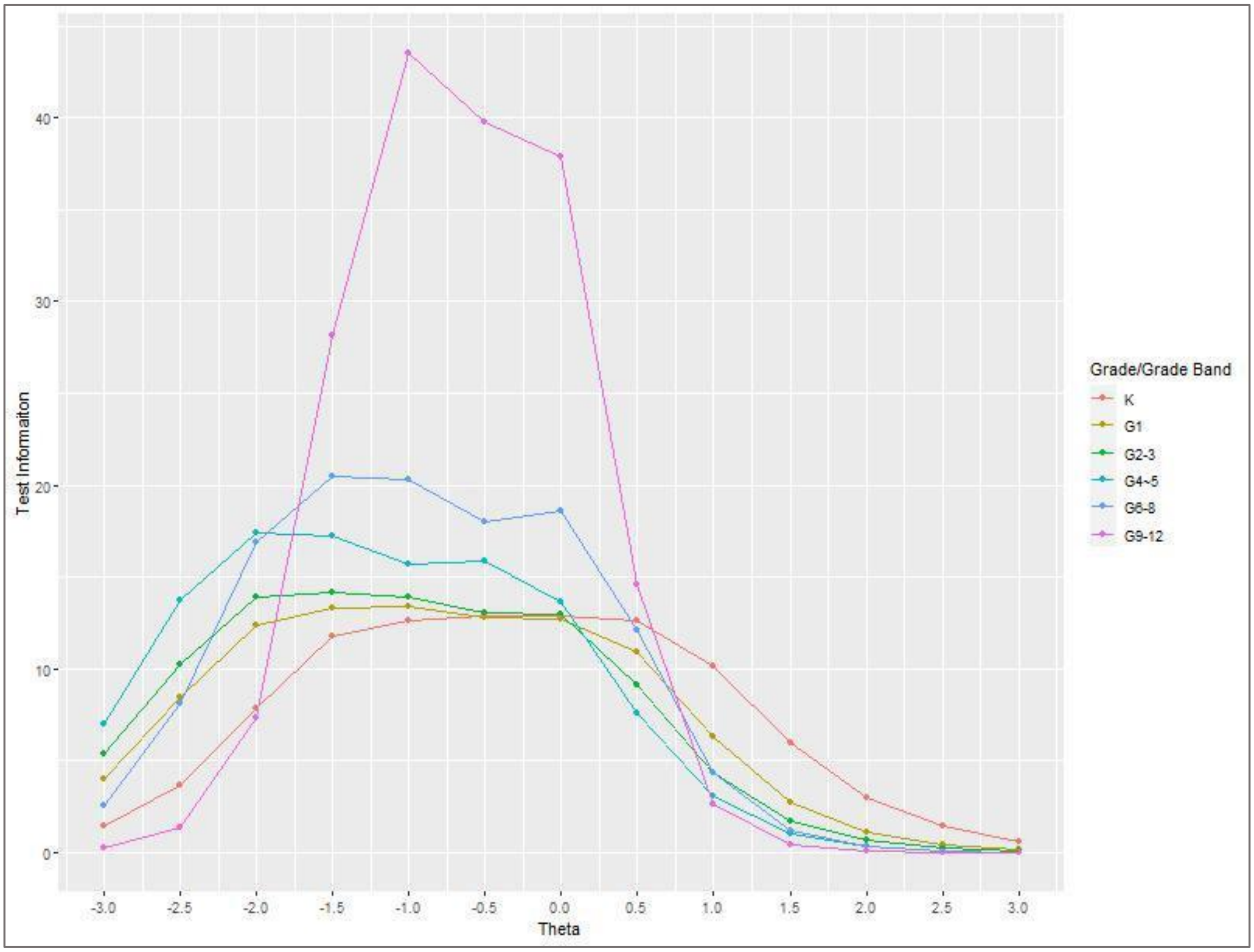


Figure IV-3: Test Information Function for Reading

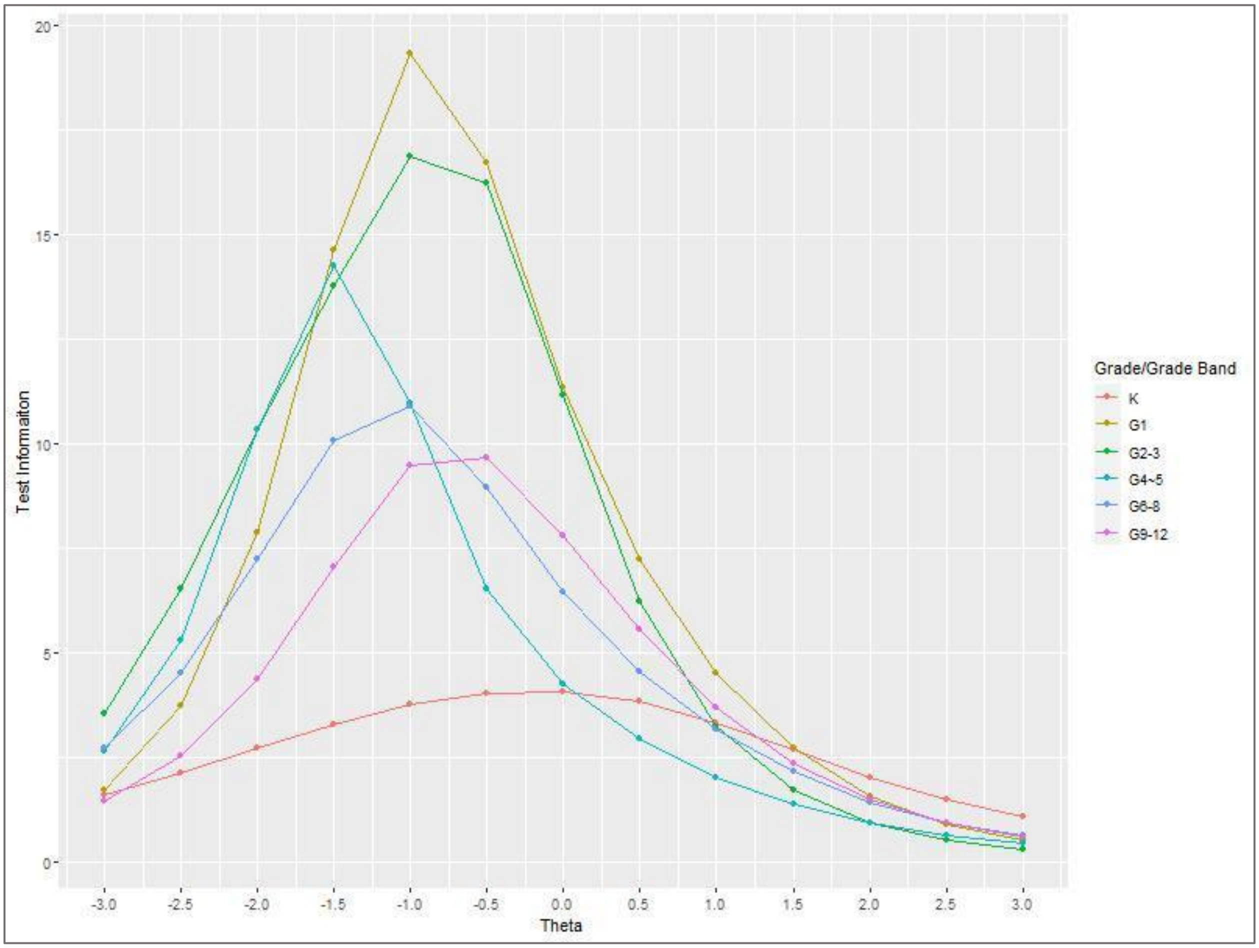
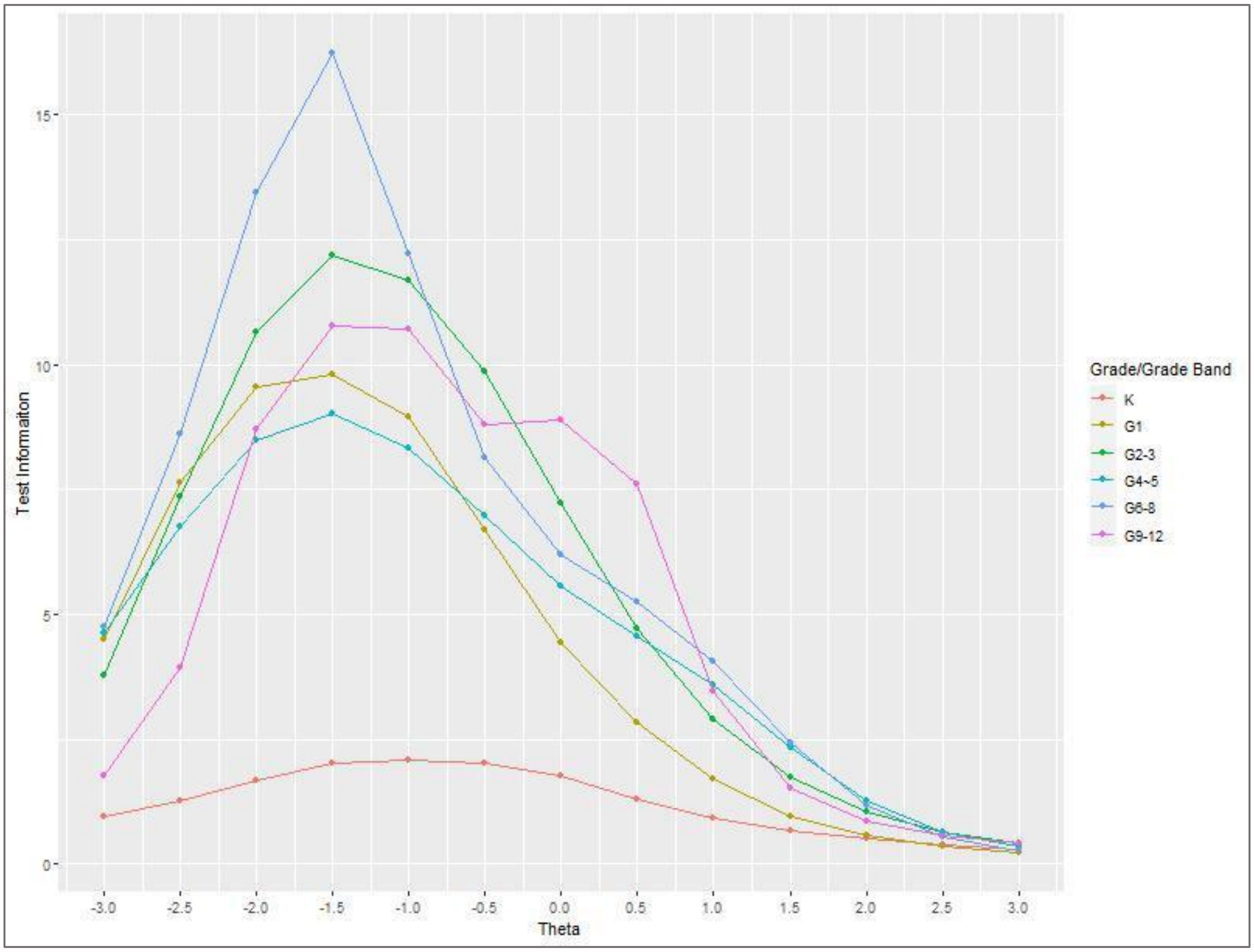


Figure IV-4: Test Information Function for Writing



In IRT, a conditional standard error of measurement (CSEM) can also be estimated for each individual level of theta across the whole performance continuum. CSEMs are computed through their inverse relationship with TIFs. The lower the CSEMs, the more accurate the theta estimates. Figures C-1 through C-4 in Appendix C present the CSEMs for theta values, ranging from −3 to 3 in increments of 0.5 for each grade or grade band in the four domains. Typical CSEM values are small at the center of the theta distribution and gradually increase toward the two ends of the scale, where thetas become very low or very high and result in a U-shaped pattern. As expected, given the TIF results, the CSEM values for KELPA shown in Figures C-1 through C-4 have smaller values at the lower end of the theta scale, which reflects greater measurement precision at the lower end of the distribution compared to the high end of the distribution. Moreover, the CSEM values for kindergarten writing and reading are larger than the CSEM values of other grades or grade bands in the same domain. Also, several KELPA theta-level cuts (presented in Table *IV-3* are at the lower end of the distribution, for example majority of Level 2 and Level 3 cuts and some of Level 4 cuts are lower than 0, the CSEMs at the theta-level cuts are small, indicating the tests can provide sufficient information at the performance-level cuts

IV.1.3 Classification Consistency and Accuracy

When an assessment uses achievement or proficiency levels as the primary method to report test results, accuracy and consistency of classification into different proficiency levels become key indicators of the quality of the assessment. As described by 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) The coefficients for both classification consistency and accuracy range from 0 to 1, with 0 representing classifications that are not consistent or accurate and 1 representing perfectly consistent or accurate classifications.

Because true scores are unobservable and repeated testing is not feasible, a true-score distribution and an observed-score distribution for an alternate parallel forms were estimated using actual observed-score distribution and reliabilities (Livingston & Lewis, 1995). The true-score distribution is used to calculate the classification accuracy, which is the probability of accurate classification between the true-score and actual observed-score distributions. The observed-score distribution for an alternate parallel form is used to calculate the classification consistency, which is the percentage of classification agreement between two observed-score distributions, in other words, the actual and alternate parallel form observed-score distributions. The results for classification consistency and accuracy for three cuts are presented in Table *IV-4*. The classification consistency and accuracy of the Level-4 cut is very important for proficiency classification because students have to be at Level 4 for all four domains to be considered proficient overall. BB-CLASS software (Brennan, 2004) was used to derive the information. Classification consistency of the KELPA domain tests have indices ranging from .66 to .98 across all cuts, grades or grand bands, and domains. Classification accuracy of the KELPA domain tests have indices ranging from .70 to .99 across the majority of cuts, grade levels or bands, and domains. The one exception was for the Level-4 cut of the writing test in grade 7 (.68). For the same grade, classification consistency and accuracy for the speaking test are higher than for the other three domain tests. Not surprisingly, speaking tests have more total score points compared to other domain tests.

Table IV-4: Classification Consistency and Accuracy by Domain and Grade

Domain and grade	Cut-score category					
	1 vs. 2, 3, 4		1, 2 vs. 3, 4		1, 2, 3 vs. 4	
	C	A	C	A	C	A
Listening						
K	.93	.95	.91	.93	.74	.80
1	.95	.96	.89	.92	.79	.85
2	.98	.99	.93	.95	.87	.91
3	.98	.99	.95	.97	.90	.93
4	.97	.98	.96	.97	.88	.92
5	.97	.98	.96	.97	.86	.90
6	.96	.97	.94	.96	.84	.89
7	.96	.97	.94	.96	.86	.90
8	.96	.97	.95	.96	.84	.89
9	.94	.96	.92	.95	.88	.92
10	.95	.97	.94	.96	.90	.93
11	.95	.97	.94	.96	.85	.90
12	.95	.96	.94	.96	.83	.88
Speaking						
K	.92	.94	.88	.91	.81	.85
1	.96	.97	.91	.94	.79	.84
2	.97	.98	.93	.95	.80	.86
3	.97	.98	.94	.96	.79	.86
4	.98	.99	.97	.98	.85	.89
5	.97	.98	.96	.97	.75	.83
6	.97	.98	.95	.97	.84	.89
7	.97	.98	.95	.97	.83	.88
8	.97	.98	.96	.97	.77	.84
9	.96	.98	.95	.97	.90	.93
10	.97	.98	.96	.97	.91	.93
11	.97	.98	.96	.97	.88	.92
12	.97	.98	.96	.97	.84	.89
Reading						
K	.72	.79	.81	.87	.90	.93
1	.90	.93	.88	.91	.89	.92
2	.89	.92	.89	.92	.89	.92
3	.92	.94	.91	.93	.86	.90
4	.93	.95	.85	.90	.80	.86
5	.91	.94	.83	.88	.78	.84
6	.94	.96	.84	.89	.81	.87
7	.92	.94	.85	.89	.78	.84
8	.92	.94	.86	.90	.72	.77
9	.86	.90	.85	.90	.87	.91

Domain and grade	Cut-score category					
	1 vs. 2, 3, 4		1, 2 vs. 3, 4		1, 2, 3 vs. 4	
	C	A	C	A	C	A
10	.87	.91	.85	.89	.85	.89
11	.87	.91	.84	.89	.83	.88
12	.87	.91	.85	.90	.84	.89
Writing						
K	.85	.90	.78	.84	.86	.91
1	.96	.97	.88	.92	.70	.74
2	.95	.96	.88	.92	.76	.83
3	.95	.96	.90	.93	.70	.74
4	.95	.96	.91	.94	.76	.83
5	.96	.97	.89	.92	.69	.77
6	.96	.97	.91	.94	.70	.76
7	.96	.98	.88	.92	.69	.68
8	.97	.98	.86	.91	.66	.70
9	.90	.93	.81	.87	.77	.83
10	.92	.94	.84	.89	.79	.85
11	.88	.92	.82	.88	.72	.80
12	.89	.92	.80	.87	.68	.75

Note. Categories 1, 2, 3, and 4 represent proficiency levels 1, 2, 3, and 4. C = consistency; A = accuracy.

IV.2 Fairness and Accessibility

From test and item development to test administration, Achievement & Assessment Institute (AAI) has taken reasonable and appropriate steps to ensure that KELPA assessments were accessible to all English learners (ELs) and fair across student groups. During item development, item-writing training, and multiple internal and external reviews, numerous checks were conducted to ensure the items were accessible and fair. Moreover, several accommodations were provided during the test to increase the fairness and accessibility of the test content, such as text-to-speech for all instructions.

IV.2.1 Fairness

According to the Standards for Educational and Psychological Testing, “the goal of fairness is to maximize, to the extent possible, the opportunity for test takers to demonstrate their standing on the construct(s) the test is intended to measure.” (APA et al., p. 51). Evidence supporting fairness of KELPA comes from several sources, such as item and test development, differential item functioning (DIF), and student-group performance.

During item development, passage development, and item writing, passage and item writers were trained in universal design (UD) principles and bias-and sensitivity-guidelines. Moreover, no item is used on the operational form for scoring that does not pass the bias-and-sensitivity reviews. Items with construct-irrelevant variance that could prevent students from demonstrating what they know and are able to do were either rejected or revised. Details about passage- and item-development training guidelines and bias-and-sensitivity reviewing criteria are in Chapter II. Assessment System Operations.

As described in Section III.3.3 Differential Item Functioning, DIF examines whether an item shows statistical difference between two groups of students after adjusting for ability. DIF was examined across gender (female vs. male) and ethnicity (Hispanic vs. non-Hispanic) groups. As shown in Section III.3.3 Differential Item Functioning, no items were identified as showing significant DIF in each of the four domains across all grades and grade bands. The absence of items with DIF or a small amount of items with DIF is evidence that AAI's efforts in proactively improving item quality has been effective. Over the years, item statistics have been used to inform and improve item-writer training and guidelines. DIF has been addressed by providing effective item bias-and-sensitivity training and guidance to item writers and item reviewers. Within AAI, the effort has resulted in a decrease in the number of DIF items over time.

The student-group test results presented in Section IV.3.4.3 Student-Group Test Results show that there are some mean-score differences among race, ethnicity, disability, and gender groups for some domain tests. Most mean-score differences fell within one standard deviation. It is important to note that performance differences across student groups does not necessarily indicate test bias. Even when a test is carefully constructed and reviewed with fairness considerations at the forefront, achievement differences or gaps may exist among student groups.

The student group test reliabilities presented in Section IV.1.1.1 Student-Group Reliability show that most test reliabilities among student groups are very similar, the differences between the lowest student-group test reliability and the highest student-group test reliability are smaller than 0.04. The similar student-group test reliabilities provide fairness evidence from a measurement accuracy perspective.

IV.2.2 Accessibility

KSDE uses the CCSSO [Council of Chief State School Officers] Accessibility Manual (Shyyan et al., 2016) to establish accommodation guidelines for all EL students, including ELs with disabilities. Accessibility guidelines were considered and carefully addressed during KELPA item writing. All items on KELPA have passed an accessibility review.

The three-tiered approach to accessibility (Shyyan et al., 2016) is currently employed by KELPA (refer to The Kansas Accessibility Manual: How to Select, Administer and Evaluate Use of Accessibility Supports for Instruction and the Assessment of All Students). This approach includes universal features (i.e., features either embedded and provided digitally through technology or nonembedded and provided locally), designated features (i.e., features for students whose needs have been indicated by educators), and accommodations (changes in procedures or materials that ensure equitable access). Kite® Student Portal (described in Section II.4 Monitoring Test Administration) has many tools available to help students navigate the online testing system. Section V.3 Accessibility Supports for KELPA has more detailed description of different tools. Some of the tools are available to all students, such as the eraser, guide line, and highlighter. These tools are available in the technology-practice tests and embedded or provided digitally through assessment technology. These tools work on laptops and desktops (Windows or Mac) and tablets (Chromebooks or iPads). Other tools are available only to students who have the need identified in their individualized education plans, Section 504 plans,[8] or statement of student needs, such as color contrast, color overlay, reverse contrast, and masking, which allows a student to mask or cover parts of the test (full list of available accommodations for KELPA included in Section V.4 Accommodations. Some tools are not available for KELPA, such as key word translation—Spanish, braille, and text-to-speech for text and test items.

UD was used as a guide during the development of items, test formats, and the online test-delivery interface. UD refers to a design framework and principles that increase access to materials, including assessments, for all students. While initially designed to meet the interests of students with special needs, universally designed assessments provide benefits to all students. Section II.2 Content Development elaborates on item-writing guidelines and training that incorporate UD principles, processes related to bias and sensitivity, and item writers' characteristics.

IV.3 Scoring and Scaling

This section discusses the procedures of scoring individual items, scoring the test as a whole, and scaling. Test-result summaries are also included.

IV.3.1 Item Scoring

Listening, reading, and a portion of writing items are machine scored. All speaking items and some writing items (i.e., constructed-response [CR] items) require local scoring by teachers. Teachers use a scoring rubric to score the speaking and CR items; the rubric ranges from 0 (no evidence of proficiency) to 3 (exhibiting proficiency).

IV.3.1.1 Machine Scoring

Machine-scored items in KELPA include both dichotomously and polytomously scored items. Dichotomously scored KELPA items are multiple-choice items for which there is only one correct answer. For kindergarten and grade 1, multiple-choice items include three answer options; for grades 2 through 12, multiple-choice items include four answer options. Polytomously scored items include multi-select multiple-choice items and technology-enhanced items. Polytomous items usually include multiple elements, and partial credit is assigned to students when they correctly answer one or more elements of a polytomous question. The online test-delivery platform (i.e., Kite) compares student responses to keys stored with the items in the system and assigns scores accordingly.

IV.3.1.2 Educator Scoring

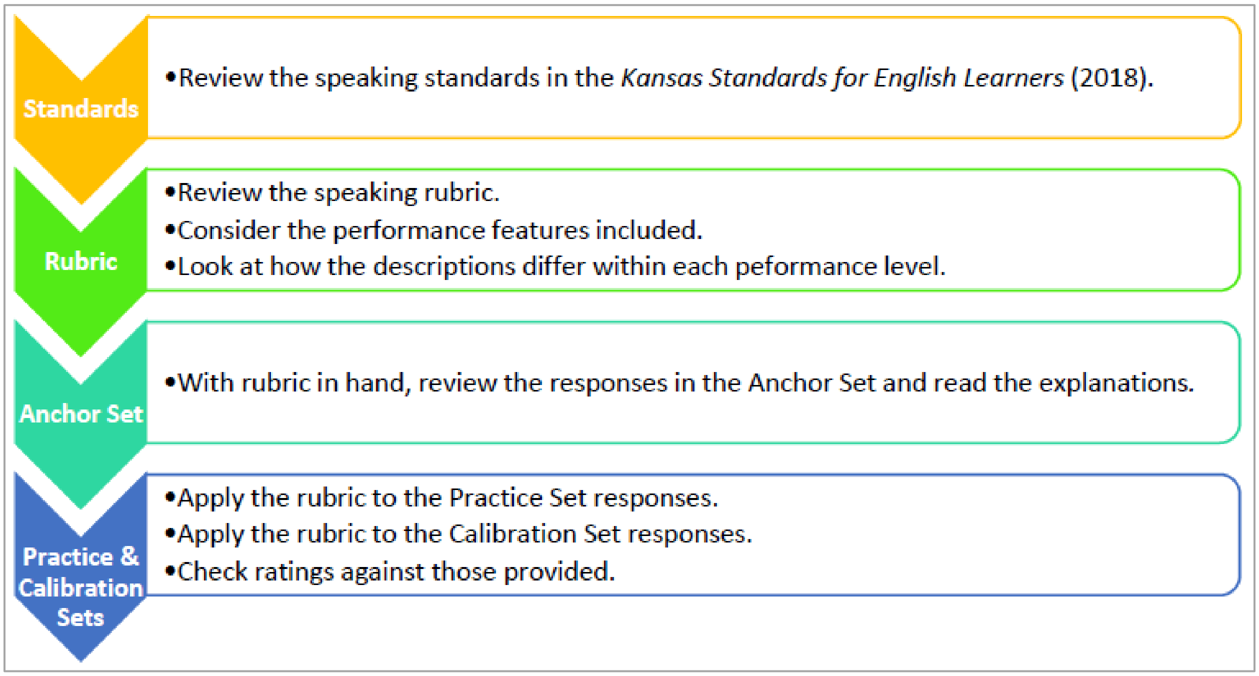
Kansas educators are responsible for scoring the speaking and CR writing items. Educators are required to have teaching licenses, complete annual scoring training before scoring KELPA items, and have accounts in Kite Educator Portal. The main purpose of the training is to allow educators to familiarize themselves with the rubrics associated with speaking and CR writing items so that they can reliably score individual CR responses.

KELPA uses a train-the-trainer model in which district test coordinators (DTCs) are required to be trained yearly in the test-coordinator training workshop held by KSDE. DTCs in turn provide scoring training in

[8] Section 504 is a federal law designed to protect the rights of individuals with disabilities in programs and activities that receive federal financial assistance from the U.S. Department of Education.

their local school districts and maintain records of rosters of educators who completed scorer training. DTCs also maintain records of issues during testing, if any, reported by the trained educators. Secured rater training materials available within the Kite Educator Portal are the designated resources for rater training. The training materials are specific to grade/grade band and domain (speaking and writing). These materials described the five-step process for rater training. Figure IV-5 shows an example training process for speaking. The rater training material include anchor responses (transcripts for writing responses and audio files for speaking responses), practice and calibration set for educators to practice applying the rubrics and calibrate their ratings with other educators in their local training. Educators must follow all guidelines and ethical practices outlined in the training.

Figure IV-5: Process for Constructed-Response Item Scorer Training for Speaking Items



Standards

- Review the speaking standards in the *Kansas Standards for English Learners* (2018).

Rubric

- Review the speaking rubric.
- Consider the performance features included.
- Look at how the descriptions differ within each peformance level.

Anchor Set

- With rubric in hand, review the responses in the Anchor Set and read the explanations.

Practice & Calibration Sets

- Apply the rubric to the Practice Set responses.
- Apply the rubric to the Calibration Set responses.
- Check ratings against those provided.

Note. Figure excerpted from KELPA Rater Training Materials for Speaking: Grade 1.

DTCs assign educators to score speaking and extended-writing items. The educators (i.e., scorers) use rubrics to assign a score to individual responses and enter the score via Educator Portal. For speaking items, Kite captures the responses as audio files to allow for educator scoring. Speaking items can be scored either in the moment that students are responding (i.e., simultaneous scoring) or later by listening to the recordings (i.e., deferred scoring). For both speaking and CR writing items, responses may be scored by individuals, pairs, or small groups of educators. DTCs are encouraged to use paired or group scoring for both speaking and CR writing items. It was highly recommended by KSDE that educators start the scoring activity with a calibration session in which multiple educators score the same set of student responses. Scorers score responses individually, discuss their individually assigned scores, and come to a consensus for a final score that is based on the scoring rubric. After this calibration session is completed, educators continue with individual scoring. Scores can be entered into Kite Ecuador Portal individually by educators or through batch loading by DTCs. For both speaking and CR writing items, one score of record per question must be entered into Kite Educator Portal.

IV.3.2 Test Scoring

Student responses to test items were then calibrated using IRT to derive a single score on each of the domain tests. Four separate unidimensional IRT model were used to score the four domain tests respectively. The decision to implement a unidimensional IRT model for each domain test was informed by the Kansas Technical Advisory Committee in collaboration with KSDE. The IRT item parameters were calibrated before test scoring. Detailed information about the IRT model and calibration procedures is in Section III.3.2 Item Response Theory and Model Assumptions. The IRT scale was established using the item parameters of the final set of retained items from the operational field-test calibration. Item scores were derived from student responses to the items by domain to produce a single score in that domain. Cut scores were then used to categorize students' domain scores into performance levels (see Chapter VI. Academic Achievement Standards and Reporting for information about standard setting) which were then used to assign students an overall proficiency level.

The IRT ability estimates (i.e., thetas) were computed using the two-parameter logistic model and the graded-response model. 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 one-to-one correspondence with raw scores (i.e., each raw score had only one matching theta). Using the test characteristic curve function of the IRT models, the theta for each raw-score point was obtained for a domain test form using the van Wijngaarden–Dekker–Brent root-finding algorithm (Press et al., 1989). The raw scores in each domain were transformed to scale scores, which were used to place students' performance into four levels; however, only the domain performance levels and the overall proficiency level determined by the domain performance levels were reported.

IV.3.3 Scaling

Scaling is the process of transforming thetas or raw scores into scale scores. The purpose of scaling is to facilitate the use and interpretation of test scores. The theoretical values of theta range from negative infinity to positive infinity; thetas can be negative values and have decimal points, which can be difficult to interpret. Therefore, it is beneficial to transform thetas to a scale composed of positive integers to make interpretation and communication of test scores easier.

Although KELPA does not include scale scores on student reports (a sample of student reports included in Section VI.5.1 Student Reports), it is easy to use a KELPA score that is not a negative value to communicate test results (e.g., standard setting, score return files to the state, and summary statistics in technical documentation). The next section addresses the procedures for constructing scale scores.

IV.3.3.1 Scale Transformation

Kolen and Brennan (2004) used the following formula (i.e., Equation IV-3 to derive scaling constants:

$$SS(y) = \frac{\sigma(SS)}{\sigma(Y)} y + \left[SS(y_1) \; - \; \frac{\sigma(SS)}{\sigma(Y)} y_1\right], (5-2), \tag{IV-3}$$

where SS(y) is the scale score, σ(SS) is its standard deviation, σ(Y) is the standard deviation 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 to

$$SS = Ay + C, (5-3), \tag{IV-4}$$

where $A = \frac{\sigma(SS)}{\sigma(Y)}$ and $C = SS(y_1) - \frac{\sigma(SS)}{\sigma(Y)} y_1$. A and C are the slope and intercept, respectively, of the scaling constants. The KELPA scale score has a distribution standard deviation equal to 100, and the distribution mean is equal to 500. The theta has a distribution standard deviation equal to 1, and the distribution mean is equal to 0. The mean of both scales is the anchor point. Thus, σ(Y) is equal to 1, σ(SS) is equal to 100, y_1 is equal to 0, and $SS(y_1)$ is equal to 500. For each domain theta in all grades or grade bands, A is equal to 100 and C is equal to 500. Then the cut scores, described Chapter VI. Academic Achievement Standards and Reporting, are applied to students' scale scores to obtain performance levels for students.

IV.3.3.2 Properties of Scale Scores

The derived scale scores are decimal numbers and must be rounded up to the nearest integer. The IRT model cannot estimate the thetas of extreme raw scores (e.g., 0 and perfect raw scores) because responses to all items are identical. A theta of −99 or 99 is typically assigned to the raw-score points. To keep the scale scores meaningful, the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS) were set to cap scale scores within a reasonable range. LOSS and HOSS are symmetric to the scale-score distribution mean. KELPA's LOSS and HOSS were set at 0 and 1,000, respectively, about five standard deviations from the mean. The choice of LOSS and HOSS makes sure that the majority of raw scores have unique corresponding scale scores instead of the same scale scores (i.e., LOSS or HOSS) for several different raw scores.

IV.3.4 Operational Test Results

The number of students who took KELPA in 2020, along with a summary of their demographic characteristics, is provided in this section. Operational test results present the summary statistics of test scores, which show the distribution of students' test scores. Statistics for test scores by domain for the whole population and different student groups were calculated and are summarized below. Also, the percentages of students in each performance level are included in this section.

IV.3.4.1 Test Enrollment Data

All students who are identified as ELs must take KELPA. For students registered in K–12 schools for the first time in Kansas, a home language survey is used to determine whether a student is a potential EL. A student who is identified by the home language survey as a potential EL is required to take a KSDE-approved EL screener to determine whether KELPA is required. A potential EL student who did not pass the screener will take KELPA in the spring. Students who scored as Proficient on KELPA in 2020 are not required to take KELPA again in the next school year.

In 2020, KELPA was administered in the four domains: listening, speaking, reading, and writing. Students who took the tests were in grades K–12. Table IV-5 shows the number and percentage of enrolled students who were tested in each grade. The students who were tested received a score report. For listening and reading tests, if a student viewed the test, they were classified as taking the domain test even if they did not answer any questions on the test. For speaking and writing tests, if a student viewed the test or teachers scored the student's test, the student is classified as taking the domain test even if they did not answer any machine-scored items on the test. Students who took at least one domain test received a score report. The tested rates for all grades were very high, ranging from 96 (grade 12) to

100% (grades 1–3). In total, 44,592 students were enrolled and 43,962 students were tested; the overall tested rate was 99%[9].

Table IV-5: Number and Percentage of Enrolled and Tested Students by Grade

Grade	Enrolled students (N)	Tested students (n)	Tested students (%)
K	4,597	4,522	98
1	4,641	4,573	99
2	4,751	4,734	100
3	4,070	4,051	100
4	3,819	3,791	99
5	3,240	3,210	99
6	2,820	2,809	100
7	2,660	2,636	99
8	2,749	2,727	99
9	3,116	3,079	99
10	3,120	3,066	98
11	2,824	2,773	98
12	2,185	2,092	96
Total	44,592	44,063	99

For all tested ELs, Table IV-6 shows the percentage of students in each student group by grade[10]. The student groups include race, ethnicity, disability status, and gender. The percentages of students in each student group were very similar across grades except there were more American Indian students in higher grades and fewer White students in higher grades. The majority race group was White, the majority ethnicity group was Hispanic, and there were about equal percentages of male and female students.

[9] The 2020 KELPA testing happened before schools were closed because of COVID-19. Thus, the participation data were not affected.

[10] Economic disadvantaged (ED) status is not shared with ATLAS to protect the privacy of students, so this student group is not included in the comparison.

Table IV-6: Percentage of Tested Students by Demographic Characteristic and Grade

Student group	Grade												
	K	1	2	3	4	5	6	7	8	9	10	11	12
Tested students (n)	4,522	4,573	4,734	4,051	3,791	3,210	2,809	2,636	2,727	3,079	3,066	2,773	2,092
Race													
Black	4.9	4.7	3.9	4.4	4.4	4.2	4.1	5.3	4.6	4.8	5.6	4.2	5.9
American Indian	6.1	7.4	6.5	7.6	8.6	9.3	10.4	11.0	12.9	15.5	17.2	19.9	19.1
Asian	11.2	11.9	11.3	10.1	9.3	7.9	7.7	6.9	7.0	7.2	8.2	9.9	10.6
NHPI	1.2	1.2	1.0	1.1	1.1	1.2	0.8	0.9	1.0	0.7	1.2	0.7	0.7
White	76.5	74.8	77.3	76.9	76.7	77.5	77	75.9	74.5	71.7	67.8	65.3	63.8
Hispanic													
Yes	77.8	78.1	79.5	80.9	81.9	83.1	83.8	84.1	84.8	85.7	83.3	82.7	79.9
No	22.2	21.9	20.5	19.1	18.1	16.9	16.2	15.9	15.2	14.3	16.7	17.3	20.1
SWD													
Yes	9.9	11.8	12.5	13.8	15.7	18.1	19.2	17.3	18.0	14.0	13.0	11.9	12.0
No	90.1	88.2	87.5	86.2	84.3	81.9	80.8	82.7	82.0	86.0	87.0	88.1	88.0
Gender													
Female	47.9	47.9	47.8	45.3	45.7	44.7	43.5	44.7	42.9	41.3	44.5	43.9	43.8
Male	52.1	52.1	52.2	54.7	54.3	55.3	56.5	55.3	57.1	58.7	55.5	56.1	56.2

Note. NHPI = Native Hawaiian and Pacific Islander; SWD = students with disability.

IV.3.4.2 Test Results for All Students

Summaries of scale scores by grade and domain are presented in Table IV-7, Table IV-8, Table IV-9, and Table IV-10. As the tables show, the minimum and maximum values were within the LOSS (i.e., 0) and the HOSS (i.e., 1,000), respectively. Although grades and domains use the same score scale with the same LOSS and HOSS, the assessments are not linked across domains and grades. Thus, the same score has different meanings across domains and grades, and scores across domains and grades should not be compared. In the summary tables below, 10th, 25th, 50th, 75th, and 90th percentiles were provided as P_{10}, P_{25}, P_{50}, P_{75}, and P_{90}, respectively. The differences between (a) P_{50} and P_{25} and (b) P_{75} and P_{50}, respectively, indicate the shape of score distributions: the larger of the two differences indicates the direction of any skewness in the distribution (i.e., a negative skew when the first difference is larger and a positive skew when the second difference is larger). If the two differences match, the distribution is symmetric. For the listening test, the distributions of scale scores were positively skewed in most grades (i.e., grades K, 1, 4, 7, 10–12) and negatively skewed in a few grades (i.e., grades 3, 5, 8); the distributions in grades 2, 6, and 9 were symmetric. For the speaking test, the distributions of scale scores were symmetric or approximately symmetric in grades 1, 3, 6, and 8–10; skewed positively in grades K, 4, 5, and 12; and skewed negatively in grade 7, 11, and 12. For the reading test, the distributions of scale scores in most grades (i.e., grades K, 1, 4, 5, and 8–12) were symmetric or approximately symmetric, and negatively (i.e., grades 3, 6, 7) and positively (i.e., grade 2) skewed in a few grades. For the writing test, the distributions of scale scores were approximately symmetric for most grades (i.e., 2, 4–7, and 10–12); slightly positively skewed in grades 1, 3, and 8; and slightly negatively skewed in grades K and 9.

Table IV-7: Scale-Score Descriptive Statistics for Listening by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	526.69	161.14	0	366	421	492	589	695	1,000
1	513.22	136.17	0	366	431	493	592	648	1,000
2	516.33	175.61	0	353	419	475	541	605	1,000
3	579.95	207.62	0	378	453	541	605	1,000	1,000
4	539.78	186.62	0	362	432	491	611	1,000	1,000
5	560.41	201.34	0	362	432	535	611	1,000	1,000
6	486.72	119.71	0	347	414	478	552	615	1,000
7	525.60	150.53	0	358	432	510	615	725	1,000
8	553.91	164.53	0	358	453	552	615	725	1,000
9	509.31	169.04	0	327	407	477	547	622	1,000
10	538.83	175.67	0	360	437	506	622	622	1,000
11	560.91	188.23	0	371	455	506	622	1,000	1,000
12	546.86	184.03	0	360	437	506	622	1,000	1,000

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

Table IV-8: Scale-Score Descriptive Statistics for Speaking by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	490.69	142.5	0	360	446	514	563	609	1,000
1	512.73	160.91	0	377	447	511	575	638	1,000
2	508.65	167.21	0	364	434	500	550	616	1,000
3	545.35	189.51	0	386	459	515	575	1,000	1,000
4	545.08	207.89	0	377	447	502	577	1,000	1,000
5	550.11	211.42	0	366	447	502	577	1,000	1,000
6	514.14	196.28	0	367	441	503	552	582	1,000
7	539.07	211.42	0	367	453	517	552	1,000	1,000
8	557.79	229.57	0	357	453	517	582	1,000	1,000
9	530.64	248.20	0	345	435	502	556	1,000	1,000
10	548.58	254.52	0	356	448	502	556	1,000	1,000
11	575.76	275.81	0	356	455	522	556	1,000	1,000
12	552.01	299.69	0	0	435	511	556	1,000	1,000

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

Table IV-9: Scale-Score Descriptive Statistics for Reading by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	504.79	138.13	0	363	432	492	552	701	1,000
1	514.89	139.93	0	381	416	492	574	648	1,000
2	492.93	136.11	0	365	392	466	564	671	1,000
3	553.79	162.22	0	379	453	536	603	671	1,000
4	505.25	128.62	0	358	422	491	557	665	1,000
5	527.98	140.56	0	373	442	521	602	665	1,000
6	484.95	118.66	0	355	407	485	541	628	1,000
7	512.26	127.31	0	355	424	511	579	699	1,000
8	538.40	137.07	0	372	443	541	628	699	1,000
9	481.02	115.34	0	359	393	469	542	631	1,000
10	507.30	125.14	0	359	424	502	566	682	1,000
11	529.15	128.37	0	377	439	521	594	682	1,000
12	523.07	136.87	0	377	424	502	594	682	1,000

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

Table IV-10: Scale-Score Descriptive Statistics for Writing by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	526.83	179.80	0	361	407	502	568	874	1,000
1	523.62	156.36	0	367	440	494	588	691	1,000
2	489.54	121.96	0	342	420	482	548	622	1,000
3	537.51	131.71	0	381	465	523	622	687	1,000
4	501.17	127.36	0	351	437	504	563	649	1,000
5	523.73	137.01	0	367	457	532	600	649	1,000
6	492.35	138.49	0	340	428	496	557	596	1,000
7	519.86	152.88	0	353	448	525	596	652	1,000
8	550.05	172.19	0	366	471	525	596	652	1,000
9	469.46	134.23	0	318	407	486	530	585	1,000
10	504.03	140.05	0	337	425	508	585	632	1,000
11	519.34	153.38	0	355	444	508	585	710	1,000
12	507.69	165.57	0	337	444	508	585	710	1,000

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

The proportion of students in each performance level[11] (i.e., Levels 1 through 4) is provided by domain and grade in Figure IV-6, Figure IV-7, Figure IV-8, and Figure IV-9. Students must obtain a Level 4 in each of the four domains to be categorized as proficient overall. The percentage of students in Level 4 ranged from 33% (kindergarten) to 74% (grade 3) across grades for listening, from 20% (kindergarten) to 54% (grade 4) across grades for speaking, from 14% (kindergarten) to 50% (grade 2) across grades for reading, and from 11% (kindergarten) to 53% (grade 1) across grades for writing.

[11] Refer to Section IV.2 Achievement Standard Setting for the KELPA performance level setting process.

Figure IV-6: Performance-Level Results for Listening

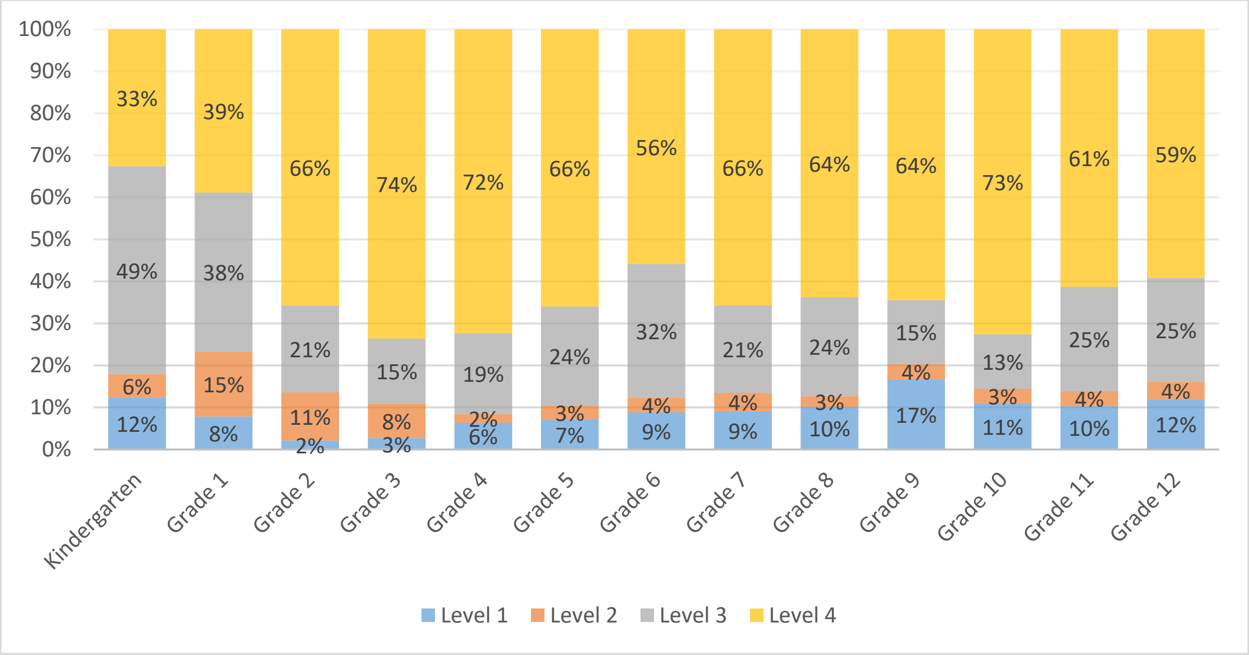


Figure IV-7: Performance-Level Results for Speaking

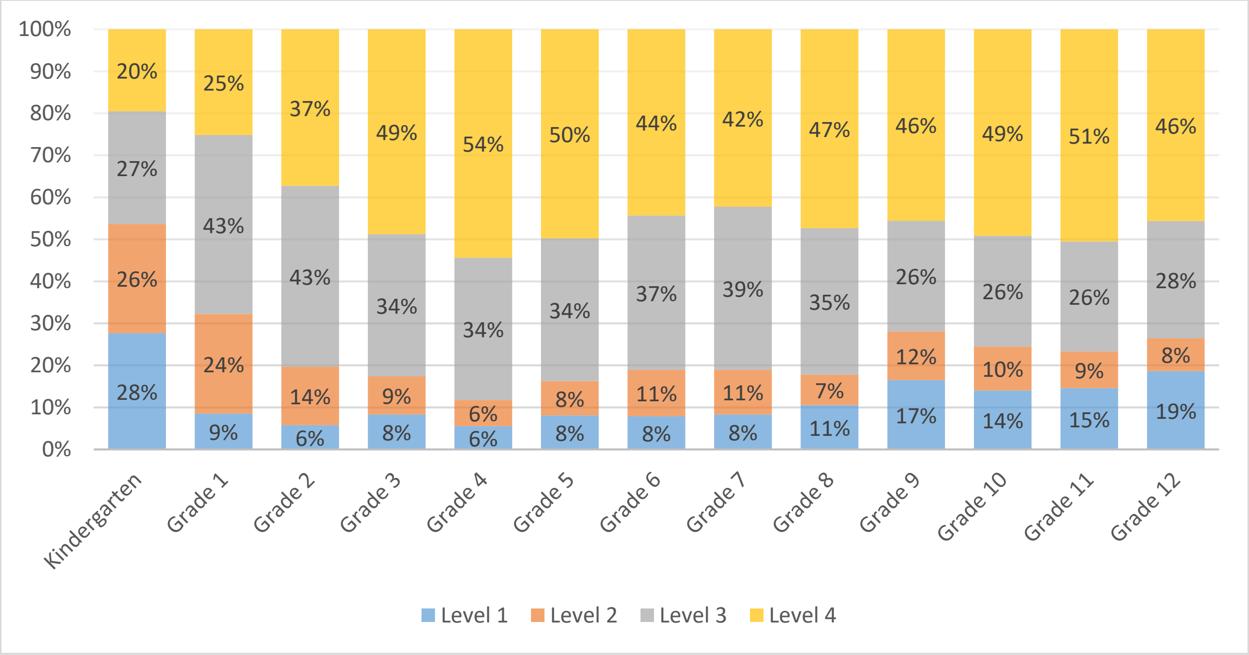


Figure IV-8: Performance-Level Results for Reading

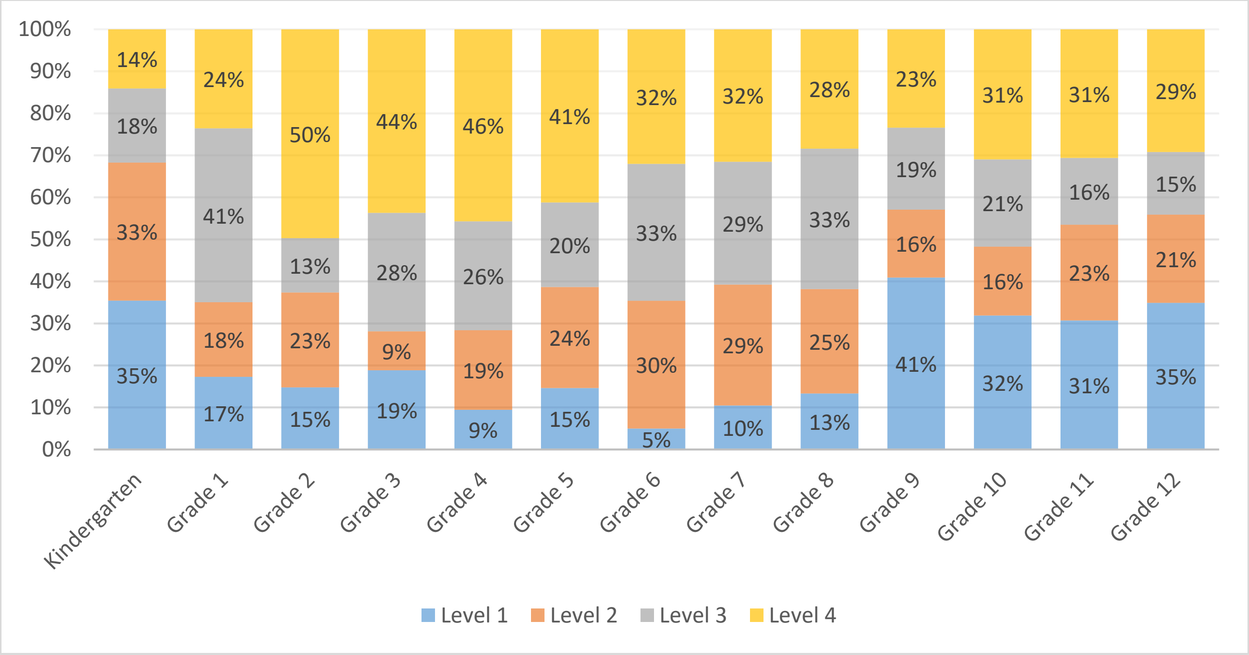
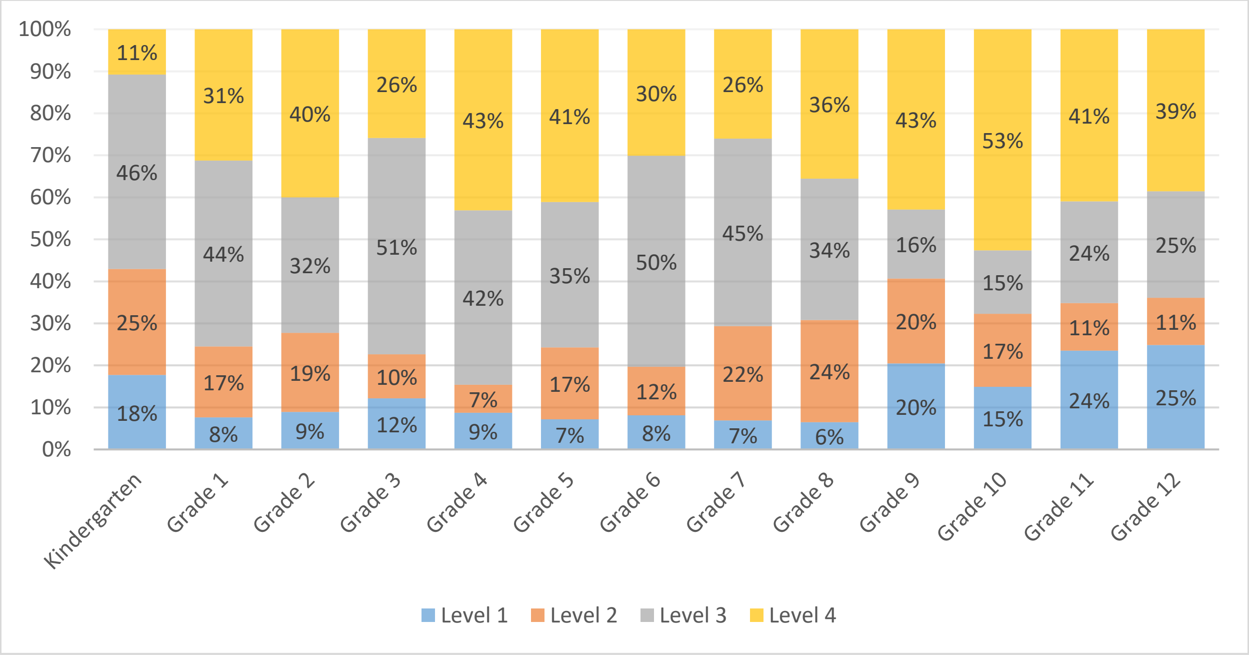


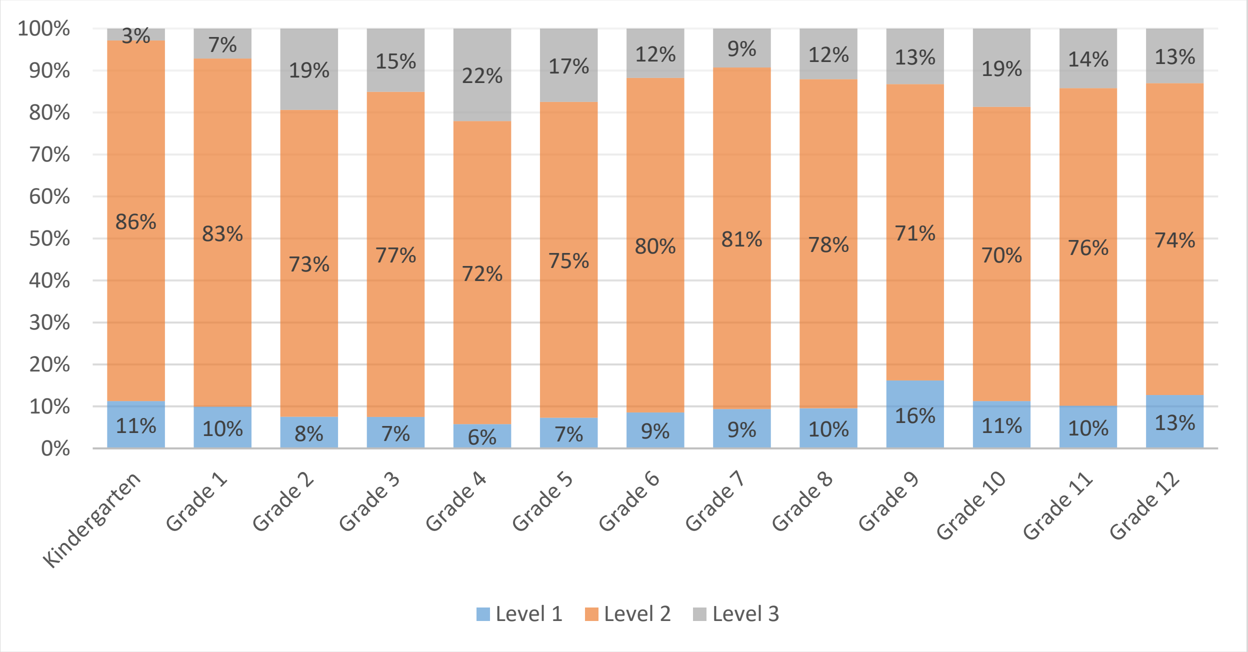
Figure IV-9: Performance-Level Results for Writing



The overall proficiency levels are determined by the patterns of the four domain performance levels. When students are categorized as Level 4 on all four domain tests, the overall proficiency level is Level 3 (i.e., Proficient). When students are at either Level 1 or Level 2 on all four domain tests, the overall proficiency level are Level 1 (i.e., Not Proficient). Students with all other domain performance-level patterns are at Level 2 (i.e., Nearly Proficient). The overall proficiency levels are presented in Figure

IV-10. Results indicate that most students were categorized as Level 2; the percentages ranged from 70% (grade 10) to 86% (kindergarten). Overall, the proficiency rates ranged from 3% (kindergarten) to 22% (grade 4).

Figure IV-10: Overall Performance-Level Results



IV.3.4.3 Student-Group Test Results

Summaries of average scale scores by demographic student groups[12] are presented in Table IV-11, Table IV-12, Table IV-13, and Table IV-14. For student-group sample size, refer to Table IV-6. In most grades and domains, Asian students had the highest mean scores. However, NHPI students had the highest mean score for the grade-8 listening test, American Indian students had the highest mean score for the grade-7 speaking test, and Black students had the highest mean score for the grade-11 speaking test. Across all domains, the mean scores of non-Hispanic students were higher than those of Hispanic students in lower grades (K–3) and were similar in other grades. Across all domains and grades, the mean scores of students without a disability were slightly higher than those of students with a disability. For listening and reading tests, the mean scores of female students were similar to those of male students in all grades. For speaking and writing tests, the mean scores of female students were higher than those of male students in all grades. However, most mean-score differences between student groups fell within the range of 1 standard deviation with one exception: the mean-score difference between Asian and NHPI students on the grade-5 listening test was slightly greater than 1 standard deviation because of the low number of NHPI students.

[12] Economic disadvantaged (ED) status is not shared with ATLAS to protect the privacy of students, so this student group is not included in the comparison.

Table IV-11: Student-Group Scale-Score Descriptive Statistics for Listening, by Grade

Student group	K		1		2		3		4		5		6		7		8		9		10		11		12	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Race																										
AI	508	155	506	143	497	161	560	208	518	184	546	206	467	126	521	153	562	178	507	158	543	172	559	185	531	189
Asian	546	166	533	158	552	193	619	225	547	181	589	212	507	125	539	159	557	155	532	176	558	180	583	188	557	185
Black	508	175	494	138	491	173	550	192	496	164	503	185	464	133	467	127	487	151	487	199	496	184	491	168	490	151
NHPI	515	140	482	125	475	159	560	198	531	200	442	89	443	95	456	167	560	180	507	185	497	175	593	210	543	157
White	526	160	511	132	514	174	577	205	543	188	562	199	489	117	529	150	555	163	507	167	538	173	562	190	555	183
Hispanic																										
Yes	522	158	511	133	510	170	576	205	541	188	561	199	485	118	528	151	555	165	506	166	538	174	561	186	551	187
No	543	170	523	148	540	195	599	218	536	181	560	211	496	128	512	148	546	161	527	185	543	183	563	200	531	170
SWD																										
Yes	489	151	461	122	455	150	506	173	476	156	491	152	447	93	485	124	519	146	464	126	480	126	489	156	477	114
No	531	162	520	136	525	177	592	210	552	189	576	208	496	123	534	154	562	168	517	174	548	180	571	190	557	190
Gender																										
Female	541	161	529	140	525	178	592	213	539	179	555	195	491	119	531	152	559	164	518	172	547	178	569	190	559	180
Male	513	160	499	131	509	173	570	203	540	193	565	206	484	120	521	149	550	165	503	167	532	174	554	187	538	187

Note. AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = students with disabilities.

Table IV-12: Student-Group Scale-Score Descriptive Statistics for Speaking, by Grade

Student group	K		1		2		3		4		5		6		7		8		9		10		11		12	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Race																										
AI	468	163	482	167	481	168	522	196	516	224	504	188	481	193	553	223	541	222	556	248	585	274	580	288	523	307
Asian	511	142	529	188	547	185	556	186	554	203	575	224	515	198	533	217	559	202	529	220	540	245	590	257	574	269
Black	495	155	531	189	498	180	544	201	544	237	507	234	489	204	482	218	536	216	517	275	509	228	591	275	520	270
NHPI	466	145	482	101	488	170	547	211	486	137	541	232	487	202	423	214	536	225	486	165	486	221	539	245	496	292
White	489	140	512	155	508	164	546	189	549	205	555	210	520	196	544	209	563	235	524	249	543	253	572	276	564	303
Hispanic																										
Yes	485	141	507	154	501	162	541	187	544	208	549	208	513	196	545	212	556	232	528	247	551	257	571	278	547	305
No	512	147	533	183	538	185	563	201	549	208	555	228	523	200	510	207	565	213	548	253	539	243	597	264	572	277
SWD																										
Yes	440	145	458	126	453	127	490	148	485	155	505	170	488	191	514	196	537	216	495	200	498	229	529	255	508	292
No	496	141	520	164	517	171	554	194	556	215	560	218	520	197	544	214	563	232	537	255	556	257	582	278	558	300
Gender																										
Female	504	138	534	173	523	173	566	199	571	221	563	220	529	197	552	222	570	233	546	259	568	261	592	273	572	301
Male	479	145	493	147	496	160	528	179	524	194	540	204	503	195	529	202	549	226	520	240	533	248	563	278	536	298

Note. AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = students with disabilities.

Table IV-13: Student-Group Scale-Score Descriptive Statistics for Reading, by Grade

Student group	K		1		2		3		4		5		6		7		8		9		10		11		12	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Race																										
AI	481	127	494	126	483	142	554	169	503	141	513	144	466	120	517	130	539	140	484	112	509	124	539	128	511	145
Asian	583	171	580	164	551	150	586	170	540	138	569	170	513	133	533	123	566	140	486	122	507	127	545	136	517	124
Black	518	154	525	142	475	129	532	161	471	127	498	143	447	117	476	128	473	119	454	125	457	119	467	119	461	118
NHPI	482	132	506	117	481	139	530	169	497	149	468	102	481	131	484	130	577	158	483	74	460	140	523	100	521	107
White	494	129	506	135	486	132	550	160	504	126	527	135	487	116	512	127	539	136	479	113	510	123	526	126	534	138
Hispanic																										
Yes	490	125	503	131	484	131	550	161	503	128	526	136	483	116	512	127	538	137	480	114	507	122	531	126	528	139
No	558	166	558	160	529	149	570	167	517	132	540	162	495	130	512	129	539	140	485	123	509	139	522	138	506	127
SWD																										
Yes	472	121	454	110	429	104	477	133	439	109	459	104	429	97	453	107	491	118	437	94	456	98	458	101	472	114
No	508	139	523	142	502	138	566	163	518	128	543	143	498	120	525	128	549	139	488	117	515	127	539	129	530	138
Gender																										
Female	513	137	520	139	500	139	559	161	509	127	524	136	488	114	516	128	540	129	481	111	506	117	528	121	521	119
Male	497	139	510	141	487	133	550	164	502	130	531	144	482	122	510	127	537	143	481	118	509	131	530	134	524	149

Note. AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = students with disabilities.

Table IV-14: Student-Group Scale-Score Descriptive Statistics for Writing, by Grade

Student group	K		1		2		3		4		5		6		7		8		9		10		11		12	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Race																										
AI	504	162	488	145	466	124	532	150	485	135	498	131	480	150	529	163	551	176	479	131	505	137	518	129	498	159
Asian	611	208	583	192	558	136	583	142	546	140	590	160	524	150	539	175	586	180	497	137	523	132	552	165	535	146
Black	526	210	516	155	480	124	530	139	478	137	477	138	467	163	473	156	471	139	404	147	450	133	478	129	455	139
NHPI	528	177	512	111	492	127	545	164	501	147	481	124	457	142	450	123	587	148	493	78	449	131	472	173	438	248
White	517	172	518	151	482	116	532	126	498	123	523	133	491	131	521	149	551	172	467	133	505	142	517	159	512	172
Hispanic																										
Yes	511	169	513	148	478	114	530	128	497	125	521	133	489	136	521	150	549	171	469	133	504	140	517	152	505	168
No	584	204	562	180	535	139	568	144	522	137	538	154	511	150	512	168	556	179	475	140	507	142	533	159	518	158
SWD																										
Yes	469	163	451	139	424	111	462	117	433	113	460	110	432	113	464	117	498	132	435	104	446	113	452	115	458	131
No	533	181	533	156	499	121	550	130	514	126	538	138	507	140	532	157	561	178	475	138	513	142	529	156	514	169
Gender																										
Female	541	179	533	157	498	123	545	133	516	133	539	140	513	146	543	164	572	185	495	136	528	141	544	155	535	168
Male	514	180	515	155	482	120	531	131	489	121	511	134	476	130	502	140	534	160	452	130	485	136	500	149	486	161

Note. AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = students with disabilities.

IV.3.5 Quality-Control Checks

Multiple quality-control processes occurred during the scoring and reporting of KELPA results. First, student testing data were monitored daily by Agile Technology Solutions (ATS) during the testing window to endure no enrollment error, such as the same student enrolled in two grades, or abnormal testing behaviors, such as testing in the evening. Second, the psychometric staff checked student-response data every three weeks during the testing window to ensure there were no machine scoring errors or duplicates, such as two different scores for the same item for one student. Third, classical item analysis was conducted by the psychometric staff during the testing window to ensure items were functioning as expected. The expectations of items for this steps include:

- Items do not have extremely low or high *p* values.
- Item-total correlations are not negative.
- Distractors for multiple-choice items do not have high correlation with total scores.

Fourth, the psychometric staff compared the IRT calibration results with the CTT statistics to confirm the accuracy of the calibration results. The IRT item-difficulty and discrimination parameters were plotted against CTT item-difficulty and discrimination parameters, respectively, by grade or grade band and domain. Both types of scatter plots indicated a strong relationship between IRT and CTT statistics, which provides evidence for the accuracy of the calibration procedures. Fifth, after the psychometric staff generated the scoring table, that is, the raw score to scale score (RSSS) conversion table. Then the psychometric team checked the reasonableness of the RSSS conversions. The RSSS conversion table is considered reasonable by meeting the following criteria.

- All domains and grade levels are represented.
- All tests are represented.
- All possible integer raw scores are represented for each test.
- No integer is missing from the raw scores, from 0 to the maximum test score.
- The scale score increases with the raw score within each test.

Sixth, the cut scores used to classify students were independently checked to ensure they were consistent with the cut scores approved by the Kansas State Board of Education by two psychometric staff. Last, the psychometric and ATS technology teams independently calculated each individual’s total score, scale score, and domain performance levels and then compared their calculations to identify any differences or calculation errors. When scoring results from both teams were equivalent, students’ score reports were generated. This quality-control process ensured the scoring results provided on students’ reports were complete and accurate.

IV.4 Full Performance Continuum

KELPA was developed with the goal of providing a reasonably precise estimation of students’ English language proficiency across the full performance continuum. The evidence from TIFs and CSEMs in Section IV.1 Reliability indicates KELPA domain tests could precisely estimate proficiency across the full ability scale, but with slightly less information provided at the two ends of the scale, as is commonly seen on large-scale assessments.

In an effort to include items on the test that covered a wide range of difficulties, there was no constraint on item *p* values or mean scores for inclusion on the final test forms. Item quality was screened through item-total correlation and distractor analyses and were confirmed by content-development staff to be free of content flaws. To confirm that the tests effectively covered the full performance continuum as expected, classical and IRT item statistics are presented in this section as evidence.

As stated in Section II.2.3.4 Data Review, 2020 KELPA assessments were operational field tests. All items reported in this section are operational items used in scoring.

IV.4.1 Classical Test Theory Item Statistics

Two CTT statistics, item difficulty and item discrimination, were calculated and provided. Item difficulty refers to how easy or difficult an item is, and item discrimination indicates the degree to which an item differentiates between students with high proficiency and students with low proficiency. Item difficulty in CTT is expressed as a *p* value or mean score. A *p* value is the percentage of students who answered the item correctly. Equation IV-5 shows the calculation of the *p* value.

$$\mathrm{p}\ value = \frac{\sum_{i=1}^{n}\ x_i}{n}, \qquad \text{(IV-5)}$$

where x refers to the observed score, i refers to student i, and n refers to the total number of students who responded to the item. For any item whose full score point is greater than 1, its *p* value is divided by the item maximum score to get an adjusted *p* value ranging from 0 to 1. The higher the *p* value, the easier the item. Table IV-15, Table IV-16, Table IV-17, and Table IV-18 summarize the CTT item difficulties of the retained items after data review, for the four domains across grades or grade bands. Across grades or grade bands and domains, the median *p* values ranged from .47 to .89 indicating that, in general, items appeared to be of medium difficulty to relatively easy for most students. The minimum and maximum *p* values ranged from .21 to .97 across grades or grade bands and domains, suggesting a decent range of possible *p* values across the entire item pool. However, some grades or grade bands had narrower ranges of item difficulties within a domain-specific test (e.g., item *p* values in speaking in grade band 4–5 ranged from .73 to .86). Across domains, listening items tended to be easier compared to items in the other three domains, indicated by the larger mean item difficulty. This finding may be expected given that listening skills tend to be acquired before speaking, reading, and writing skills.

Table IV-15: Summary Statistics for Classical Test Theory Item Difficulties for Listening

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	23	.78	.13	.45	.73	.80	.88	.91
1	25	.73	.17	.33	.58	.81	.86	.93
2–3	25	.82	.10	.60	.74	.83	.87	.97
4–5	25	.84	.12	.53	.81	.89	.93	.96
6–8	25	.78	.16	.33	.74	.83	.90	.94
9–12	24	.80	.12	.43	.76	.81	.89	.94

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

Table IV-16: Summary Statistics for Classical Test Theory Item Difficulties for Speaking

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	10	.59	.06	.43	.58	.60	.63	.64
1	10	.71	.04	.62	.71	.72	.73	.78
2–3	10	.76	.06	.61	.74	.77	.78	.84
4–5	10	.81	.04	.73	.78	.81	.83	.86
6–8	9	.75	.03	.69	.74	.75	.78	.80
9–12	10	.73	.02	.70	.72	.73	.74	.78

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

Table IV-17: Summary Statistics for Classical Test Theory Item Difficulties for Reading

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	19	.49	.15	.21	.35	.47	.63	.71
1	25	.64	.18	.30	.56	.69	.78	.86
2–3	24	.69	.16	.46	.56	.70	.79	.96
4–5	22	.70	.17	.38	.55	.72	.83	.93
6–8	21	.65	.17	.28	.57	.66	.77	.88
9–12	23	.61	.14	.31	.51	.64	.71	.79

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

Table IV-18: Summary Statistics for Classical Test Theory Item Difficulties for Writing

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	8	.55	.25	.22	.36	.56	.71	.55
1	13	.73	.21	.41	.54	.82	.91	.73
2–3	19	.66	.18	.31	.56	.70	.80	.93
4–5	17	.72	.16	.44	.58	.75	.84	.91
6–8	18	.79	.12	.58	.69	.79	.89	.94
9–12	17	.64	.15	.27	.55	.68	.72	.88

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

Item discrimination reflects an item’s ability to differentiate students of high proficiency from those of low proficiency. The Pearson product-moment correlation coefficient between student item scores and domain-test total scores (excluding the studied item score) is used to calculate CTT item discrimination, which ranges from −1.0 to 1.0. Positive values indicate that students with higher proficiency levels are more likely to answer an item correctly than are those with lower proficiency levels; negative values indicate the opposite. The magnitude of the value indicates the degree of discrimination: items with higher values have better discrimination power. CTT item discrimination does not provide information on measuring the full performance continuum directly, but a test in which most items have high item discrimination will provide more-accurate measures of proficiency than a test with few discriminating items. Table IV-19, Table IV-20, Table IV-21, and Table IV-22 summarize the CTT item discrimination of the retained items after data review for the four domains across grades or grade bands. Across grades or grade bands and domains, the median item-discrimination values ranged from .27 to .84 indicating that, in general, items appeared able to discriminate well for most tests. The minimum and maximum item discrimination ranged from .11 to .87 across grades or grade bands and domains, suggesting some tests had items with very high item discrimination. However, some grades or grade bands had some items with low item discriminations (e.g., items on the kindergarten writing and reading tests). As more items are added to the kindergarten writing test, an increase in items with higher item discrimination is expected. Across domains, speaking items tended to have higher item-discrimination values compared to items in the other three domains, indicated by the larger mean item-discrimination. This finding may be expected given that all items on the speaking tests are polytomously scored items, which tend to differentiate students' ability levels better than the dichotomous items.

Table IV-19: Summary Statistics for Classical Test Theory Item Discrimination for Listening

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	23	.41	.10	.24	.34	.41	.48	.57
1	25	.40	.07	.28	.36	.41	.45	.55
2–3	25	.45	.09	.29	.40	.44	.53	.60
4–5	25	.47	.11	.20	.41	.51	.55	.63
6–8	25	.46	.12	.20	.36	.47	.56	.66
9–12	24	.49	.11	.28	.44	.51	.58	.67

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

Table IV-20: Summary Statistics for Classical Test Theory Item Discrimination for Speaking

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	10	.68	.03	.63	.65	.68	.71	.73
1	10	.69	.04	.58	.68	.70	.70	.72
2–3	10	.69	.05	.59	.66	.70	.72	.74
4–5	10	.73	.02	.71	.71	.73	.75	.77
6–8	9	.78	.01	.76	.77	.78	.78	.80
9–12	10	.84	.02	.79	.83	.84	.86	.87

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

Table IV-21: Summary Statistics for Classical Test Theory Item Discrimination for Reading

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	19	.30	.08	.14	.23	.30	.37	.42
1	25	.48	.10	.20	.45	.50	.54	.60
2–3	24	.50	.11	.32	.41	.51	.59	.66
4–5	22	.39	.09	.21	.31	.39	.45	.57
6–8	21	.41	.12	.17	.33	.42	.51	.58
9–12	23	.43	.09	.21	.36	.45	.49	.57

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

Table IV-22: Summary Statistics for Classical Test Theory Item Discrimination for Writing

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	8	.30	.15	.11	.19	.27	.39	.54
1	13	.43	.17	.21	.29	.45	.62	.64
2–3	19	.46	.15	.16	.35	.50	.56	.67
4–5	17	.44	.14	.15	.34	.43	.54	.64
6–8	18	.50	.12	.30	.41	.51	.58	.68
9–12	17	.39	.16	.15	.28	.33	.45	.68

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

IV.4.2 Item Response Theory Item Statistics

The two-parameter logistic IRT model and its polytomous counterpart, the graded-response model, were used to fit data. For these two models, both item-difficulty and discrimination parameters were freely estimated. IRT item-difficulty parameters ranged from negative infinity to positive infinity. Different from CTT item-difficulty parameters, the higher the IRT item-difficulty parameter, the harder the item is. The number of IRT difficulty parameters of a polytomous item is equal to the score points it has minus 1 (i.e., excluding the score point 0). Most KELPA items are dichotomous with one *b* parameter, but some are polytomous items with as many as 10 score categories (thus, nine *b* parameters); therefore, the number of *b* parameters can be different from the number of items. Table IV-23, Table IV-24, Table IV-25, and Table IV-26 summarize the IRT difficulty (i.e., *b* parameter) for the four domains across grades or grade bands of the retained items after data review. Across grades or grade bands and domains, the median *b* parameters range from −1.91 to −0.26 indicating that, in general, items appeared to be of medium difficulty to relatively easy for most students. The minimum and maximum *b* parameters ranged from −9.13 to 3.51 across grade or grade bands and domains, suggesting a decent range of possible *b* parameters across the entire item pool. However, some grades or grade bands had narrower ranges of item difficulties within a domain-specific test (e.g., item *b* parameters in speaking in grade band 9–12 ranged from −1.53 to 0.16). As with CTT item difficulties across domains, listening items tended to be easier compared to items in the other three domains, indicated by the smaller mean *b* parameter. Because during item calibration the theta scale is fixed and item scale is centered on the theta scale without any constrains or priors, some IRT item difficulties can be very low especially in the listening domains. For example, a *b* parameter of −9.13 is for the lowest possible score point of a grade band 4–5 listening item with a *p* value of .96.

Table IV-23: Summary Statistics for Item Response Theory Item Difficulties for Listening

Grade or grade band	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	31	-1.49	1.01	-3.66	-1.79	-1.48	-1.05	1.16
1	33	-1.49	1.40	-6.39	-2.01	-1.41	-0.54	1.01
2–3	36	-1.90	1.31	-6.02	-2.35	-1.50	-1.08	0.10
4–5	36	-2.12	1.80	-9.13	-2.31	-1.91	-1.12	0.10
6–8	34	-1.48	1.20	-5.36	-1.93	-1.47	-1.07	1.31
9–12	30	-1.39	0.63	-2.42	-1.85	-1.37	-1.11	0.44

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

Table IV-24: Summary Statistics for Item Response Theory Item Difficulties for Speaking

Grade or grade band	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	30	-0.30	0.94	-1.60	-1.25	-0.26	0.49	1.92
1	30	-0.86	0.93	-2.14	-1.77	-0.91	0.15	0.53
2–3	30	-1.10	0.94	-2.64	-1.98	-1.18	-0.19	0.66
4–5	30	-1.29	0.87	-2.61	-2.07	-1.43	-0.38	0.24
6–8	27	-0.94	0.79	-2.04	-1.64	-1.05	-0.10	0.27
9–12	30	-0.69	0.51	-1.53	-1.17	-0.73	-0.14	0.16

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

Table IV-25: Summary Statistics for Item Response Theory Item Difficulties for Reading

Grade or grade band	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	32	-0.40	1.24	-3.00	-1.16	-0.52	0.46	2.61
1	30	-0.66	0.81	-2.20	-1.18	-0.79	-0.27	1.20
2–3	29	-0.99	0.89	-3.47	-1.63	-0.86	-0.41	0.17
4–5	23	-0.90	0.83	-2.26	-1.46	-0.95	-0.40	0.82
6–8	31	-0.78	1.79	-6.82	-1.48	-0.92	-0.12	3.51
9–12	25	-0.63	1.01	-4.08	-0.85	-0.63	-0.20	1.10

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

Table IV-26: Summary Statistics for Item Response Theory Item Difficulties for Writing

Grade or grade band	No. of *b* parameters	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	14	-0.71	1.69	-3.08	-2.10	-0.79	0.57	2.51
1	27	-1.39	1.07	-3.48	-2.14	-1.42	-0.72	0.59
2–3	34	-0.86	1.00	-2.29	-1.52	-1.00	-0.38	1.92
4–5	28	-1.16	1.12	-3.20	-1.95	-1.21	-0.40	0.96
6–8	28	-1.29	1.08	-3.57	-1.87	-1.45	-0.72	0.98
9–12	27	-0.89	1.10	-3.39	-1.57	-0.97	-0.17	2.08

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

The IRT item-discrimination parameter reflects an item’s ability to differentiate students of high ability from those of low ability. Usually IRT item-discrimination parameters are positive. Items with higher values have better discrimination power, and one item has only one discrimination parameter. The IRT item-discrimination parameter does not provide information on measuring the full performance continuum directly, but a test in which most items have high IRT item-discrimination parameters will provide more-accurate measures of ability than a test with few discriminating items. Table IV-27, Table IV-28, Table IV-29, and Table IV-30 summarize the IRT item-discrimination values for the four domains across grades or grade bands of the retained items after data review. Across grades or grade bands and domains, the median a parameters ranged from 0.78 to 3.83, indicating that, in general, items appeared able to discriminate well for most tests. The minimum and maximum a parameters ranged from 0.23 to 4.54 across grade or grade bands and domains, suggesting some tests had items with very high item discrimination. However, consistent with the CTT results, some grades or grade bands had some items with low item discriminations (e.g., items on the kindergarten writing and reading tests). As more items are added to the pool, particularly in kindergarten reading and writing domains, an increase in items with high a parameters is expected. Similar to the CTT item-discrimination results across domains, speaking items tended to have higher item discrimination compared to items in the other three domains, indicated by the larger mean a parameter.

Table IV-27: Summary Statistics for Item Response Theory Item Discrimination for Listening

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	23	1.57	0.82	0.57	0.97	1.31	2.20	3.32
1	25	1.28	0.38	0.73	0.90	1.30	1.49	1.94
2–3	25	1.56	0.55	0.88	1.14	1.36	1.84	2.66
4–5	25	1.68	0.65	0.60	1.12	1.75	1.95	3.31
6–8	25	1.59	0.67	0.59	0.89	1.70	2.05	2.93
9–12	24	1.79	0.67	0.77	1.28	1.87	2.24	2.97

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

Table IV-28: Summary Statistics for Item Response Theory Item Discrimination for Speaking

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	10	2.14	0.22	1.87	1.94	2.17	2.28	2.47
1	10	2.17	0.25	1.55	2.11	2.22	2.30	2.42
2–3	10	2.22	0.38	1.59	1.97	2.22	2.52	2.76
4–5	10	2.47	0.21	2.21	2.32	2.44	2.60	2.85
6–8	9	2.85	0.14	2.67	2.78	2.82	2.83	3.13
9–12	10	3.85	0.45	2.95	3.61	3.83	4.19	4.54

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

Table IV-29: Summary Statistics for Item Response Theory Item Discrimination for Reading

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	19	0.86	0.36	0.31	0.59	0.78	1.13	1.56
1	25	1.77	0.66	0.51	1.30	1.69	2.28	3.32
2–3	24	1.79	0.55	0.86	1.44	1.84	2.16	2.66
4–5	22	1.48	0.87	0.52	0.86	1.21	1.52	3.70
6–8	21	1.37	0.64	0.36	0.98	1.30	1.81	2.68
9–12	23	1.30	0.44	0.53	0.94	1.28	1.59	2.33

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

Table IV-30: Summary Statistics for Item Response Theory Item Discrimination for Writing

Grade or grade band	No. of items	*M*	*SD*	Min	P_{25}	Median	P_{75}	Max
K	8	0.89	0.58	0.23	0.50	0.78	1.10	2.03
1	13	1.55	0.71	0.56	1.05	1.59	2.02	2.87
2–3	19	1.46	0.65	0.44	0.98	1.56	1.81	2.89
4–5	17	1.36	0.55	0.36	0.89	1.46	1.82	2.17
6–8	18	1.79	0.79	0.78	1.08	1.83	2.33	3.79
9–12	17	1.24	1.00	0.37	0.58	0.78	1.17	3.45

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

IV.5 Continuous Program Improvement

This section summarizes the ongoing improvements for KELPA. An independent alignment study to establish the alignment between KELPA items and the 2018 Standards and the correspondence between the 2018 Standards and Kansas content standards in English language arts, mathematics, and science is planned in spring of 2021. A rater agreement study to collect information on the reliability of educator-scored CR items is also planned in spring of 2021 using data from the 2020–2021 KELPA administration. With the rater agreement study, the Kite Educator Portal user interface will be updated to collect information for CR items (speaking and writing) scoring method. An optional post-administration teacher survey will be given to educators to document their experience in administering and scoring KELPA.

V. Inclusion of All Students

This chapter begins with a general introduction to the accessibility framework in Kansas assessments, with a focus on KELPA. It then elaborates on the guidelines and procedures for selecting accommodations on KELPA and ends with a summary of the frequency of use of accommodations on KELPA.

V.1 Inclusion of All English Learners in KELPA

As described in Section I.3. Intended Population, all students who are identified as English learners (ELs) must take KELPA, according to the requirement by the Elementary and Secondary Education Act Title I. ELs who have significant cognitive disabilities also participate in KELPA. Accessibility tools and accommodations are available either within or outside the Kite® system. The inclusion of students with disabilities in KELPA is achieved by providing guidelines for educators to register their students with different needs through a Personal Needs Profile (PNP) in Kite Educator Portal. The KELPA Examiner's Manual describes step-by-step registration procedures for students who need accommodations.

V.2 Kansas Accessibility Framework

As discussed in The Kansas Accessibility Manual, a three-tier accessibility framework is applied in Kansas state assessments, which provides for:

- accessibility supports, including both embedded (i.e., digitally provided) and non-embedded (i.e., non-digitally or locally provided) universal features that are available to all students
- designated features that are available for students whose need has been identified by an informed educator or team of educators
- accommodations that are available for students and documented on an Individualized Education Program (IEP), Section 504 plan, or individual learning plan (ILP)

Note that, depending on the focal construct of an assessment and instruction, the same accessibility supports may be considered universal in one assessment and an accommodation in another. For example, a specific EL support on an academic assessment might be part of the accessibility supports on an English language proficiency assessment.

It is also important to note that universal design principles are intended to help make the assessment accessible to as many students as possible, and accommodations are implemented during instruction and assessment as needed. With increased technology capabilities, some accommodations are embedded into the design and may be included in the online delivery of an assessment. This flexibility of technology enables more features to be available as accessibility options for more students.

V.3 Accessibility Supports for KELPA

Accessibility tools are available for all students taking various components of the Kansas assessments. Accessibility tools available for students vary by testing programs under the Kansas Assessment Program (KAP). The accessibility supports available to all students who take KELPA, as well as their descriptions and recommendations for use, are described in Table 0-1.

Table 0-1: Accessibility Tools for KELPA

Tool	Description
Eraser	Removes highlighting and striker marks from the screen.
Guide Line	When selected, follows the student’s pointer and lightly highlights the text of a reading passage line by line. This tool 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	When selected by test takers, changes the item number indicator at the top of the screen to blue with an accompanying flag graphic.
Notes	Presents a yellow rectangle on the screen where students can type notes about the test content.
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 available tags are: Main Idea, Supporting Details, Key Word, Evidence, Reread This, and Help.
Text-to-Speech audio (TTS)—directions	Allows students to choose to have a synthetic voice read directions aloud on the assessment.
Whole Screen Magnification	Allows students to magnify the screen up to four levels.
Sketch Pad	Allows students to draw, write, create shapes, etc.

The Kansas Accessibility Manual introduces a five-step decision-making process of selecting accessibility supports for instruction and assessments; it also elaborates on the principles of accessibility selection and provides tools to facilitate selection. This general process and principles are applicable to accessibility selection for KELPA. It is ensured that accommodations used on the state assessment have been a regular part of instruction. In addition, to ensure that students have had prior experience with the testing format being used, Technology Practice Tests and Subject-Oriented Practice Tests are used to familiarize students and teachers with the assessment format (including accessibility and accommodation tools) and the procedures for answering the different types of KELPA items.

V.4 Accommodations

Assessment accommodations are practices and procedures that provide equitable access for students with disabilities during assessments. These accommodations may not alter the assessment’s validity, score interpretation, reliability, or security. Accommodations on KELPA, documented in students’ IEPs, Section 504 plans, or ILPs, should reflect those that are provided during instruction. That is, students are provided accommodations during assessments that they use in their regular instruction. Some accommodations that are appropriate for instructional uses may not be appropriate for use on standardized assessments. For example, during instructional activities, a student with low vision may use read aloud, text-to-speech, and magnifying devices to access written materials. However, for KELPA, reading passages aloud to a student on the reading portion of the test 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, on the other hand, is an acceptable accommodation. According to the 2019–2020 KELPA Examiner's Manual (p. 15), prohibited practices include

- reading to students any text (including isolated words) in a KELPA domain assessment, unless directly specified by the KELPA Test Administration and Scoring Directions[13]
- translating passages, test questions, answer choices, labels, or other items into the student's native language
- teachers and students bringing pregenerated organizers, journals, logs, or notes into a test session

The 2019–2020 KELPA Examiner's Manual provides more details regarding accommodations in KELPA, including an overview, prohibited practices, and recording accommodations used during testing (e.g., most testing accommodations should be entered into the student's PNP). Additional information about accommodations or Kite tools can be found in the Kite Educator Portal Manual for Test Coordinators. Table 0-2 presents the accommodations available for KELPA.

[13] KELPA Test Administration and Scoring Directions for each grade are available for download from the Help tab in Kite Educator Portal. Because the documents contain scoring information and are considered secure materials, a link is not provided to the documents.

Table 0-2: Accommodations Available for KELPA

Tool	Description
Auditory calming	Provides relaxing, peaceful music that can play while the student takes the test.
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.
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/key instead of a mouse.
Whole screen magnification[a]	Allows students to magnify screen according to what has been set up in the PNP.

[a] When the whole screen magnification tool is set up through the Personal Needs Profile (PNP), the educator can set a default magnification level for each student.

V.4.1 Selection of Accommodations

According to the 2019–2020 KELPA Examiner's Manual, IEPs, 504 plans, services for English for speakers of other languages, and Student Improvement Team plans may use only accommodations documented on those plans. Accommodations must be recorded in a PNP or Access Profile in Educator Portal (for more information about setting options in the PNP, refer to the Kite Educator Portal Manual for Test Coordinators). To use an accommodation other than one listed in Tools and Accommodations for the Kansas Assessment Program (KAP), the examiner should contact the District Test Coordinator (DTC), and the DTC will send the request to KSDE. If the accommodation changes the construct being tested for a student, the test will not be valid for the student.

A few guidelines apply to every available accommodation on KELPA. First and foremost, only accommodations that have been used regularly in instruction may be used on state assessments. For accommodations to be available, teachers must submit accommodation requests through a student's PNP in Educator Portal. EL teams, IEP teams, and educators for 504 plans make decisions about accommodations. For ELs with disabilities, these teams should include an expert in the area of English language acquisition. These decision makers provide evidence of the need for accommodations and ensure that they are noted on the IEP, ILP, or 504 plan. Decision makers are responsible for entering information on accessibility features and accommodations from the IEP, ILP, or 504 plan into the planning tool so that all needed features and accommodations can be activated for the student. In general, refer to The Kansas Accessibility Manual for more in-depth information on accommodations for instruction and assessments.

A student's IEP will guide which accommodations to use for KELPA. Accommodations should be set in the PNP in Educator Portal before testing. As KELPA evaluates English language proficiency, accommodations of translation and text-to-speech[14] are not available during the test but can be used during instruction. Braille forms were not provided in the 2020 operational field-test administration. Spanish translation, although available in the Kite system, is not available for KELPA. Directions may be read to students in English; this feature is part of the text-to-speech function in Kite.

V.4.2 Frequency of Accommodations

Test administrators provide some accommodations that are allowed locally for KELPA, but other accommodations are built-in features in the Kite system. Because features in Kite are activated according to students' needs, teachers are required to mark those needs in the PNP. The PNPs submitted by teachers determine the availability of test accommodations for individual students. Table 0-3 presents the number of students who took KELPA in Kansas in 2020 and had PNP accommodation requests[15] for each accommodation. The summary in the table shows no accommodation requests for kindergarten, only a few requests for most grades, and slightly more requests for grade bands 6–8 and 9–12; whole screen magnification was the most commonly requested accommodation. Any nonstandard accommodation requests and approvals were handled by KSDE.

Table 0-3: Number of Students With Accommodation Requests by Grade or Grade Band

Grade or grade band	No. of requested accommodations						
	Auditory calming	Color contrast	Color overlay	Masking	Reverse contrast	Switches	WSM
K	0	0	0	0	0	0	0
1	1	0	0	0	0	0	3
2–3	6	4	1	0	0	1	4
4–5	11	6	4	1	0	7	8
6–8	16	8	4	1	1	5	15
9–12	23	12	2	2	3	6	26

Note. WSM = whole screen magnification.

V.4.3 Domain Exemptions

Students with specific disabilities such that no appropriate accommodations can be made for the students to access the domain test(s) may be exempted from testing in such domain(s). School districts may contact KSDE to request exemption for specific domain(s). Exempted domains will not be taken into account for overall proficiency. For example, deaf/hard of hearing students may be exempted from the listening test. For these students, overall proficiency will be determined by speaking, reading, and

[14] Text-to-speech is available for directions for all students.

[15] Some of the PNP requests may not be delivered via Kite.

writing performance, and students will be considered proficient if they score at the proficient level in the speaking, reading, and writing domains.

VI. Academic Achievement Standards and Reporting

This chapter mainly describes the standard-setting method used to set cut scores for KELPA and procedures used in the KELPA virtual standard-setting meeting; it also briefly presents the standard-setting results and score reporting. The standard-setting event was composed of two major activities: the panelist advance training and assignments and the virtual panel meetings of setting cut scores. For both activities, the Moodle learning management system was used as a digital platform to host, save, and deliver materials needed for the KELPA standard-setting activities. For detailed information regarding the KELPA standard-setting event, refer to the 2020 KELPA Standard-Setting Technical Report.

During the virtual standard-setting meeting that occurred in October 2020, the recruited Kansas educators set cut scores corresponding to the domain performance levels. This chapter summarizes procedures used to establish cut scores. The cut scores recommended by the standard-setting panels were approved by the Kansas State Board of Education (the Board hereafter) on January 12, 2021.

VI.1 State Adoption of Performance Standards for All English Learners

The performance standards for English learners (ELs) serve as a foundation for successful English language instruction. They depict expectations of what a student needs to know and be able to do to demonstrate adequate mastery of English language skills and knowledge to access and achieve grade-level content. The performance standards for KELPA are adapted from the previous English language proficiency assessment in Kansas, Kansas English Language Proficiency Assessment 2 (KELPA2). The domain performance levels reported in KELPA2 included five levels: Level 1—Beginning, Level 2—Early Intermediate, Level 3—Intermediate, Level 4—Early Advanced, and Level 5—Advanced. The major adaptation of KELPA performance standards from KELPA2 is the removal of Level 5—Advanced from each domain. Other than removing Level 5 performance, the definition of student proficiency on KELPA is consistent with KELPA2. That is, students must achieve performance Level 4 in each domain to be considered proficient (i.e., Level 3 on overall performance). Refer to Section VI.5.1 Student Reports for details about the relationship between domain performance and overall proficiency levels.

VI.2 Achievement Standard Setting

The objective for the standard-setting meeting was to set cut scores for each domain of KELPA at each grade or grade band. Grade-banded tests, such as grades 2–3, grades 4–5 and grades 6–8, include grade-specific cuts at each grade within the grade bands. The grades 9–12 test includes cut scores applicable for grade band 9–10 and grade band 11–12. The 2020 KELPA standard setting occurred virtually via Zoom meetings during a 2-week time frame: October 6–9 (Week 1) and October 12–16 (Week 2). Three panels were held each week; the panels for kindergarten and grades 2–3 and 4–5 were held virtually the first week, and those for grades 1, 6–8, and 9–12 took place virtually the second week. The kindergarten and grade-1 panels occurred over 3 days, the grade 6–8 panel over 5 days, and the other panels needed 4 days to complete all the standard-setting activities. The standard-setting event included advance (i.e., pre-meeting) training sessions (see Panelist Materials in the 2020 KELPA Standard-Setting Technical Report for training materials), advance assignments, multiple rounds of bookmark procedures (see Mitzel et al., 2001) for each grade or grade band in each domain during the meeting, and a vertical-articulation session after the meeting. The main goal of the KELPA standard setting was to establish three cut scores that differentiate four proficiency levels in each domain of the assessment at every grade (i.e., kindergarten through grade 8) or grade band (i.e., grade bands 9–10, 11–12). The panelists' recommended cut scores were presented to the Board for review and approval.

VI.2.1 Overview of the Bookmark Method

The Bookmark standard-setting method, which is widely used in K–12 educational assessment contexts, was used to establish cut scores. The Bookmark method is designed to generate cut scores based on panelists' review of collections of test items (Cizek & Bunch, 2007). In this method, an ordered item booklet (OIB) displays items ranked from easiest to hardest according to empirical item data (e.g., item response theory [IRT] item-parameter estimates). 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) would have a specific probability (i.e., 67%) of answering the item correctly.

The Bookmark method takes advantage of IRT scaling, which places students and items on the same scale. Based on the assumptions of the IRT model, a theoretically known probability for the student answering a given multiple-choice item correctly or obtaining a given score point (as in polytomously scored, e.g., constructed-response items) can be determined from a student's test score.

According to Cizek and Bunch (2007), the Bookmark method is commonly used for several reasons. First, from a practical perspective, it can be used for complex, mixed-format assessments, and panelists using it consider selected-response and constructed-response items together. Second, from the perspective of those who will be asked to make judgments, it presents a relatively simple task to participants. Third, in addition to being easy for participants, the Bookmark method is also comparatively easy for those who must implement the procedure. Finally, from a psychometric perspective, the method has certain advantages because of its foundation in IRT analysis and because of the fidelity of the method to the test-construction techniques that were used in developing the assessment. Bookmark standard setting relies on a reasonably large population of students taking the assessment that represents the full range of performance and an adequate number of items meeting criteria to become operational. In spring 2020, before testing in the state was cancelled because of the COVID-19 pandemic, KELPA was administered to a reasonably large population of students across grades K–12, using an adequate number of assessment items in each domain. Therefore, in consultation with Kansas State Department of Education (KSDE) and the Technical Advisory Committee (TAC), the Bookmark method was determined to be a reasonable method for establishing KELPA cut scores.

VI.2.2 Ordered Item Booklet

The OIB can contain both dichotomously scored items (e.g., multiple choice) and polytomously scored items (e.g., items with partial-credit scoring). Each dichotomously scored item appears in the OIB once, in a location determined by its difficulty (IRT *b* value). Each polytomously scored item appears several times in the booklet, once for each of its score points. Each dichotomous item will have one associated difficulty index, and each polytomous item will have as many difficulty indexes as it has score points (excluding zero). Also, each page in the OIB corresponded to a scale score. For KELPA standard setting, items included in the OIB were from the intact, operational forms. The same OIB was used for a given domain for all grades in a grade band.

VI.2.3 Panelist Recruitment

The standard-setting process relies on the expertise of educators. The goal of recruiting educators to participate in the KELPA virtual standard-setting meeting was to obtain a representative sample of Kansas educators who had experience teaching ELs and were able and willing to participate in a completely virtual event. To obtain a large and diverse pool of applicants, KSDE began recruitment efforts early in 2020. A recruitment letter and an event interest survey (see Appendix B: Standard-Setting Panelist Recruitment Letter and Survey in the 2020 KELPA Standard-Setting Technical Report) were sent via email distribution lists to curriculum leaders, test coordinators, and educators who provide English language instruction or services to ELs. KSDE staff also contacted individual educators in the field in an effort to promote and encourage participation in the event.

While the recruitment process was uniquely challenging because of the pandemic and the time and resource constraints teachers faced, every effort was made to encourage and support participation in the event. In total, KSDE recruited 55 educators to potentially serve as panelists for the event. All interested educators were asked to complete the interest survey (i.e., Appendix B in the 2020 KELPA Standard-Setting Technical Report). Survey items included basic demographic information, as well as criteria for participation identified by KSDE (described below). In addition, educators were asked to indicate whether they were willing and able to commit to up to 6 hours of advance training before the virtual standard-setting meeting and whether they would be available to attend one of two weeks of virtual standard-setting panel meetings in 2020: October 6–9 (Week 1 panels) or October 12–16 (Week 2 panels).

Several criteria were identified before recruitment to ensure selected panelists represented the following areas to the greatest extent possible:

- all 10 State Board districts
- a cross section of the state’s large and small districts, and rural and urban districts
- a range of length of teaching experience (i.e., new and veteran teachers)
- experience in the grade level of nomination
- experience with academic content areas
- experience in providing EL services or working with ELs in academic content areas
- English for Speakers of Other Languages (ESOL) endorsement
- diversity in ethnicity, race, and gender

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

- availability for a multiple-day virtual event (number of days varied by panel), plus approximately six hours of advance online training and activities via a Moodle course
- comfort level with participating in online video meetings
- 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 (broadband wired or wireless [3G or 4G/LTE]) and the following features:
    - participant’s email
    - ability to participate in an online Zoom meeting, including:

        - speakers and a microphone
        - video capability
    - ability to run Kite® Student Portal software, including:
        - desktop or laptop running Windows 8.1 or 10 or macOS 10.13–10.15
        - one of the following browsers: Firefox, Chrome, Edge, or Safari

Forty-three educators were selected to serve on one of the six grade-level or grade-band panels. Each panel represented one grade level or band (i.e., kindergarten, 1, 2–3, 4–5, 6–8, 9–12), as shown in Table 0-1. Grade-band panels were responsible for setting cut scores for each grade within their assigned band.

Table 0-1: Grade or Grade-Band Panels

Panel	Grade or grade bands in which cut scores were set
Kindergarten	Kindergarten
Grade 1	1
Grade band 2–3	2, 3
Grade band 4–5	4, 5
Grade band 6–8	6, 7, 8
Grade band 9–12	9–10, 11–12

On the first day of the Week 2 meetings, two panelists (one assigned to the grade 1 panel and the other to the grade band 6–8 panel) were unable to participate because of personal emergencies. As a result, 41 educators participated in the standard-setting event, with six panelists in the 6–8 panel and seven panelists in each of the other panels. As shown in Table 0-2, most panelists were female (93%), White (80%), from rural areas (51%), and had ESOL endorsement (95%) and 10 or more years of experience with ELs (76%) and/or English language arts (66%). In terms of other content areas, most panelists had only 0–2 years' experience with science (49%) and/or mathematics (41%). 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 KELPA standard-setting panel (i.e., 80% White and 93% female) approximately represented the demographic characteristics of the Kansas public school teacher population.

Table 0-2: Panelist Demographic Characteristics (N = 41)

Subgroups	Group (n)	%
Gender		
Female	38	93
Male	3	7
Race		
Black	1	2
Hispanic, Latino, or Spanish origin	5	12
Asian	2	5
White	33	80
Hispanic		
Yes	7	17
No	34	83
Area		
Rural	21	51
Suburban	13	32
Urban	7	17
English learners experience (years)		
3–5	3	7
6–9	7	17
10 or more	31	76
English language arts experience (years)		
0–2	4	10
3–5	1	2
6–9	9	22
10 or more	27	66
Science experience (years)		
0–2	20	49
3–5	2	5
6–9	7	17
10 or more	12	29
Mathematics experience (years)		
0–2	17	41
3–5	3	7
6–9	7	17
10 or more	14	34
ESOL endorsement		
Yes	39	95
No	2	5
Role		
Building administrator	1	2
Classroom teacher	9	22
District staff	11	27
Other	20	49

Note. ESOL = English to Speakers of Other Languages.

VI.2.4 Performance Level Descriptors

Policy performance level descriptors (PLDs) describe the expected English proficiency standards at each performance level. They guided the development of threshold PLDs used by panelists when setting cut scores. The policy PLDs for listening, speaking, reading, and writing were determined by KSDE and are the same across grades and grade bands. These policy PLDs define the general expectations for students' English proficiency within four different levels of performance. There are four performance levels for each of the four domains measured by the assessment.

- Level 1 Beginning: The student displays few grade-level English language skills and will benefit from EL program support.
- Level 2 Early Intermediate: The student presents evidence of developing grade-level English language skills and will benefit from EL program support.
- Level 3 Intermediate: The student applies some grade-level English language skills and will benefit from EL program support.
- Level 4 Early Advanced: The student demonstrates English language skills required for engagement with grade-level academic content instruction at a level comparable to non-ELs.

For standard setting, panelists used more-detailed descriptions of students' knowledge, skills, and abilities at each domain and grade level to help set cut scores that differentiate students' performance in the four performance levels. These more-detailed descriptions are referred to as threshold PLDs or standard-setting PLDs (see Appendix A in the 2020 KELPA Standard-Setting Technical Report for an example of listening threshold PLDs for kindergarten). Based on the policy PLDs and student expectations described in the 2018 Standards, the threshold PLDs for Level 4 were developed by one content expert at KSDE, three Kansas educators, and two Achievement and Assessment Institute (AAI) content-development staff. The threshold PLDs for Levels 2 and 3 were developed by the two AAI content-development staff only. The threshold PLDs are intended to reflect the minimum key knowledge and skills of for students in each performance level for each grade or grade band. They also 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 just barely meets the threshold (i.e., cut score) of the given level. Panelists used these definitions throughout the entire standard-setting process.

VI.2.5 Standard-Setting Procedure

The standard-setting event was conducted virtually using Zoom. There was one panel for each grade or grade-band test, for a total of six panels; three panels were conducted each week. The event included two principal activities: panelist advance training and assignments and the virtual standard-setting meeting. A Moodle course, developed for this event, contained the majority of the materials needed for both activities. Following a chronological order, this section describes activities both before and during the virtual standard-setting meeting. An example meeting agenda for the virtual panel meetings can be found in Appendix G in the 2020 KELPA Standard-Setting Technical Report.

VI.2.5.1 Panelist Advance Training and Assignments

Panelist advance training and assignments were a combination of synchronous and asynchronous activities conducted within the Moodle course (with one exception described below) before the virtual standard-setting meeting. They included three main parts in the Moodle course: an orientation meeting, training videos and a quiz, and assignments. The advance training and assignment were held in the weeks preceding the virtual panel meetings. For a calendar view of the advance training and assignments leading up to the virtual standard-setting meeting, see Appendix H in the 2020 KELPA Standard-Setting Technical Report.

VI.2.5.1.1 Orientation Meeting

For each week, a 30-minute, synchronous session was held late in the afternoon (i.e., after school hours); the sessions were recorded for anyone who could not attend the live session. The purpose of the orientation meetings was to review expectations for the panelist advance training and assignments, answer initial questions, and introduce support staff.

VI.2.5.1.2 Training Videos and Quiz

Advance panelist training was conducted asynchronously within the Moodle course. All panelists were required to watch all training videos and complete the online quiz and questionnaire as well as the confidentiality form. Completing the Moodle course took panelists approximately two hours and was self-paced. The four training videos included:

- Video 1: KELPA background, test design, and policy PLDs—About 20 minutes long, this video presented the purpose of KELPA, KELPA test takers, an overview and background of KELPA, KELPA test design, and scoring and reporting of KELPA.
- Video 2: Standard-setting overview—This video was about 20 minutes long and covered the purpose of the standard-setting meeting and the Bookmark method.
- Video 3: Standard-setting meeting step-by-step procedures—This video, about 30 minutes long, described several activities that happened during the virtual meeting but before the bookmark-placement process began, an overview of the bookmark-placement process, and the step-by-step process.
- Video 4: Meeting attendees' roles and responsibilities—This video was about 15 minutes long. It discussed panelists' roles during the standard-setting meeting, staff roles during the meeting, materials to be used for the meeting, the importance of material security, and the consent to confidentiality.

Panelists were required to respond to a short quiz of six questions covering critical points from the videos to ensure they had completed all training videos (see Appendix J in the 2020 KELPA Standard-Setting Technical Report). Panelists needed to answer all questions correctly before starting the actual standard-setting event. They were encouraged to review relevant parts of the training videos for questions answered incorrectly before retaking the quiz. Panelists could retake the quiz as many times as needed to achieve 100%. Nine (21%) of 43 panelists who attempted to take the quiz needed to take the quiz more than once; among them, six scored 100% after taking it twice and three required at least three attempts to achieve 100%.

Panelists also responded to two open-ended questions regarding any outstanding questions they had from training or other areas in which they wanted more information. Responses from the open-ended questionnaire were used to inform additional training at the start of the virtual standard-setting meeting. AAI Staff discussed and developed responses to submitted questions and comments before the virtual standard-setting meeting.

In addition, virtual office hours were available during designated times with AAI staff 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; AAI staff monitored the chat twice each day (once in the morning and once in the afternoon) during the 2-week training windows. Answers to the submitted questions, developed collaboratively by AAI staff, were posted on the Moodle course for panelists' review.

VI.2.5.1.3 Assignments

Panelists needed to complete two assignments in the Moodle course before attending the virtual standard-setting meeting. They were asked to complete the first assignment before working on the second one.

- Assignment 1: Taking the operational test—Panelists took a live, proctored KELPA test matching their assigned grade or grade band via Zoom. The purpose of participating in this testing session was to allow panelists to consider the items and test from the students' perspective and to think about the kinds of knowledge and skills measured by each item. Several Zoom meeting sessions were available and hosted by AAI staff for panelists to log in to (with webcam enabled). Once panelists were in the Zoom meeting, they used demo account information to log in to the Student Portal and take the test. Panelists had been instructed on how to download the Student Portal software and factors to consider during the operational test. Panelists were instructed to submit questions through a Moodle message to the facilitator. The facilitator compiled and brought all the questions for discussion on Day 1 of the virtual meeting.
- Assignment 2: Just-barely-student activity—After watching a short training video about the purpose and development of threshold PLDs and how to read the draft PLD documents, panelists continued to review hard copies of the draft threshold PLDs. They were instructed to write notes on the just-barely-student activity worksheet where draft threshold PLDs were provided and prepare to discuss rationales for any suggested changes on Day 1 of the virtual meeting. Instructions to panelists emphasized that the draft threshold PLDs had been written by Kansas educators, so significant changes to the PLDs were not expected.

VI.2.5.2 Virtual Standard-Setting Meeting

Separate Zoom panel meetings were set up, one per panel per week, during the virtual standard-setting meeting. Panelists were required to have their webcams turned on during the meetings. Three additional training presentations were provided to panelists throughout the virtual, multiday standard-setting meeting. The purpose of these additional training presentations was to review information and knowledge critical to the standard-setting process. The three additional training presentations are described below.

- Presentation 1, delivered during the morning orientation on Day 1, focused on KELPA standards, test design, and the PLDs. This training helped panelists distinguish the differences between the performance expectations in the 2018 Standards and the KELPA domain performance levels.
- Presentation 2, delivered immediately before panelists practiced placing bookmarks, provided detailed information about the OIBs, different item types within OIBs (i.e., dichotomous and polytomous items), and the item-map tables. In addition, scoring rubrics

for writing and speaking were reviewed to allow panelists to understand performance expectations at various score levels for the educator-scored items.

- Presentation 3, delivered after setting cut scores for the highest grade within a grade band, provided an overview of the modified bookmark procedure. The modified bookmark procedure included only two rounds of bookmark placement, and panelists started with the cut-score recommendations for the highest grade as a starting point for setting cut scores for the lower grade levels within a grade band.

VI.2.5.2.1 Group Discussion of Just-Barely Students for Each Domain

Within each panel, the facilitator delivered Presentation 1. Next, panelists discussed the draft threshold PLDs, using their notes from the second pre-meeting assignment. Facilitators reminded panelists that the draft threshold PLDs had been developed by Kansas educators and were not intended to be significantly revised. To improve the definition of just-barely students, however, panelists could suggest changes to better differentiate performance expectations among performance levels and grade levels. This discussion of just-barely students led to a consensus decision of the final definitions of just-barely students for each performance level and each domain. The final just-barely student definitions were uploaded to the Moodle course site for panelists' electronic access.

The discussion of just-barely students for the meeting was planned to be completed for all domains at once on the first day of the meeting. As planned, panelists discussed just-barely students for all domains on the first day of the first week of the virtual meeting. However, although just-barely student definitions were reviewed the first morning of the meeting during the first week, panelists spent additional time refining the just-barely definitions for each domain right before the first round of bookmark placements. Therefore, experiences from the first week of the meeting indicated that reviewing and updating the just-barely student definitions, one domain at a time and immediately before each domain's bookmark-placement activity, was more helpful for panelists in placing bookmarks; the modified discussion procedures were implemented in Week 2.

VI.2.5.2.2 Review of Item Maps and Scoring Rubrics

The facilitator in each panel delivered Presentation 2 (i.e., descriptions of different types of items in the OIB). They then led the panelists to review the item map or maps and interpreted how item maps could be used to facilitate their bookmark placements. Panelists were also instructed to review the scoring rubrics for writing and speaking to help them understand performance expected to earn each of the score points on educator-scored items. All documents were stored electronically in the Moodle course.

VI.2.5.2.3 Bookmark Practice

The purpose of this practice was to allow panelists to familiarize themselves with the bookmark procedures. Using a practice OIB of 10 writing items at the upper end of the difficulty range and the practice item-map table, the panelists reviewed the first half of the practice OIB and completed the following steps:

- Panelists answered the question "What does the student have to know and be able to do to answer each item or score point correctly?"
- Using both the updated, just-barely Level-4 descriptions, panelists answered the question "Would 20 out of 30 just-barely Level-4 students be able to answer each item correctly?"

- For the polytomous items, panelists use the just-barely Level-4 descriptions, as well as the scoring rubric to answered the question “Would 20 out of 30 just-barely Level-4 students be able to earn this score point or higher?”
- Facilitators asked panelists to use the Thumbs Up button on Zoom to respond to the question.
- Panelists shared their thoughts and rationales for their answers to the questions.

Panelists moved to the second half of the practice OIB and repeated the steps above. Practice was done only for Level 4. Panelists were told that, when doing their official placements, they would start with Level 4 and then return to the first item in the OIB to continue the process for Level 2 and then Level 3.

VI.2.5.2.4 Review the Ordered Item Booklet

Using the actual OIB in the Moodle course, the purpose of this facilitator-led exercise was to provide panelists the opportunity to holistically review the items in the OIB to avoid one item having too much influence during the actual bookmarking process; that is, panelists should consider the group of items within which the cut score should be placed. As panelists reviewed each item in order in the OIB, they thought about three questions:

- What does the student have to know and be able to do to answer this item correctly?
- What makes this item more difficult than the ones preceding it?
- Are there points in the OIB where the just-barely student in each level would not be able to answer an item correctly?

VI.2.5.2.5 Readiness Poll

Before starting the first round of ratings, panelists were asked to participate in a readiness check-in poll via Zoom. The purpose of the check-in poll was to ensure that panelists felt confident they understood the process and were ready to proceed with individual bookmark placements. If any panelists did not feel confident with the bookmarking tasks, additional training would be provided until all felt ready to begin. The readiness poll survey posed two questions: (a) Do you clearly understand your role in this event and what you are being asked to do? and (b) Are you ready to start Round 1? All panelists responded affirmatively.

VI.2.5.2.6 Setting Cut Scores

After panelists were comfortable with the rating procedures, they began the first round of item review and bookmark placements for the highest grade or grade band they were assigned to (e.g., grade 5 for the 4–5 panel). Panelists were instructed to refer to their materials organizer, which listed the materials they needed and where to find each one. For each domain, panelists placed bookmarks using three rounds of ratings. Panels completed the entire process for one domain before continuing to the next. The order of domains was listening, speaking, reading, and writing, which is the order in which students acquire English language skills. Because the grade-band tests have identical content for each grade within the band, but different performance expectations are associated with 2018 Kansas standards for ELs across grade levels, panelists used the cut scores set for the highest grade level or band within their band (e.g., grade band 11–12 in high school) as the starting point for adjusting their bookmark placements downward for the next grade.

For kindergarten and grade 1, as well as for the highest grade level or band within each grade band (i.e., grades 3, 5, 8, 11–12), three rounds of bookmark placements were implemented. Procedures in the next sections describe the three bookmark-placement rounds for kindergarten, grade 1, and the highest grade level or band within each grade band.

For grades 2, 4, 6, 7, and 9–10, modified bookmark placements (which included only two rounds) were carried out where Round 2 and Round 3 placements, described below, were implemented. In other words, Round 1 procedures were not applied to these lower grades within each grade band. After panels completed bookmark placements for the highest grade in a grade band and before they moved on to the next grade, a short presentation (i.e., Presentation 3 described in the beginning of Section VI.2.5.2 Virtual Standard-Setting Meeting) was given to the panelists to review the modified standard-setting process. The two main modifications were that panelists were instructed to start their bookmark placement from the final bookmark placement of the grade level they had just completed, and panelists were provided impact data (i.e., the percentage of students who would be categorized in each performance level based on the panel's median bookmark placements from the round) after completion of Round 1 placement. Panelists were instructed to place bookmarks for the target grade below the highest grade in the grade band unless there was a compelling content-related reason for the two grade levels to have the same cut score.

Round 1 Placement. Panelists were asked to work individually using the OIB, item-map table, and just-barely student definition. They were also asked to use the 2018 Standards, scoring rubrics (for speaking and writing), item stimuli (for reading and listening) and example student responses (for speaking and writing) when needed. The facilitator instructed panelists to work alone, avoiding discussion with others, to ensure that the Round 1 bookmark placements were established independently.

Starting with Item 1 in the OIB, panelists reviewed each item individually and determined whether a just-barely Level-4 student would answer the item correctly (or earn the score point or higher for polytomous items). Panelists were instructed to place bookmarks where two-thirds of the just-barely Level 4 students at each level 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 ensure correct bookmark placement by making sure that just-barely Level-4 students would not be able to answer the next few items correctly. Once panelists placed their bookmarks for just-barely Level 4, they returned to the first item and repeated the bookmark-placement process for just-barely Level-2 students. They then placed their bookmark for just-barely Level 3 between just-barely Level-2 and just-barely Level-4 students.

Once panelists finished making 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. Facilitators displayed (i.e., shared their screens in Zoom) the Round 1 summary results derived from panelists' bookmark placements. Table 0-3 and Figure 0-1 provide example results. Facilitators pointed out that the bar chart (see Figure 0-1) was intended to show all the individual placements for the panel. Panelists compared their own results with those of the other panelists and were asked to think about two questions:

- Am I relatively strict or lenient in relation to others?
- Am I consistently strict or lenient across all three levels?

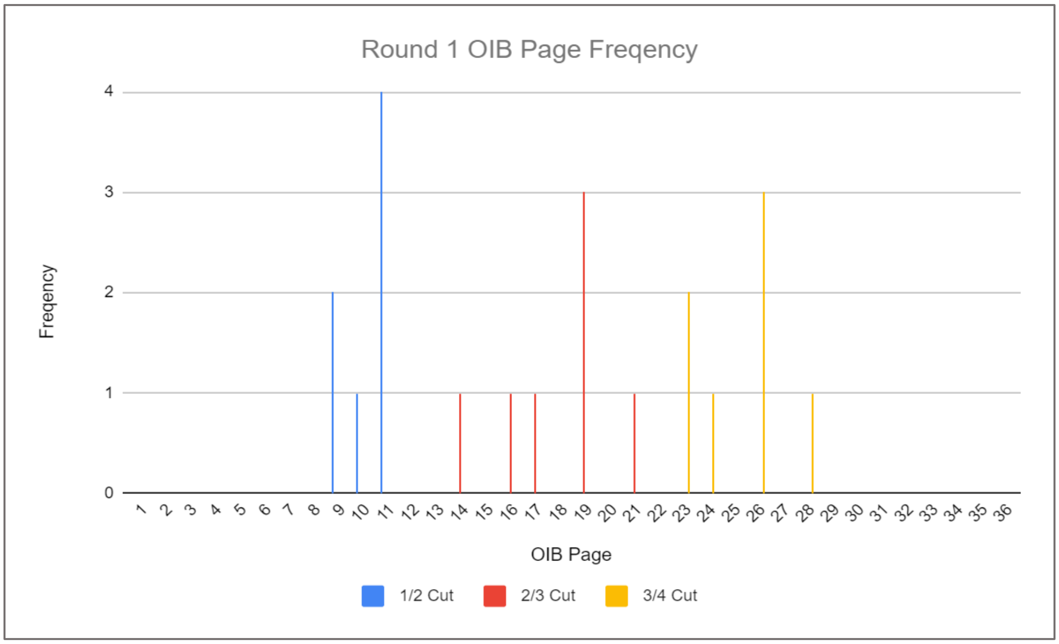
After individuals had reviewed their results in relation to the panel summary results, discussions began. The facilitator prompted panelists to review the Level-4 placements and asked them to share their thoughts and rationales with the group. For example, each of the seven panelists in one panel placed bookmarks for the Proficient cut score for grade-3 listening at items 18, 18, 19, 21, 21, 22, and 23, respectively. Discourse centered on what students should know to attain a given achievement level. In this example, the panelists had collectively identified a range from items 18 to 23 in which the Proficient cut score should fall. The group's consideration of the skill level and knowledge that should be mastered for the Proficient performance level (i.e., Level 4) needed only focused on the content in this six-item range. This process of discussing items within the panel's identified range was repeated for each cut point within the given domain.

Table 0-3: Example of Panel Summary Results for Bookmark Placement

	Level 1/2 cut	Level 2/3 cut	Level 3/4 cut
Minimum OIB page	9	14	23
Median OIB page	11	19	26
Maximum OIB page	11	21	28

Note. OIB = ordered item booklet. The numbers in the table are OIB page numbers with bookmarks.

Figure 0-1: Example Frequency of Round 1 Ordered Item Booklet Page Numbers With Bookmarks

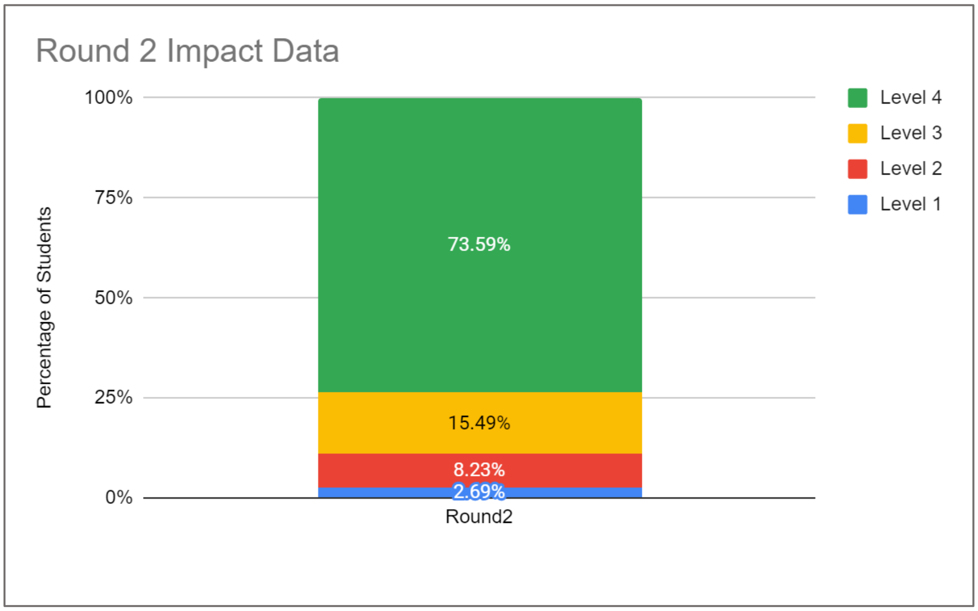


Round 2 Placement. After discussion, panelists were directed to repeat the Round 1 procedures with consideration of Round 1 feedback results and the panel discussions. In this step, panelists were told not to feel compelled to conform to the panel's results. Rather, they were to use it as additional information to reconsider their placements, but did not have to change their bookmark placements. At the end of Round 2, panelists submitted their Round 2 placements electronically through a Google Form. The same summary table and chart prepared for Round 1, mentioned above, were also prepared for Round 2. In addition to these results, the impact data was generated by the data tool.

Round 2 Results and Discussion. Facilitators displayed (i.e., shared their screens in Zoom) the summary of Round 2 results (refer to Table VI-3 for an example) and a frequency bar chart of bookmark placements (refer to Figure VI-1 for an example) for Round 2 results. Facilitators pointed out that the range of bookmark placements that was the focus of discussion should likely be narrower in Round 2 than in Round 1.

In addition, facilitators displayed (i.e., shared their screens in Zoom) the percentage of students who would be categorized in each performance level according to the panel's median bookmark placements from Round 2 (i.e., impact data). Figure 0-2 provides an example of the impact data. Panelists were instructed to think about whether the impact data were consistent with content-based expectations at each performance level. In other words, given what they knew about the student population and the knowledge and skills needed at each level, were the percentages of students at each level reasonable.

Figure 0-2: Example Round 2 Impact Data



Panelists again compared their own bookmarks with those of others in the panel and considered whether they were relatively strict or lenient in relation to others. Panelists were guided to think about three questions and shared their thoughts and rationales with the group:

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

Round 3 Placement. Panelists were instructed to use all available information to guide their decisions: individual and median bookmark placements over the two rounds, policy-domain PLDs, just-barely student knowledge, and any notes from reviewing the OIB, and the input of their colleagues through discussion. They could use the available information to reconsider their bookmark placements but did not have to change them. Panelists independently recorded their final cut scores in the Google Form. Facilitators presented (i.e., shared their screens in Zoom) the same summary table and chart (described earlier for Round 2) prepared for Round 3. The median bookmark placements for the third round were presented to panelists as the final panel recommendation to be shared with the vertical-articulation panel (described in Section VI.2.6 Vertical-Articulation Procedure and KSDE.

Evaluation Forms. After bookmark placement for Round 3, panelists completed the cut-score evaluation form (see Appendix E in 2020 KELPA Standard-Setting Technical Report) for the domain (i.e., the first domain was listening) for which they had just completed their cut-score recommendations. The steps for Rounds 1–3 were repeated for the remaining domains (i.e., speaking, reading, writing). Discussions about the cut-score evaluation results are in Section VI.3.1 Results of Cut-Score Evaluations. After completing all four domains in kindergarten and grade 1, panelists completed the standard-setting process evaluation form (see Appendix F in 2020 KELPA Standard-Setting Technical Report) and adjourned. The other panels continued working on setting cut scores for the lower grade(s) in the grade band. Section VI.3.2 Panelist Training and Meeting Evaluations discusses the results of the standard-setting process evaluation.

VI.2.6 Vertical-Articulation Procedure

After the standard-setting panel meetings ended, a vertical-articulation panel was conducted to evaluate the consistency of cut scores across all grade levels. The vertical-articulation panel addressed any issues associated with unintended or inappropriate reversals of performance expectations from one grade level to the next; these can sometimes occur when separate panels set cuts for different grade levels. Representatives from each panel (one each from the kindergarten and grade 9–12 panels and two from each of the other panels) participated in the vertical-articulation panel and discussed the panel-recommended cut scores during standard setting across all grade levels. These representatives (i.e., the vertical-articulation panelists) were recommended by the standard-setting panel facilitators from each panel because they were very engaging and contributed significantly to the standard-setting panel discussions. Articulation panelist first review cut scores recommended by standard-setting panels across all six grades or grade bands and identified cut scores leading to inconsistency of performance expectations across grades. Then, panelists discussed possible adjustments to these cuts, one domain at a time. Group consensus was required for any recommended adjustments.

VI.2.7 Standard-Setting Results

The cut scores recommended by the standard-setting panel represented the median OIB page number for each performance level, domain, and grade or grade band. Each page of the OIB represented one item with a calibrated item difficulty on an IRT scale. The calibration was conducted using the operational field-test data. Note that, because items in each of the domain tests within a grade or grade band were calibrated separately, the IRT scale of each domain test is different. Thus, the cut scores are not directly comparable across domains within a grade or grade band.

Table 0-4 presents the scale-score cuts corresponding to the panel-recommended cuts. Table 0-5 presents the scale-score cuts recommended by the articulation panel. As highlighted in Table 0-5, across all grades or grade bands and domains, the articulation panel made changes to six cut scores (three in the listening domain, one in the speaking domain, one in reading, and one in writing). The round-by–round results are presented in Chapter V of the 2020 KELPA Standard-Setting Report, and the evidence of increased consensus on bookmark placements over rounds can be found in Chapter VI of the 2020 KELPA Standard-Setting Report. A presentation of the proportion of students in each performance level can be found in Chapter IV. Technical Quality—Other of this manual.

Table 0-4: Scale-Score Cuts Recommended by KELPA Standard-Setting Panel by Domain and Grade

Grade or	Listening			Speaking			Reading			Writing		
grade band	L2	L3	L4	L2	L3	L4	L2	L3	L4	L2	L3	L4
K	377	396	556	449	519	611	447	536	630	406	496	682
1	337	416	545	359	462	558	404	448	579	337	422	579
2	255	372	446	327	419	524	394	432	480	340	427	519
3	301	381	458	365	427	527	419	460	546	395	459	601
4	318	348	452	324	395	493	353	431	494	340	396	510
5	333	367	464	354	414	515	391	473	532	345	443	554
6	347	367	457	344	414	514	334	435	539	322	410	544
7	350	388	469	352	422	520	372	471	566	322	452	569
8	363	396	517	359	428	533	384	491	609	326	475	610
9–10	361	393	431	396	444	504	449	501	547	373	462	494
11–12	375	406	482	403	447	513	464	532	574	431	485	541

Note. L2= Level 2 cut; L3= Level 3 cut; L4= Level 4 cut.

Table 0-5: Scale-Score Cuts Recommended by KELPA Articulation Panel

Grade or	Listening			Speaking			Reading			Writing		
grade band	L2	L3	L4	L2	L3	L4	L2	L3	L4	L2	L3	L4
K	377	396	556	449	519	587	447	536	630	406	496	682
1	337	416	545	359	462	558	404	448	579	337	422	579
2	275	372	446	327	419	524	367	432	480	340	427	519
3	301	381	458	365	427	527	419	460	546	395	459	601
4	318	348	452	324	395	493	353	431	494	340	396	510
5	333	367	464	354	414	515	391	473	532	345	443	554
6	347	367	457	344	414	514	334	435	539	322	410	544
7	350	388	469	352	422	520	372	471	566	322	452	569
8	363	396	489	359	428	533	384	491	609	326	475	590
9–10	361	393	449	396	444	504	449	501	547	373	462	494
11–12	375	406	482	403	447	513	464	532	574	431	485	541

Note. L2= Level 2 cut; L3= Level 3 cut; L4= Level 4 cut. The highlighted cells indicate the cut scores that were adjusted during articulation.

VI.3 Evaluations

As described above, panelists completed two types of evaluation forms during the standard-setting meeting: cut-score evaluation forms (one for each domain) and a standard-setting-process evaluation form. The cut-score evaluation forms had two versions, one for grades that implemented three rounds of bookmark placements (K, 1, 3, 5, 8, and 11–12), and another version for grades that implemented two rounds of bookmark placements (2, 4, 6, 7, and 9–10). A total of 15 evaluation strands were included (i.e., seven in the cut-score evaluation forms and eight in the standard-setting-process evaluation form) so that a variety of event aspects could be evaluated. For more information about evaluation forms, refer to Chapter III in the 2020 KELPA Standard-Setting Technical Report. In addition, a TAC member observed the virtual standard-setting meeting during the first week of the process to evaluate fidelity and provide feedback on the process.

VI.3.1 Results of Cut-Score Evaluations

The cut-score evaluation form consisted of two main sections: panelists' perceptions of influential factors in their bookmark placements and their perceptions of the panel bookmark-placement results in each domain. Results show several major findings, described next.

VI.3.1.1 Influential Factors

Panelists' ratings of influential factors varied for each round and domain. The pattern was that, for grades K, 1, 3, 5, 8, and 11–12, the just-barely student definitions at Levels 2, 3, and 4 were considered the most influential factor for Round 1 across all four domains, as well as in Round 2 for the listening domain. Group discussion was rated as the most influential factor in Round 2 for the speaking, reading, and writing domains, as the second-most influential factor in Round 2 for the listening domain, and as the most influential factor across all four domains in Round 3 (see Tables L-1, L-3, L-5, and L-7 in Appendix L of the 2020 KELPA Standard-Setting Technical Report for more details). For grades 2, 4, 6, 7, and 9–10, the just-barely definitions at Levels 2, 3, and 4 were rated the most influential factor for Round 1 across all four domains and the second-most influential factor for speaking, reading, and writing domains in Round 2. Group discussion was rated the most influential factor in Round 2 across all four domains. The least influential factors across rounds and domains included the policy PLDs and the 2018 Standards. (See Tables L-9, L-11, L-13, and L-15 in Appendix L of the 2020 KELPA Standard-Setting Technical Report for more details.)

VI.3.1.2 Bookmark-Placement Results

Panelist responses to bookmark-placement results were generally positive. The majority of panelists agreed or strongly agreed that

- the summary panel results from Rounds 1 and 2 (for lower grades in grade bands, Round 1 only) for each domain were clear and useful.
- the impact results for each level were reasonable.
- the cut score for each level was appropriate based on the policy PLDs and just-barely student definitions.
- the cut score for each level was defensible because of panelists' adherence to procedures.

Results showed that the percentage of panelists who agreed with these statements increased as they worked through the domains and that the percentage who agreed with them increased as they completed additional grade levels. (See Tables L-2, L-4, L-6, L-8, L-10, L-12, L-14, and L-16 in Appendix L of the 2020 KELPA Standard-Setting Technical Report for more details.)

VI.3.2 Panelist Training and Meeting Evaluations

This subsection summarizes the main findings from responses to the standard-setting-process evaluation. Detailed findings can be found in Tables L-17 through L-24 in Appendix L of the 2020 KELPA Standard-Setting Technical Report.

First, the majority of panelists (70%–95%) agreed or strongly agreed with statements about the effectiveness, clarity, and organization of the advance training and assignments. For example, most panelists agreed or strongly agreed the advance training and activities helped them get ready for the standard-setting event (90%) and clearly explained the meeting procedures (95%). The majority of panelists (93%–100%) also agreed or strongly agreed with statements about the effectiveness, importance, and organization of the welcome and orientation sessions, the group discussions, and the practice sessions. For example, all panelists agreed or strongly agreed that the welcome and orientation session was well organized, and nearly all of them (98%) agreed or strongly agreed that the practice bookmark-placement activity helped them understand the procedures.

Furthermore, most panelists believed the amount of time used for advance training and assignments (85%) and the additional training during the standard-setting meeting (93%) was about right. 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.

Panelists' feedback on the group discussion was also positive. For example, all panelists agreed or strongly agreed that the discussions about the just-barely student definitions were helpful and that the discussions after each round of ratings were helpful. Similarly, the majority of panelists found the policy PLDs to be the most useful (95%) and that the other technology features (including features in Moodle, Zoom, and Google Forms) were effective or easy to use (93%–100%). Finally, nearly all panelists (95%–100%) found the support staff (e.g., the lead panel facilitator) to be moderately helpful or very helpful.

VI.3.3 Technical Advisory Committee Feedback

As mentioned, a member from the Kansas TAC observed the virtual standard-setting meeting during the first week of the process, as well as the subsequent articulation-panel meeting. The TAC member was able to join each of the three panel meetings at different times throughout each meeting day, along with the staff debrief meetings each day. The TAC member found that panelists were engaged in the process, followed standard-setting procedures, and were well supported by staff. In a memorandum, the TAC member concluded that the event was implemented with fidelity to the procedures and that the panel-recommended cut scores were derived from a well-implemented meeting. The standard-setting results and the memorandum were shared and discussed with the full TAC in November 2020. The TAC memorandum is available in Appendix N of the 2020 KELPA Standard-Setting Technical Report.

VI.4 Challenging and Aligned Achievement Standards

KELPA achievement standards represent challenging expectations for ELs, and they are aligned with the state standards for ELs. The 2018 Standards include performance-level rubrics in each stage of language acquisition and use of language to construct meaning, convey ideas, and engage in grade-level content. The performance-level rubrics described in the 2018 Standards are the foundations for the policy-domain PLDs and connect to expectations for acquiring content-area knowledge and skills. The policy-domain PLDs define the general expectations for student performance relative to the expectations expressed in the set of content standards for ELs. KELPA grade-level, domain-specific threshold PLDs were drafted to align with the grade- and domain-specific content standards for ELs. AAI content experts worked alongside KSDE staff and Kansas educators to define the threshold PLD for Level 4, and AAI staff developed the Level 3 and Level 2 threshold PLDs. The outcome of the collaboration was a set of draft threshold PLDs that describe student-performance expectations that reflect the content and rigor of the

2018 Standards. As described in Section VI.2.5 Standard-Setting Procedure, the threshold PLDs were reviewed and revised by the panelists and used as the foundation for setting the cut scores.

The final recommended cut scores from the standard-setting and articulation-panel meetings were presented to KSDE on December 8, 2020, followed by a period for public comment and review of the recommended cut scores. On January 12, 2021, the Board approved the final recommended cut scores for KELPA.

VI.5 Reporting

The 2020 KELPA testing window closed on March 13, and the scoring window closed on March 27. Normally, KELPA student reports are delivered before the end of the spring semester. However, because of COVID-19, the standard-setting event was delayed until October 2020, thereby delaying the Board's approval of the new cut scores. As a result, the delivery of KELPA student reports to educators was postponed from summer 2020 to January 2021.

The KELPA provides separate score reports to students, schools, and districts in an understandable and uniform format. These reports include the overall proficiency level and the domain performance levels that are used to determine the overall proficiency level. Students must attain Level 4 (Early Advanced) in all domains to be considered proficient. To assist readers in interpreting the information in the reports, nontechnical language is used and descriptions of what students should know and be able to do at each performance level are provided. In addition, the KELPA Educator Guide and the KELPA Parent Guide are provided to assist the interpretation of the score reports.

VI.5.1 Student Reports

Performance levels for listening, speaking, reading, and writing were used to determine the overall proficiency level. Overall proficiency levels were defined by KSDE.

- Level 1 Not Proficient: Students who are not yet proficient have not attained a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels of Beginning and Early Intermediate in all four domains. Not Proficient students are eligible for ongoing program support.
- Level 2 Nearly Proficient: Students are nearly proficient when they approach a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by students' attaining performance levels above Early Intermediate but not meeting the requirements to be proficient. Nearly Proficient students are eligible for ongoing program support.
- Level 3 Proficient: Students are proficient when they attain a level of English language skills necessary to independently produce, interpret, collaborate on, and succeed in grade-level, content-related academic tasks in English. This is indicated by attaining performance levels of Early Advanced in all domains.

To be considered proficient (i.e., Level 3 on overall performance) and eligible to exit the EL program, students must receive 4s on all domain scores. Students who receive all 1s or 2s on the domain scores are considered not proficient, in other words, Level 1 on overall proficiency. Students who do not meet the criteria for either Level 1 or Level 3 are considered nearly proficient, that is, Level 2 on overall proficiency. A sample of a KELPA Student Report is shown in Appendix D.

VI.5.2 Interpretive Guides

The KELPA Educator Guide and the KELPA Parent Guide (see Appendix E and Appendix F) are available to download from the Kansas Assessment Program website. These guides explain the scores presented in the report and how the overall proficiency level and domain performance levels are determined. They also help readers understand students’ progress toward proficiency.

References

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51.

Brennan, R. L. (2004). *BB-CLASS* (Version 1.1) [Computer software]. University of Iowa, Center for Advanced Studies in Measurement and Assessment. https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs#class

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Center for Applied Linguistics. (2017). *Annual technical report for ACCESS for ELLS® 2.0 online English language proficiency test, series 401, 2016-2017 administration*. https://sde.ok.gov/sites/default/files/documents/files/ACCESS%20Online%20Technical%20Report.pdf

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289. https://doi.org/10.3102/10769986022003265

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/BF02310555

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 37-50.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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

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

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

Magilner, A., & Magilner, T. (2006). *Children's writers word book* (2nd ed.). Writer's Digest Books.

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

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

National Teacher and Principal Survey. (2020b). *Average and median age of public school teachers and percentage distribution of teachers by age category, sex, 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_fltable02_t1s.asp

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

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Shyyan, V., Thurlow, M., Christensen, L., Lazarus, S., Paul, J., & Touchette, B. (2016). *CCSSO accessibility manual: How to select, administer, and evaluate use of accessibility supports for instruction and assessment of all students*. Council of Chief State School Officers. https://www.ccsso.org/sites/default/files/2017-10/CCSSO%20Accessibility%20Manual.docx

Taylor, S. E., Nieroroda, B. W., & Birsner, E. P. (Eds.). (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Steck-Vaughn.

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Appendix A: KSDE Presentation to Kansas Board of Education

6/23/2021

Kansas Standards for English Learners

August 14, 2018

Video from Committee

Add link

EL Standards Document

- Crosswalk
- Stages of Language Acquisition
- Performance Level Rubric
- Link to the Kansas State Literacy Plan
  - Vocabulary
- Classroom Supports
- Standards: K-8, 9-10, and 11-12
  - 4 Domains – Reading, Writing, Speaking, Listening
  - Level 5 – ELA standard
- Q&A and References

Public Comments

- Public Hearings
  - Wichita – July 24
    - 14 attendees
    - 9 districts
  - Topeka – August 7
    - 7 attendees
    - 5 districts
- Written Comments
  - 1 district
  - 1 higher education

Public Comments

Recommendations (included in updated draft standards)

- Add Guiding Principles from current ELPA 21 Standards
- Clarify reference coding

Public Comments

Likes

- Organization
- Grade specific
- For use by all teachers, not isolated
- Teacher friendly
- Reference to ELA Standards
- Addition of grade level vocabulary
- Crosswalk
- Performance level rubric

1

Appendix B: Item Statistics Flagging Criteria

Table A-1: Item Statistics Flagging Criteria

Statistic	Criterion
Omit	Omit correlation > .1 Omit percentage > .05
Differential item functioning	Gender R^2 change > 0.035 Race R^2 change > 0.035 EL R^2 change > 0.035
Item-total correlation	Item-total correlation ≤ .249
p value	*p* value = 0
Item response theory—discrimination	a < 0.3 and abs(b1 . . . b10) ≤ 5 a < 0.3 and abs(b1 . . . b10) > 5 0.3 ≤ a ≤ 0.699
Item response theory—difficulty	abs(b1 . . . b10) > 3.5
Item-total correlation of keyed response for selecting-key items	Correlation of keyed response < 0
Item-total correlation of distractors for selecting-key items	Correlation of keyed response < 0 and correlation of distractors > 0
	Correlation of distractors > .10 and correlation of distractors > correlation of keyed response

Appendix C: Conditional Standard Error of Measurement

Figure C-1: Conditional Standard Error of Measurement for Listening

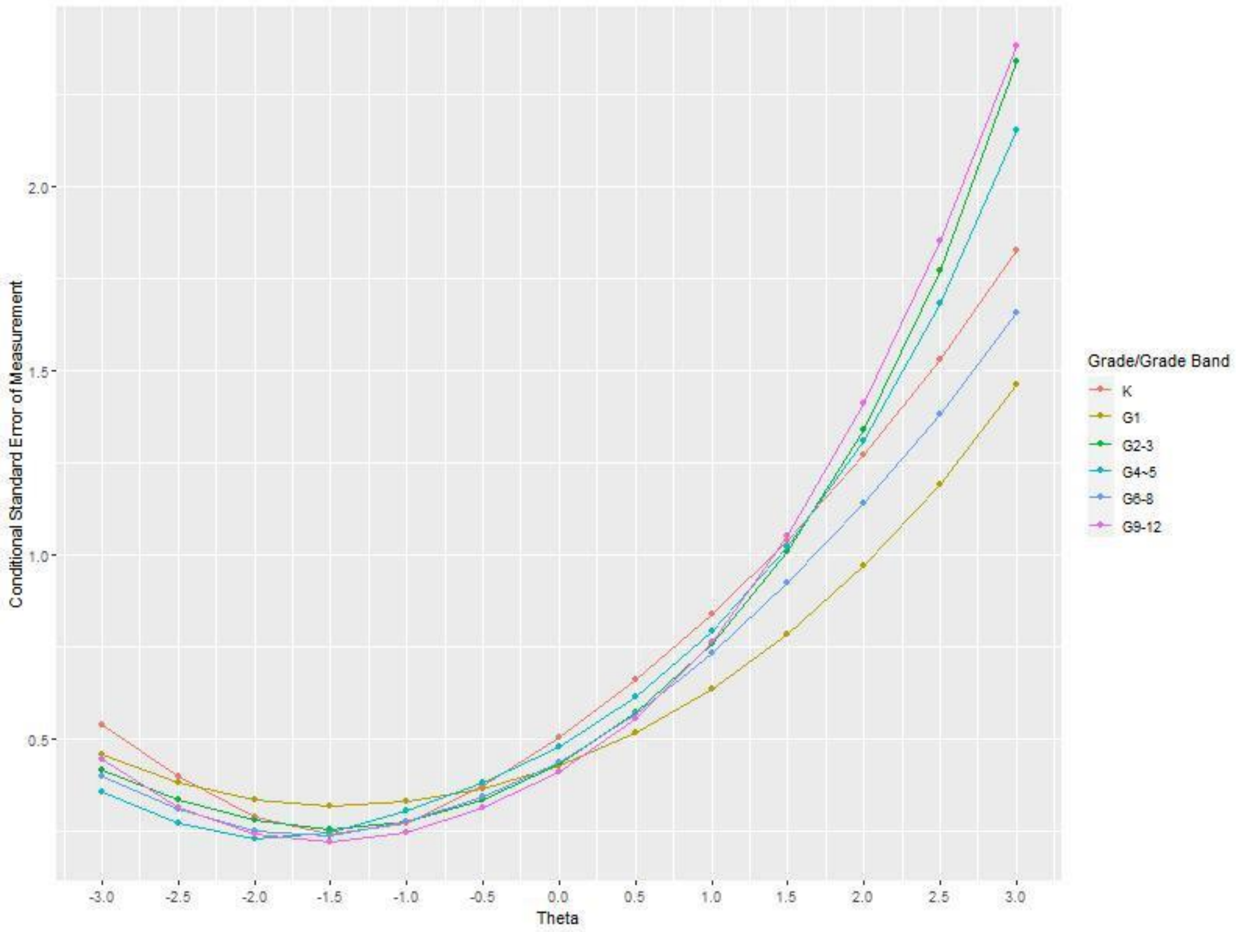


Figure C-2: Conditional Standard Error of Measurement for Speaking

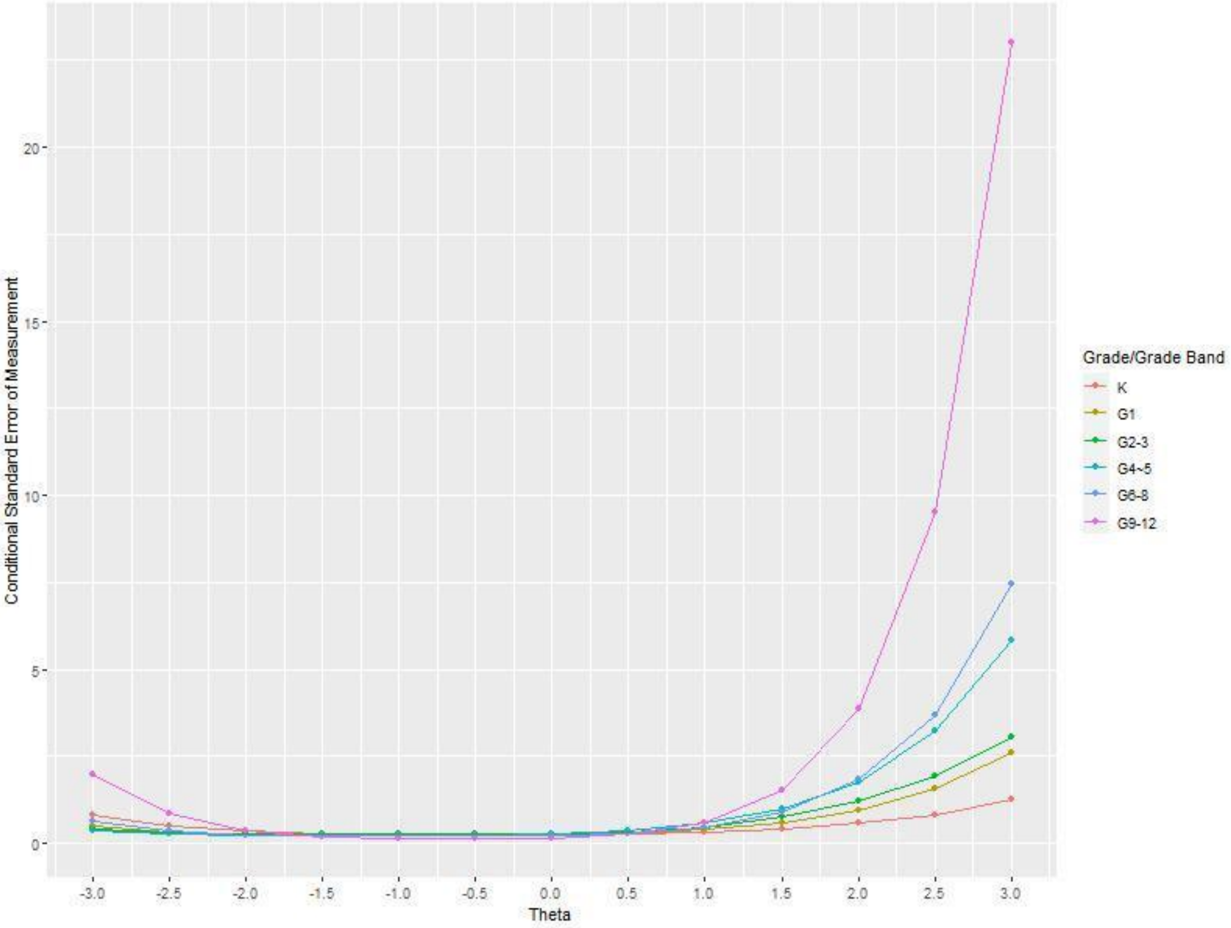


Figure C-3: Conditional Standard Error of Measurement for Reading

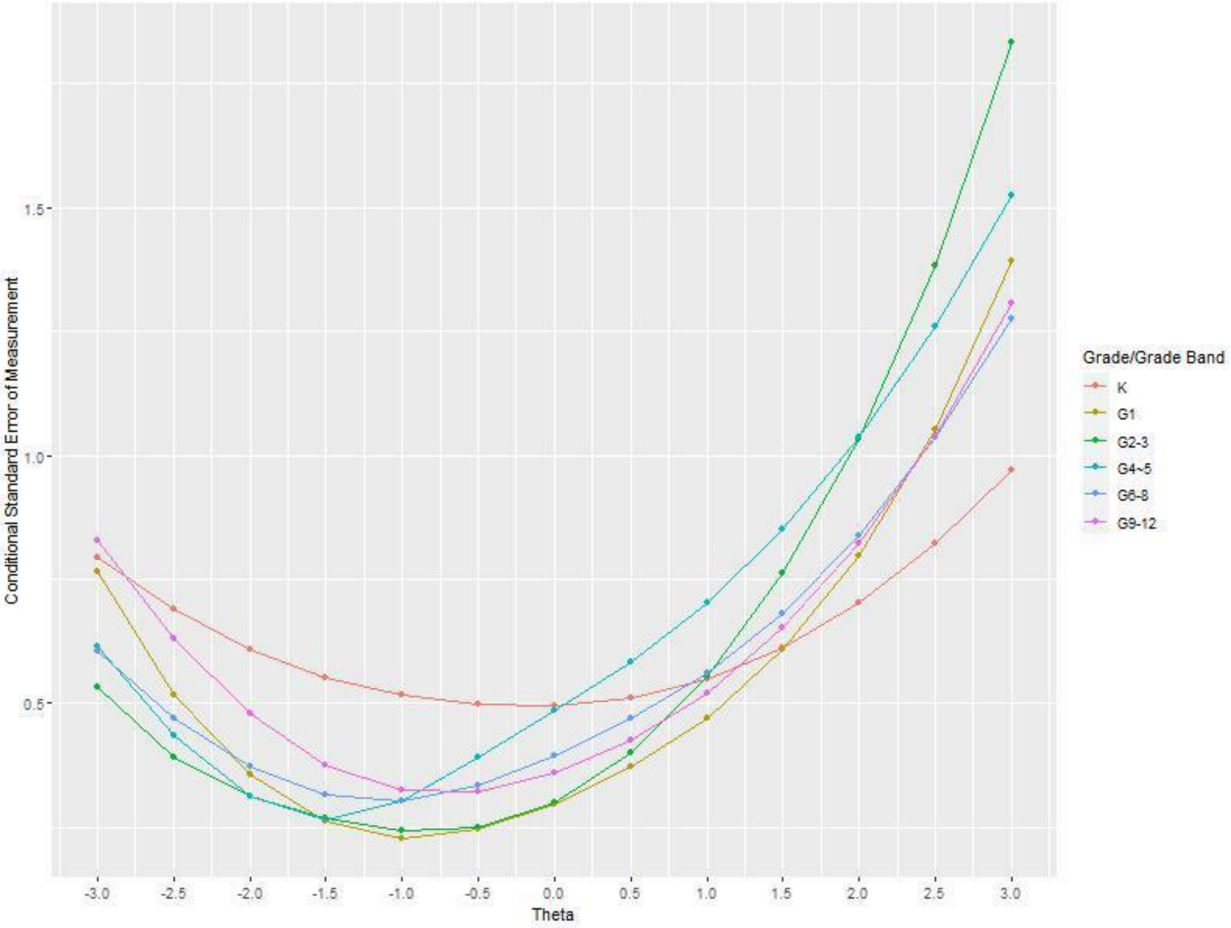
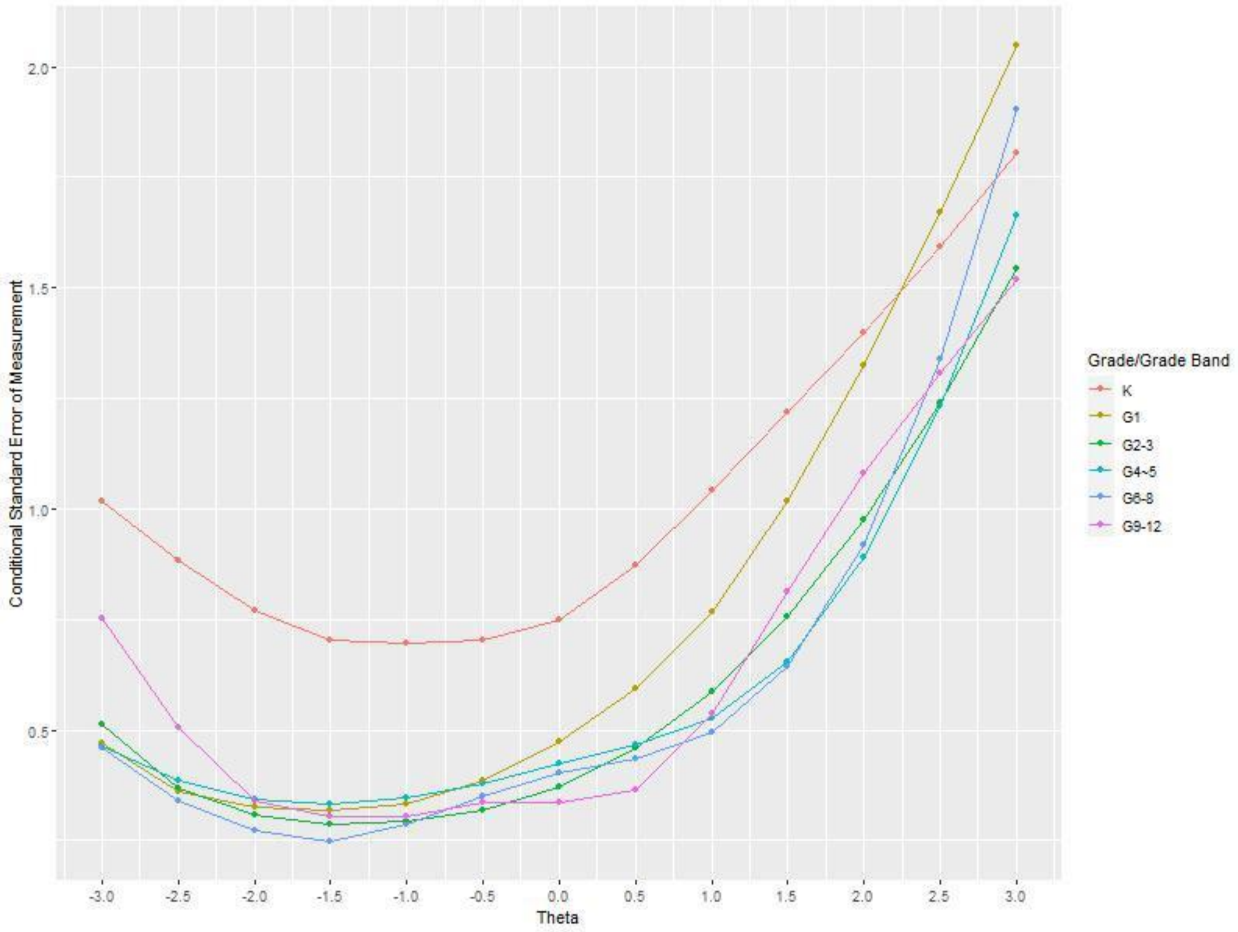
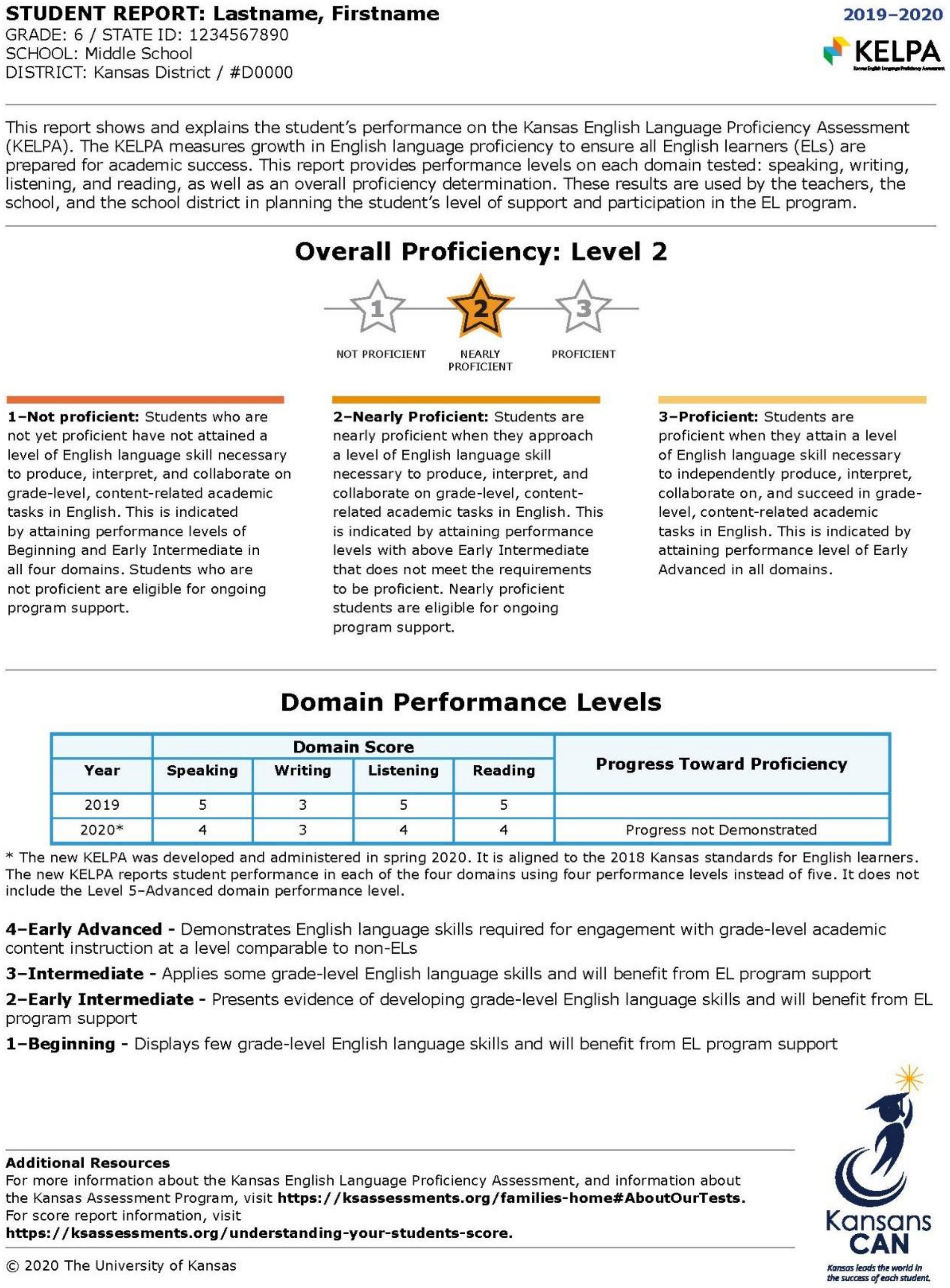


Figure C-4: Conditional Standard Error of Measurement for Writing



Appendix D: Sample KELPA Student Report



STUDENT REPORT: Lastname, Firstname
GRADE: 6 / STATE ID: 1234567890
SCHOOL: Middle School
DISTRICT: Kansas District / #D0000

2019–2020

KELPA

This report shows and explains the student's performance on the Kansas English Language Proficiency Assessment (KELPA). The KELPA measures growth in English language proficiency to ensure all English learners (ELs) are prepared for academic success. This report provides performance levels on each domain tested: speaking, writing, listening, and reading, as well as an overall proficiency determination. These results are used by the teachers, the school, and the school district in planning the student's level of support and participation in the EL program.

Overall Proficiency: Level 2

1 NOT PROFICIENT — 2 NEARLY PROFICIENT — 3 PROFICIENT

1–Not proficient: Students who are not yet proficient have not attained a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels of Beginning and Early Intermediate in all four domains. Students who are not proficient are eligible for ongoing program support.

2–Nearly Proficient: Students are nearly proficient when they approach a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels with above Early Intermediate that does not meet the requirements to be proficient. Nearly proficient students are eligible for ongoing program support.

3–Proficient: Students are proficient when they attain a level of English language skill necessary to independently produce, interpret, collaborate on, and succeed in grade-level, content-related academic tasks in English. This is indicated by attaining performance level of Early Advanced in all domains.

Domain Performance Levels

	Domain Score				Progress Toward Proficiency
Year	Speaking	Writing	Listening	Reading	
2019	5	3	5	5	
2020*	4	3	4	4	Progress not Demonstrated

* The new KELPA was developed and administered in spring 2020. It is aligned to the 2018 Kansas standards for English learners. The new KELPA reports student performance in each of the four domains using four performance levels instead of five. It does not include the Level 5-Advanced domain performance level.

4–Early Advanced - Demonstrates English language skills required for engagement with grade-level academic content instruction at a level comparable to non-ELs

3–Intermediate - Applies some grade-level English language skills and will benefit from EL program support

2–Early Intermediate - Presents evidence of developing grade-level English language skills and will benefit from EL program support

1–Beginning - Displays few grade-level English language skills and will benefit from EL program support

Additional Resources
For more information about the Kansas English Language Proficiency Assessment, and information about the Kansas Assessment Program, visit **https://ksassessments.org/families-home#AboutOurTests.**
For score report information, visit
https://ksassessments.org/understanding-your-students-score.

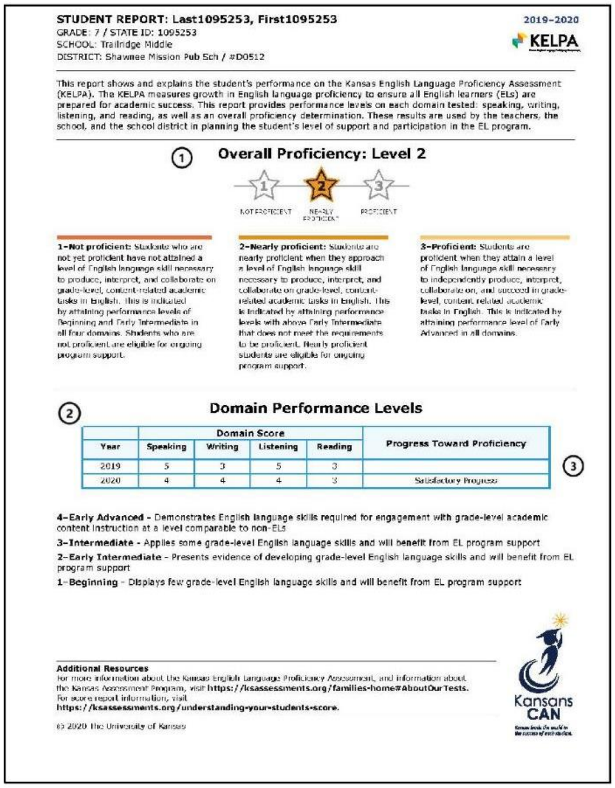
Kansans CAN
Kansas leads the world in the success of each student.

Appendix E: KELPA Educator Guide

KELPA
Kansas English Language Proficiency Assessment

Educator Guide

This score report shows and explains your student's performance on the Kansas English Language Proficiency Assessment (KELPA). Score reports can be accessed under Reports > English Language Learners Assessment in Kite® Educator Portal.



STUDENT REPORT: Last1095253, First1095253 2019–2020
GRADE: 7 / STATE ID: 1095253
SCHOOL: Trailridge Middle
DISTRICT: Shawnee Mission Pub Sch / #D0512

This report shows and explains the student's performance on the Kansas English Language Proficiency Assessment (KELPA). The KELPA measures growth in English language proficiency to ensure all English learners (ELs) are prepared for academic success. This report provides performance levels on each domain tested: speaking, writing, listening, and reading, as well as an overall proficiency determination. These results are used by the teachers, the school, and the school district in planning the student's level of support and participation in the EL program.

① **Overall Proficiency: Level 2**

1–Not proficient: Students who are not yet proficient have not attained a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels of Beginning and Early Intermediate in all four domains. Students who are not proficient are eligible for ongoing program support.

2–Nearly proficient: Students are nearly proficient when they approach a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels with above Early Intermediate that does not meet the requirements to be proficient. Nearly proficient students are eligible for ongoing program support.

3–Proficient: Students are proficient when they attain a level of English language skill necessary to independently produce, interpret, collaborate on, and succeed in grade-level, content related academic tasks in English. This is indicated by attaining performance level of Early Advanced in all domains.

② **Domain Performance Levels**

	Domain Score				
Year	Speaking	Writing	Listening	Reading	Progress Toward Proficiency
2019	5	3	5	3	
2020	4	4	4	3	Satisfactory Progress

③

4–Early Advanced - Demonstrates English language skills required for engagement with grade-level academic content instruction at a level comparable to non-ELs

3–Intermediate - Applies some grade-level English language skills and will benefit from EL program support

2–Early Intermediate - Presents evidence of developing grade-level English language skills and will benefit from EL program support

1–Beginning - Displays few grade-level English language skills and will benefit from EL program support

Additional Resources
For more information about the Kansas English Language Proficiency Assessment, and information about the Kansas Assessment Program, visit **https://ksassessments.org/families-home#AboutOurTests.**
For score report information, visit
https://ksassessments.org/understanding-your-students-score.

© 2020 The University of Kansas

Kansans CAN

① This represents the overall proficiency score for the assessment administration.

② Speaking, writing, listening, and reading domain test scores are used to determine the overall proficiency score.

Students must receive all 4s on the domain scores (speaking, writing, listening, reading) to be considered proficient.

③ Progress Toward Proficiency is determined for each student who did not score proficient. Domain scores for the current year's KELPA assessment are compared to the previous year's KELPA assessment. Students may earn either Satisfactory Progress or Progress Not Demonstrated according to the comparison of the domain scores.

To evaluate student progress, domain performance levels 4 and 5 on the 2019 KELPA2 program are considered level 4 performance on the 2020 KELPA assessment. Students are considered to be making satisfactory progress when they make net progress over the four domains. Net progress is shown by comparing 2019 and 2020 levels.

Consider an example student. On the 2019 KELPA2, the example student performed at levels 2, 3, 3, and 5 in speaking, writing, listening, and reading, respectively. On the 2020 KELPA, the same student performed at levels 3, 3, 3, and 4 in the same domains and the student's status is Satisfactory Progress.

The student's 2019 KELPA2 levels are converted to the equivalent KELPA levels. First, the sum of the 2019 KELPA2 levels is calculated: 2 + 3 + 3 + 4 = 12. Next, the sum of the 2020 KELPA levels is calculated: 3 + 3 + 3 + 4 = 13. So, the example student's status is Satisfactory Progress because the 2020 score is higher than the 2019 score. If the sum of the student's 2020 KELPA levels is equal to or less than that of their 2019 equivalent KELPA levels, the student's status is Progress Not Demonstrated.

If 2019 performance data are not available for a student, the space under Progress Toward Proficiency will be left blank.

Appendix F: KELPA Parent Guide

KELPA
Kansas English Language Proficiency Assessment

Parent Guide

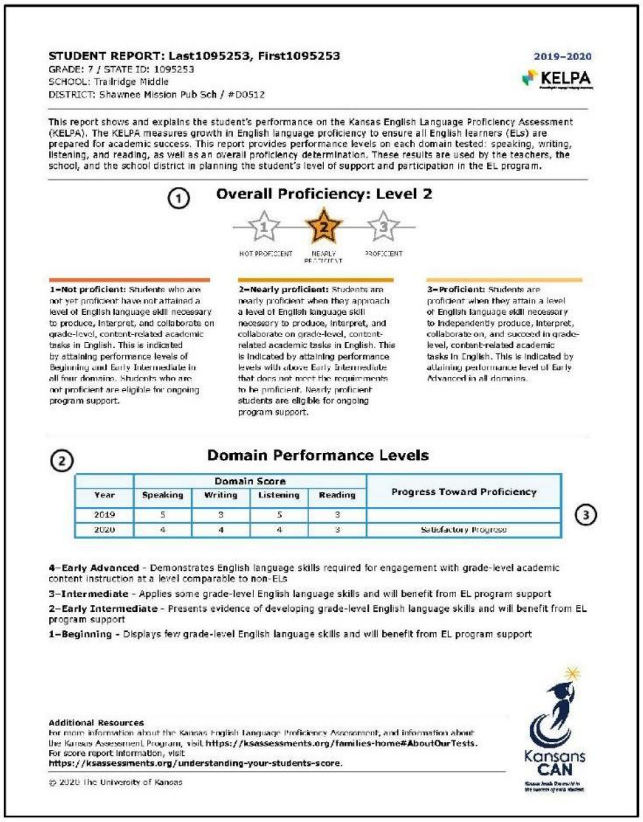
This score report shows and explains your student's performance on the Kansas English Language Proficiency Assessment (KELPA). Current and historic KELPA score reports are available to view in Kite® Parent Portal. Access is managed by your child's school district. Please contact your school district for information on logging in to Kite Parent Portal.

① This represents the overall proficiency score for the assessment administration.

② Speaking, writing, listening, and reading domain test scores are used to determine the overall proficiency score.

Students must receive all 4's on the domain scores (speaking, writing, listening, reading) to be considered proficient.

③ Progress toward proficiency is determined for each student who did not score proficient. Domain scores for the current year's KELPA assessment are compared to previous year's KELPA assessment. Students may earn either satisfactory progress or progress not demonstrated based on comparison of the domain scores. For further information speak with your child's teacher.



STUDENT REPORT: Last1095253, First1095253
GRADE: 7 / STATE ID: 1095253
SCHOOL: Trailridge Middle
DISTRICT: Shawnee Mission Pub Sch / #D0512

2019–2020

KELPA

This report shows and explains the student's performance on the Kansas English Language Proficiency Assessment (KELPA). The KELPA measures growth in English language proficiency to ensure all English learners (ELs) are prepared for academic success. This report provides performance levels on each domain tested: speaking, writing, listening, and reading, as well as an overall proficiency determination. These results are used by the teachers, the school, and the school district in planning the student's level of support and participation in the EL program.

① **Overall Proficiency: Level 2**

1–Not proficient: Students who are not yet proficient have not attained a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels of Beginning and Early Intermediate in all four domains. Students who are not proficient are eligible for ongoing program support.

2–Nearly proficient: Students are nearly proficient when they approach a level of English language skill necessary to produce, interpret, and collaborate on grade-level, content-related academic tasks in English. This is indicated by attaining performance levels with above Early Intermediate that does not meet the requirements to be proficient. Nearly proficient students are eligible for ongoing program support.

3–Proficient: Students are proficient when they attain a level of English language skill necessary to independently produce, interpret, collaborate on, and succeed in grade-level, content-related academic tasks in English. This is indicated by attaining performance level of Early Advanced in all domains.

② **Domain Performance Levels**

	Domain Score				Progress Toward Proficiency
Year	Speaking	Writing	Listening	Reading	
2019	5	3	5	3	
2020	4	4	4	3	Satisfactory Progress

③

4–Early Advanced - Demonstrates English language skills required for engagement with grade-level academic content instruction at a level comparable to non-ELs

3–Intermediate - Applies some grade-level English language skills and will benefit from EL program support

2–Early Intermediate - Presents evidence of developing grade-level English language skills and will benefit from EL program support

1–Beginning - Displays few grade-level English language skills and will benefit from EL program support

Additional Resources
For more information about the Kansas English Language Proficiency Assessment, and information about the Kansas Assessment Program, visit **https://ksassessments.org/families-home#AboutOurTests.**
For score report information, visit
https://ksassessments.org/understanding-your-students-score.

© 2020 The University of Kansas

Kansans CAN