KELPA Technical Manual

October 2021

Table of Contents

Table of Tables

Table of Figures

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 grades and grade 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.

Important Note on the COVID-19 Pandemic

The 2020-2021 academic school year was significantly affected by the COVID-19 pandemic. After complete school and district closures and halting of assessment administration in spring 2020, the reopening of schools in the fall 2020 was characterized by variations of remote, in-person, and hybrid instructional models both within and across states. In many states and districts, the degree to which these instructional models were utilized changed over the course of the school year and was dependent on multiple factors including, COVID-19 case counts, district size, ages of students within schools, local policy, student needs, and parent choice.

Although state and local education agencies made every effort to ensure all students had access to instruction and instructional materials regardless of learning environment, it is well acknowledged that changes to learning inevitably occurred during the 2020–2021 academic year. Recognizing both the variability of instructional access and state and local need for data on student achievement, on February 22, 2021, the U.S. Department of Education, Office of Elementary and Secondary Education provided states with guidance regarding assessment, accountability, and reporting requirements for the 2020–2021 school year. The department's guidance, as it relates to assessments, offered states the option to apply for a one-year waiver from accountability requirements as well as flexibility in assessment administration. The types of flexibility described in the department's letter included: administering shorter versions of state assessments, offering remote administration where feasible, and extending testing windows. The guidance further explained that the focus of this year's assessments is "to provide information to parents, educators, and the public about student performance and to help target resources and supports" (Rosenblum, 2021).

I.1. Overview of English Language 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 English language arts (ELA)—reading, speaking, listening, and writing—are the foundation for the 2018 Standards. The 2018 Standards reflect the continual improvement associated with specific, grade-level ELA standards within these four domains. 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 new KELPA first administrated in 2020. Refer to *2020 KELPA Technical Manual* (Achievement and Assessment Institute [AAI], 2021a) for more details about the 2018

Standards. The 2021 administration was the second administration of KELPA that was aligned with the 2018 Standards.

I.2. Test Purposes and Uses

KELPA is a yearly summative assessment for students in grades K–12 who are identified as not proficient in English, whether or not they receive English for speakers of other languages (ESOL) 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 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 the ESOL services and support that ensure students can acquire 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 (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 ESOL program services.

Beyond understanding common English usage, ELs need to understand the language used for grade-level instruction in ELA, mathematics, science, and social studies. 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 ESOL services, but if the student is identified as an EL, he or she is still required to take KELPA. Detailed information about participation in ESOL services and the KELPA program can be found in ESOL Program Guidance.

When applicable, a student's Individualized Education Program is used to guide accommodations use for KELPA. For more information, refer to the *2020 KELPA Technical Manual* (AAI, 2021a). A detailed summary of accommodations is in Chapter V. Inclusion of All Students in this technical manual.

[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.4. Overview of Technical Manual Updates

A complete technical manual was created for the first year of operational administration in 2020. This technical manual provides updates for the 2021 administration; therefore only sections with updated information are included in this manual. This current chapter recaptures the alignment between KELPA and the 2018 Standards, the purposes of KELPA and its intended population. Chapter II. Assessment System Operations provides updates on KELPA design and development, administration, and test security. Chapter III. Technical Quality—Validity provides validity evidence collected during 2020–2021 school year, i.e., validity evidence based on (a) test content evaluated by an alignment study, (b) relations to other variables evaluated by relationships between KELPA domain scale scores and KAP subject scale scores, and (c) consequences of testing supported by a summary of the 2021 KELPA Teacher Survey (0). Chapter IV. Technical Quality—Other provides updated evidence related to technical qualities, including reliability-related evidence, test-results summary, and ongoing program improvement. Chapter V. Inclusion of All Students provides an updated summary of the accommodations requested in 2021 KELPA administration and information about domain exemption in future KELPA administrations. Chapter 0.

Academic Achievement Standards and Reporting provides the updates about the 2021 KELPA student score report. For a complete description of KELPA, refer to the *2020 KELPA Technical Manual* (AAI, 2021a).

II. Assessment System Operations

This chapter provides updated information about KELPA design and development, administration, and test security. For more details (e.g., monitoring test administration), refer to Chapter II in the *2020 KELPA Technical Manual* (AAI, 2021a).

II.1 Test Design and Development

KELPA, part of the Kansas Assessment Program, is entirely computer based for students in grades 2 through 12. Students in kindergarten and grade 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 University of Kansas's Achievement and Assessment Institute (AAI) worked with the Kansas State Department of Education (KSDE) to determine the content to be assessed by the KELPA tests for each domain and grade or grade band. The developmental milestones leading to the 2020 KELPA test administration can be found in Table II-1 of the *2020 KELPA Technical Manual* (AAI, 2021a). The *2020*

KELPA Technical Manual (AAI, 2021a) also provides detailed information about KELPA test blueprints (i.e., Section II.1.1 Test Blueprints), test design (i.e., Section II.1.2 Test Design), and test construction (i.e., Section II.1.3 Test Construction).

II.2 Content Development

Content development entails various efforts to ensure item quality, including ongoing research into best practices for assessing English learners' 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. Section II.2 Content Development in the *2020 KELPA Technical Manual* (AAI, 2021a) includes detailed descriptions of the typical procedures for different stages of content development:

- Section II.2.1 Passage Development
- Section II.2.2 Item Writing
- Section II.2.3 Item Review

This section provides updated information about the development of the rubric and rater-training materials.

II.2.1 Rubric Development

KELPA rubric development is described in Section II.2.4 in the 2020 KELPA Technical Manual. The same rubrics developed for 2020 administration were used in 2021. Refer to Section II.2.4 Rubric Development in the *2020 KELPA Technical Manual* (AAI, 2021a) for detailed activities of rubric development by phase. To help support rater use of the rubrics in kindergarten and grade 1, a supplementary document was added to the rater-training materials to provide additional, more specific guidance on using the writing rubrics in those grades.

II.2.2 Development of Rater-Training Materials

This section describes the development of updated rater-training materials for the 2021 KELPA administration as well as plans for a staged roll-out in 2022–2023 of prompt-specific exemplar responses for every constructed-response item on the assessment.

II.2.2.1 Materials for 2021 Administration

KELPA rater-training materials were updated for the 2021 administration. Previously, the rater-training materials included the current, holistic rubrics but used items that did not reflect current content of the assessment. Additionally, because some exemplar student responses had been gathered via a small-scale pilot project, they did not include responses at each score point for all prompts. To ensure the training materials provided examples of each score point, to increase relevance, and to better assist educators in scoring operational constructed-response items, all prompts and responses from the 2020 materials were removed and replaced with one set of exemplar student responses for an operational constructed-response (CR) item in each grade or grade band per content domain (i.e., speaking and writing).

Student responses to one operational writing CR item and one speaking CR item in each grade were obtained from the 2020 administration. AAI content-development staff evaluated responses according

to each rubric and then selected three sets of responses per item to utilize in the materials for district and building coordinators to train and calibrate local raters. Those three sets consisted of an anchor set, a practice set, and a calibration set. The anchor set contains three responses for each score point (0–3) on the rubric to identify how the holistic rubrics are applied to a variety of student responses. There are 12 responses in the anchor set for the same item; each anchor-set response is accompanied by an explanation for the assigned score point. Each practice and calibration set has 10 responses for the same item, generally with three responses at score points 3, 2, and 1 and one response at 0. Thus, for each item (or prompt) in each grade or grade band, there are 32 responses: 12 anchor responses and 10 in practice and calibration sets, respectively. Both the calibration and practice sets are intended to help local raters practice; that is, they aid raters in developing an understanding of how to operationalize the rubrics by evaluating student-response examples at each score point.

In December 2020, the rater-training materials went through an external review by Kansas educators and KSDE staff. There were three panels (i.e., kindergarten and grade 1, grades 2–5, grades 6–12) for the review, and each panel looked at both writing and speaking rater-training materials. Two educators served on each panel, as well as two or three KSDE staff members. The panelists asynchronously reviewed all responses (i.e., the anchor, practice, and calibration sets) in their grade or grade-band grouping for both domains and sent their feedback to AAI's content-development staff, including whether they agreed with the assigned score point and whether they felt any revision was needed for the anchor-set explanations. Panels then met for a synchronous discussion of that feedback. Panelists discussed responses that they had rated differently from the rating given in the materials. When the panelists agreed that a response was not suitable for the assigned score point, AAI content-development staff showed (i.e., for writing) or played (i.e., for speaking) other preselected options based on scoring notes from the earlier process of response evaluation for the sets until one of the new responses was determined by the panel to accurately demonstrate the knowledge and skills associated with the score point.

Based on feedback in the synchronous discussions as well as asynchronous feedback on anchor-set explanations, AAI content-development staff made changes to the materials (mainly, replacing responses, revising explanations, reordering anchor-set responses) and finalized the documents for the 2021 administration of KELPA.

II.2.2.2 Materials for 2022 and 2023 Administrations

AAI content-development staff is in process of developing additional sets of rater-training materials so that there will be a prompt-specific set of exemplar responses for every CR item on the assessment. The staged roll-out of those materials will occur in 2022–2023. The 2021 materials contained three sets (anchor, calibration, and practice) for one operational CR item in each grade or grade band in speaking and writing. A validation set of 10 responses will be added for those CR items. The materials will also be expanded to include four sets of example student responses for all operational CR items by the 2023 administration.

The development process for 2022 and 2023 materials will be similar to the process used for the 2021 materials. Content-development staff will select responses for all sets and write explanations for the anchor-set responses. During external reviews, educators will review anchor and calibration sets, and KSDE staff will review all sets. AAI content-development staff will use synchronous and asynchronous feedback to select and determine any needed replacements, which KSDE will review.

II.3 Test Administration and Scoring

The 2021 KELPA testing window was open to students from February 15 through March 31, 2021. Educators were able to enter scores for CR items until April 20, 2021. Additional information about scoring can be found in the *KELPA Scoring Manual*. For an overview of KELPA administration and scoring, refer to the introductory paragraphs of Section II.3 Test Administration and Scoring in the *2020 KELPA Technical Manual* (AAI, 2021a).

Kansas uses a train-the-trainer model in which District Test Coordinators (DTCs) receive training directly from KSDE and, in turn, train educators in their local school districts in test administration and scoring. 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. For more information about this model and training details, refer to Section II.3.1 Test-Administrator and Scorer Training of the *2020 KELPA Technical Manual* (AAI, 2021a). The training webinars, recorded and posted on the DTC Virtual Training Webinars site, are provided and updated every year. The training slides, frequently asked questions, and responses to these questions are also posted on the DTC Virtual Training Webinars site.

The standardized test-administration procedures provided for districts, schools, and teachers are described in the *2020–2021 KELPA Examiner's Manual* (*Examiner's Manual* hereafter). The *Examiner's Manual* also provides guidance and procedures related to administration of KELPA in 2020–2021, for example, procedures and information needed to prepare students and administrators before, during, and after KELPA (Sections 4, 5, and 6, respectively). A summary of these details is in Section II.3.2 Test-Administration Procedures of the *2020 KELPA Technical Manual* (AAI, 2021a).

II.3.1 KELPA Teacher Survey

At the beginning of the KELPA testing window, a teacher survey (see 0) about the 2021 administration was sent to educators via Kansas State Department of Education email distribution lists. At the same time, an announcement about the teacher survey was posted on the Educator Portal. The survey was available until April 26, about a week after the testing window closed. The survey included questions about teachers' background information and their experience with Kite®, scoring, and test administration, as well as students' testing experience and supporting materials (e.g., the 2020–2021 KELPA Examiner's Manual, KELPA Test Administration and Scoring Directions for speaking and writing, etc.); 146 educators (12% of active Educator Portal users who had student(s) rostered to them to take 2021 KELPA) responded to the survey. Tables B-1 through B-13 (see 0) summarize teachers' responses to the survey questions.

The results in Table B-1 show that about half (51%) of the participating educators who responded to the survey were teachers (i.e., classroom, Title 1, special education, EL) who administered KELPA. Many of these educators had 10 or more years of experience in ELA (55%), mathematics (45%), science (36%), and/or with ELs (58%; see Table B-3). They were well spread across different grades or grade bands (see Table B-2).

The percentage of educators who thought it somewhat easy or very easy to use Kite Educator Portal ranged from 18% (i.e., uploading batch student scores, assigning raters [as a DTC]) to 86% (i.e., managing user accounts). The low percentages were because of the fact that not all the tasks in the survey questions applied to all the participants. For example, 74% of the participating educators selected Not Applicable in response to the question about their experience assigning raters as a DTC in the Educator Portal and, therefore, only 18% thought it was somewhat easy or very easy to complete

this task. Refer to Table B-4 for educators' responses regarding user experience of other aspects of Educator Portal. Educators' experience of using Student Portal were better than with Educator Portal. For example, 96% thought it somewhat easy or very easy to submit a completed test. Table B-5 shows details about educators' responses regarding user experience of different aspects of Student Portal. Most educators (61%–72%) agreed or strongly agreed that both the technology practice test and the KELPA practice tests familiarized students and teachers with the technologies, format, and procedures of the real tests (see Table B-6 and Table B-7).

Over 80% of educators agreed or strongly agreed that the training materials for scoring were helpful and the scoring window was sufficient (see Table B-8). Most educators (73% or more) had positive feedback on rater-training workshops (see Table B-9). The majority of educators (84% or more) had positive test-administration experience both in general and with each domain test (see Table B-10 and Table B-11). Most educators agreed or strongly agreed that their students had positive experiences with KELPA (nearly 70% or more; see Table B-12) and that support materials were helpful (76% or more; see Table B-13).

II.4 Test Security

Test security is maintained by protecting the integrity and confidentiality of test materials, test-related data, and personally identifiable information. For a summary of KSDE's plan for ensuring the security and confidentiality of state testing materials, refer to Section II.5 Test Security of the *2020 KELPA Technical Manual* (AAI, 2021a). For more details about security requirements, refer to the *2020–2021 Kansas Assessment Fact Sheet: Test Security and Ethics* and the *Kansas State Department of Education Test Security Guidelines*. Sections II.5.1 through II.5.4 of the *2020 KELPA Technical Manual* (AAI, 2021a) provide detailed information about and requirements for test-materials security, test-related data security, security of personally identifiable information, and accommodations-related security.

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). *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. The test forms in 2021 were the same as the operational forms in 2020; therefore, the evidence from the model calibration and differential item functioning analysis did not need to be updated. For details about validity evidence based on internal structure and other additional evidence, refer to Chapter III Technical Quality — Validity in the *2020 KELPA Technical Manual* (AAI, 2021a). This chapter presents validity evidence collected or evaluated during the 2020–2021 school year.

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. The interpretation and use of KELPA results relies on the correspondence between items and the 2018 Standards, as well as between the test and test blueprint. This section focuses on evidence from the KELPA external alignment study.

The Human Resources Research Organization (HumRRO) conducted independent external study with Kansas educators in spring 2021 to examine the extent of alignment between KELPA, the 2018 Standards, and the academic content standards (Sinclair et al., 2021). The independent study collected information to address six claims:

1. KELPA items are aligned to 2018 Standards.
2. KELPA items represent the 2018 Standards.
3. KELPA meets test blueprints, representing a balanced assessment.
4. KELPA domain-level tests are reliable[2].
5. KELPA includes items representing a range of linguistic difficulty levels.
6. Language proficiency requirements of the academic standards are addressed by the 2018 Standards.

Educators made up panels for the following grades or grade bands: kindergarten, grade 1, grade band 2–3, grade band 4–5, grade band 6–8, and grade band 9–12. There were seven participants in each of the kindergarten, grade 1, and grade band 2–3 panels, six in each of the grade band 4–5 and grade band 9–12 panels, and five in the grade band 6–8 panel. The study consisted of two main parts. The first part of the study, the items-to-standard alignment activity, focused on individual KELPA items that address Claims 1–5. The second part of the study, the standards-correspondence activity, focused on the 2018 Standards and the academic content standards that address Claim 6. Panelists also responded to an evaluation form where they indicated level of agreement with statements relating to each of the two panel activities.

[2] Claim 4 of the alignment study is addressed by domain-level test reliabilities reported in the 2020 KELPA technical manual.

During the items-to-standard alignment activity, panelists were asked to review individual KELPA items and select the 2018 Standard that best aligns with each item. For machine-scored items, panelists were asked to align each item to one standard, and for constructed-response items, panelists were asked to align each item to up to two standards. Panelists rated items as Not Aligned, Partially Aligned, or Fully Aligned to the standards. Panelists were also asked to rate the linguistic difficulty level (LDL) for each item, from Level 1 (least linguistically difficult) to Level 3 (most linguistically difficult; Johnson, 2005). Panelists matched KELPA items to 2018 Standards and rated the LDL for each item independently before discussing ratings as a group. After this initial discussion, panelists viewed the item metadata, which contained the 2018 Standard and the LDL rating assigned to the item, to inform group discussion and consensus. Consensus was defined as agreement by most of the panelists (i.e., five of the six panelists). If consensus was not reached, the majority rating (i.e., four of the six panelists) was used. If half the panelists disagreed and the standard selected by half the panelists matched the standard in the metadata, the standard identified in the metadata would be selected by the facilitator.

During the standards-correspondence activity, panelists first examined language proficiency expectations specified in Kansas's Standards for English Learners Performance Level Rubric (part of the 2018 Standards) to identify the English language skills needed for English learners (ELs) to be able to demonstrate the knowledge and skills reflected in the grade-level academic content standards in English language arts (ELA), mathematics, and science (from Level 0 to Level 5). Because of time constraints, consensus was not required for this part of the alignment study (the most frequently selected panelist ratings were used for data analyses).

The following sections include a brief summary of results from the HumRRO alignment study. For full results, refer to the *Kansas English Language Proficiency Assessment: Alignment Study* (Sinclair et al., 2021). A summary description of plans to address the HumRRO findings is also provided.

III.1.1 Items-to-Standard Alignment Activity Results and Corrective Actions

Claim 1

The criterion applied to Claim 1 was that no items were rated Not Aligned to a standard. This criterion was met. Most items were rated Fully Aligned; however, a few items were rated as Partially Aligned, including grade-1 reading (32%); grade band 2–3 writing (26.3%); grade band 6–8 speaking (22%); and grade band 9–12 listening (41.7%), speaking (30%), and reading (60.9%).

Additionally, the percentage of items for which the primary standard that the panel identified matched the standard in the item metadata ranged from 40% (grade band 9–12 speaking) to 100% (grade band 6–8 speaking).

Claim 2

The criterion applied to Claim 2 was that a minimum of 50% of domain-specific standards (based on the number of standards indicated in the test blueprint) were represented by standard-linked items on domain-level tests. The criterion for Claim 2 was met for most grades or grade bands and domains as the majority of standards were represented by items on most tests. The exceptions were grade band 2–3 listening (43%) and grade band 4–5 reading (40%).

Claim 3

The criterion applied to Claim 3 was "KELPA domain-level tests meet blueprint specifications" (Sinclair et al., 2021). This was determined through comparing the number of score points per cluster (group of standards) based on panel linkage data to the range of score points described on the test blueprint for each cluster. If the panel-determined score points fell within the specified range for a cluster then the criterion for Claim 3 was met in that grade or grade band and domain. Linkages to a secondary standard were only used if linkages to the primary standards did not fall within the specified range. Ultimately, results for Claim 3 were mixed. The criterion was met in all four domains for grade band 6–8. The criterion was met for all domains except reading in grade band 9–12. Two of the four domains met the criterion in grade 1 and grade bands 2–3 and 4–5.

Claim 4

The internal consistency reliability estimates (i.e., coefficient alpha) for each domain are reported for each grade/grade band in the *2020 KELPA Technical Manual* (AAI, 2021a). The criterion for meeting Claim 4 is that the internal consistency reliability estimates for each domain are .70 or above. Reliability coefficients are all above .70 indicating acceptable levels of reliability across domains for all grade/grade band assessments.

Claim 5

The criterion applied to Claim 5 was that each domain had items identified at all LDLs and that more than 50% of the items were identified at level 2 or higher. There were no items identified at LDL 1 in eight of the 24 domain and grade or grade band combinations: speaking in kindergarten, grade 1, and grade bands 6–8 and 9–12; reading in grade 1 and grade band 9–12; and writing in grade bands 6–8 and 9–12 (in four of these instances, however, there were no items at LDL 1 in the metadata). Thus, a portion of the criterion for Claim 5 was not met. However, the highest proportion of LDL 1 was 28%, thus meeting the criterion that 50% or more of the items be identified at LDL 2 or higher.

Additionally, panel LDL ratings overall were in agreement with the LDL ratings in the metadata, although there were some divergences. The data revealed larger divergences in grade-1 and grade band 2–3 speaking (40% of items were rated at an LDL lower than the metadata), grade-1 listening, grade band 2–3 writing, and grade band 9–12 reading (40% of items were rated at an LDL higher than the metadata).

Table III-1 presents overall findings from the items-to-standard alignment activity. Additionally, as indicated by the evaluation form results, panelists tended to agree that "items assess the depth and breadth of the KELP Standards across all proficiency levels."

Table III-1. Alignment Criteria Results for Claims 1–3 and 5 for Grade or Grade-Band Tests by Domain

Grade or grade band	Domain	Claim 1: KELPA items are aligned to KELP standards	Claim 2: KELPA items represent KELP standards	Claim 3: KELPA meets test blueprint, representing a balanced assessment	Claim 5: KELPA domain-level tests include a range of LDLs
Kindergarten	Listening	Met	Met	Not met	Met
	Speaking	Met	Met	Not met	Not met
	Reading	Met	Met	Not met	Met
	Writing	Met	Met	Met	Met
1	Listening	Met	Met	Met	Met
	Speaking	Met	Met	Not met	Not met
	Reading	Met	Met	Not met	Not met
	Writing	Met	Met	Met	Met
2–3	Listening	Met	Not met	Not met	Met
	Speaking	Met	Met	Met	Met
	Reading	Met	Met	Not met	Met
	Writing	Met	Met	Not met	Met
4–5	Listening	Met	Met	Met	Met
	Speaking	Met	Met	Not met	Met
	Reading	Met	Not met	Met	Met
	Writing	Met	Met	Not met	Met
6–8	Listening	Met	Met	Met	Met
	Speaking	Met	Met	Met	Not met
	Reading	Met	Met	Met	Met
	Writing	Met	Met	Met	Not met
9–12	Listening	Met	Met	Met	Met
	Speaking	Met	Met	Met	Not met
	Reading	Met	Met	Not met	Not met
	Writing	Met	Met	Met	Not met

Note. KELP = Kansas English language proficiency. Adapted from Sinclair et al., 2021.

III.1.2 Standards-Correspondence Activity Results and Corrective Actions

Claim 6

The criterion for Claim 6 was that at least 70% of the academic content standards in ELA, mathematics, and science were rated at requiring a language proficiency of Level 4 or lower. The criterion was met for all grade/grade bands and academic content areas with the exception of grade-1 mathematics (where 42% of the standards were rated as requiring Level 5).

Table III-2 presents overall findings from the standards-correspondence activity. Additionally, as indicated by the evaluation form results, panelists tended to agree that students with Level 4 language skills could demonstrate their achievement on the academic content standards.

Table III-2. Alignment Criterion Results for Claim 6 for Grade or Grade-Band Tests by Academic Content Area

Grade or grade band	English language arts	Mathematics	Science
Kindergarten	Met	Met	Met
1	Met	Not met	Met
2–3	Met	Met	Met
4–5	Met	Met	Met
6–8	Met	Met	Met
9–12	Met	Met	Met

Note. Adapted from Sinclair et al., 2021.

III.1.3 Summary of Next Steps

After considering the findings and recommendations from the alignment study along with the intended test design, a few next steps are planned and summarized below. For more information about follow-up analyses that were conducted and details about the next steps for each of the claims, see the response memo (0).

III.1.3.1 Claim 1

Evaluate the current metadata and alignment-study panelist ratings of items to standards for all items that did not receive a Fully Aligned rating and update metadata as needed. Make decisions about whether some more-complex items should be aligned to dual standards (primary and secondary alignments).

III.1.3.2 Claim 2

Analyze cluster-level content coverage of blueprints if any metadata is revised. The HumRRO alignment study was designed to evaluate coverage of all domain-specific standards and did not consider that KELPA blueprints were developed from clusters of standards. Therefore, after any update to metadata, cluster-level content coverage of blueprint will be evaluated.

III.1.3.3 Claim 3

Reanalyze test blueprint coverage will be reanalyzed taking into account potential dual alignments to standards for selected items (follow-up action for Claim 1). Per HumRRO recommendations, blueprints will be presented as a proportion of items instead of score-point ranges.

III.1.3.4 Claim 5

No further action will be taken for operational assessment. Given the request in the field to shorten the test length (i.e., in comparison to the previous version of the test) and the need to maximize to test information at the proficiency cut score, adding LDL 1 items to the test is not consistent with the intended test design and purpose.

III.1.3.5 Claim 6

Establish criterion to review grade-1 mathematics standards and correspondence between standards and performance levels of ELs. Convene educators to review language demands of grade-1 mathematics standards and correspondence between these standards and performance levels of ELs.

III.2 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). This kind of evidence refers to external evidence. Three types of external evidence are convergent, discriminant, and criterion related (either predictive or concurrent). Convergent evidence is provided by relationships between students' performance on different assessments measuring similar constructs. Discriminant evidence is provided by relationships between students' performance on different assessments measuring different constructs. Criterion-related evidence is provided by relationships between students' test scores on one test and those on another test of a related attribute (Cronbach, 1971; Messick, 1989).

The external assessments used in this study are the KAP ELA and mathematics assessments, which are administered annually to students in grades 3–8 and 10, as well as the KAP science assessment, which is administered annually to students in grades 5, 8 and 11. The Pearson product-moment correlations between KELPA domain scale scores and KAP ELA, mathematics, or science scale scores can provide validity evidence based on relations to other variables. The effect size is considered small if a correlation coefficient is less than .30, large if equal to or greater than .50, and medium if in between (Cohen, 1988). Relationships between KAP-subject scale scores and KELPA-domain scale scores were examined because ELs' proficiency in each KELPA domain may have a different impact on their performance in the grade-level academic tests.

Table III-3 presents correlation coefficients between KELPA domain scores and KAP ELA scores. The strongest correlations were between KAP ELA and the KELPA reading domain, ranging from .58 (grade 10) to .67 (grade 3); the weakest correlations were observed between ELA and the speaking domain, ranging from .24 (grade 6) to .33 (grade 3). Correlation coefficients between KAP ELA and KELPA speaking domain across grades were small (with the exception of grade 3); medium to large coefficients were seen between KAP ELA and the other KELPA domains. For relationships between KAP ELA and KELPA listening, reading, and writing, medium to large correlation coefficients were found across grades.

Table III-3. Correlations Between KELPA Domain Scores and KAP English Language Arts (ELA) Scores by Grade

Grade	Correlation between KAP ELA and			
	Listening	Speaking	Reading	Writing
3	.49	.33	.67	.61
4	.57	.28	.66	.62
5	.51	.27	.65	.53
6	.56	.24	.64	.48
7	.55	.28	.63	.49
8	.58	.28	.64	.49
10	.42	.27	.58	.49

Table III-4 present correlations between KELPA domain scores and KAP mathematics scores. Compared to relationships with KAP ELA, relationships between KELPA domain scores and KAP mathematics scores were weaker in all domains. The strongest correlation was between KAP mathematics and KELPA reading domain, ranging from .31 (grade 10) to .56 (grade 3); the weakest correlation was between KAP mathematics and KELPA speaking domain, ranging from .13 (grade 10) to .30 (grade 3). Relationships between KAP mathematics and KELPA speaking domain across grades were weak (with the exception of grade 3); weak to medium relationships were seen between KAP mathematics and the other KELPA domains in most grades.

Table III-4. Correlations Between KELPA Domains Scores and KAP Mathematics Scores by Grade

Grade	Correlation between KAP mathematics and			
	Listening	Speaking	Reading	Writing
3	.45	.30	.56	.54
4	.48	.19	.49	.48
5	.38	.17	.43	.39
6	.36	.15	.42	.28
7	.40	.21	.43	.34
8	.38	.19	.39	.31
10	.24	.13	.31	.28

Table III-5 present correlations between KELPA domain scores and KAP science scores. The strongest correlation was between KAP science and speaking, ranging from .45 (grade 8) to .52 (grade 5); the weakest correlation was between science and reading, ranging from .16 (grade 11) to .24 (grade 5). The strength of most relationships between KAP science and KELPA domains across grades was medium.

Table III-5. Correlations Between KELPA Domains Scores and KAP Science Scores by Grade

Grade	Correlation between KAP science and			
	Listening	Speaking	Reading	Writing
5	.51	.52	.24	.43
8	.40	.45	.21	.29
11	.29	.47	.16	.32

III.3 Validity Evidence Based on Consequences of Testing

Details about validity evidence based on consequences of testing are described in Section III.5 in the *2020 KELPA Technical Manual* (AAI, 2021a). An additional piece of evidence base on consequences of testing was collected, using a teacher survey, during the 2021 KELPA administration. Responses to one of the survey questions indicated that 81% (*n* = 118) of the participating educators believed that the content of KELPA measured important English language proficiency knowledge, skills, and abilities. For a complete summary of the teacher survey results, refer to Section II.3.1 KELPA Teacher Survey in the current manual.

IV. Technical Quality—Other

This chapter provides updated evidence related to the technical quality of KELPA administrated in 2021, including reliability-related evidence, a summary of test results, and a description of ongoing program improvement. For technical quality related evidence (e.g., information of fairness and accessibility, full performance continuum, quality-control steps), refer to Section IV.4 Full Performance Continuum in the *2020 KELPA Technical Manual* (AAI, 2021a).

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)

The reliability estimates for KELPA were reported in two ways: reliability coefficients from classical test theory (CTT) and IRT information functions as well as conditional standard error of measurement. CTT reliability coefficients are sample dependent and were updated using the 2021 data. IRT reliability does not change by test sample and only changes by test form. Because same test forms were used in 2021 as in 2022, the IRT reliability is not provided in this section. For the detailed information about the IRT reliability, refer to Section IV.1 Reliability of the *2020 KELPA Technical Manual* (AAI, 2021a). For the CTT reliability coefficients, the student-group reliabilities were also calculated. Indices of classification consistency and accuracy of different domain performance levels and interrater agreement on speaking and writing constructed-response (CR) are also provided in this section.

IV.1.1 Test Reliability

Because KELPA uses only one fixed form for each domain test at each grade 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:

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

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 .79 to .97 across the

majority of grade levels or bands and domains. The exceptions were in kindergarten for reading (.72) and writing (.55). Test length and test reliability are closely related, and shorter tests are usually less reliable. Compared to other domains, writing tests across grades and grade bands had lower reliabilities because these tests had the fewest score points. Table II-13 in the *2020 KELPA Technical Manual* (AAI, 2021a) indicates the test lengths and total score points for all domain tests.

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

Grade or grade band	Listening α	Speaking α	Reading α	Writing α
K	.86	.92	.72	.55
1	.84	.91	.89	.81
2–3	.88	.91	.90	.86
4–5	.88	.93	.82	.83
6–8	.86	.94	.84	.86
9–12	.88	.97	.85	.79

IV.1.1.1 Student-Group Reliability

Reliability estimates were also calculated by student group and are presented in Table IV-2. Results show that the student-group reliabilities were similar within a domain and at most grades or grade bands; the exceptions were kindergartens in speaking and writing (consistent with the domain-level coefficient alphas). Also, the student-group reliabilities were similar to the overall reliabilities, with the majority of the estimates in the .80s to .90s; reading in kindergarten (mostly in the .70 range or lower) and writing in kindergarten (mostly in the .50 range or lower) and grade band 9–12 (mostly in the .70 range) had lower reliabilities. The sample size of each student group can be found in Section IV.2.1.1 Test Enrollment Data of the current document.

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	SWOD
Listening								
Kindergarten	.86	.86	.86	.88	.86	.88	.88	.86
1	.83	.85	.84	.86	.84	.86	.87	.84
2–3	.87	.88	.87	.89	.87	.90	.89	.87
4–5	.87	.89	.87	.90	.88	.90	.88	.88
6–8	.86	.85	.85	.87	.86	.86	.82	.86
9–12	.87	.89	.88	.89	.88	.88	.86	.88
Speaking								
Kindergarten	.92	.91	.91	.92	.91	.92	.93	.91
1	.91	.91	.91	.93	.91	.92	.92	.91
2–3	.91	.91	.91	.92	.91	.93	.92	.91
4–5	.94	.93	.93	.94	.93	.93	.92	.93
6–8	.95	.94	.94	.94	.94	.93	.92	.95
9–12	.97	.97	.97	.97	.97	.97	.96	.97
Reading								
Kindergarten	.72	.72	.68	.79	.66	.81	.71	.72
1	.89	.90	.88	.91	.88	.91	.88	.89
2–3	.90	.90	.90	.90	.90	.90	.89	.90
4–5	.81	.83	.81	.85	.82	.84	.81	.81
6–8	.83	.84	.83	.84	.84	.85	.80	.84
9–12	.84	.86	.85	.86	.85	.87	.82	.85
Writing								
Kindergarten	.54	.56	.52	.62	.51	.64	.56	.55
1	.81	.80	.80	.82	.80	.82	.81	.80
2–3	.86	.86	.86	.87	.86	.87	.85	.85
4–5	.82	.83	.82	.85	.82	.85	.81	.82
6–8	.86	.85	.85	.87	.86	.87	.82	.86
9–12	.80	.78	.79	.80	.79	.81	.75	.80

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

IV.1.2 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.

The detailed descriptions of the calculation of two indexes can be found in Section IV.1.3 Classification Consistency and Accuracy in the *2020 KELPA Technical Manual* (AAI, 2021a). The results for classification consistency and accuracy for three cuts are presented in Table IV-3. 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. Classification consistency indices for the KELPA domain tests ranged from .68 to .98 across the majority of cuts and grades or grand bands. Classification accuracy indices for the KELPA domain tests ranged from .75 to .99 across the majority of cuts and grade levels or bands.

Table IV-3. Classification Consistency (C) and Accuracy (A) 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						
Kindergarten	.93	.95	.91	.94	.77	.83
1	.93	.95	.86	.90	.82	.87
2	.97	.98	.91	.94	.85	.90
3	.98	.99	.95	.96	.89	.92
4	.96	.97	.95	.96	.84	.89
5	.97	.98	.95	.97	.85	.90
6	.95	.97	.93	.95	.81	.86
7	.95	.97	.93	.95	.83	.88
8	.95	.96	.94	.96	.81	.87
9	.94	.96	.92	.94	.87	.91
10	.94	.96	.93	.95	.89	.92
11	.94	.96	.93	.95	.84	.89
12	.94	.96	.93	.95	.83	.88
Speaking						
Kindergarten	.92	.94	.88	.92	.82	.86
1	.96	.97	.91	.94	.80	.85
2	.97	.98	.93	.95	.80	.86
3	.97	.98	.95	.96	.79	.86
4	.97	.98	.95	.97	.86	.90
5	.97	.98	.96	.97	.77	.84
6	.97	.98	.94	.96	.84	.89
7	.97	.98	.95	.96	.83	.88
8	.97	.98	.96	.97	.76	.83
9	.97	.98	.96	.97	.91	.94
10	.97	.98	.96	.97	.92	.94
11	.97	.98	.96	.97	.89	.92
12	.97	.98	.97	.98	.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
Reading						
Kindergarten	.70	.78	.84	.89	.92	.94
1	.86	.90	.88	.91	.92	.94
2	.86	.90	.89	.92	.90	.93
3	.91	.94	.90	.93	.87	.91
4	.91	.94	.83	.88	.81	.86
5	.90	.93	.83	.88	.79	.85
6	.95	.97	.84	.89	.83	.88
7	.92	.94	.85	.89	.80	.86
8	.93	.95	.85	.90	.76	.82
9	.85	.89	.84	.89	.87	.91
10	.86	.90	.85	.89	.84	.89
11	.87	.91	.84	.89	.83	.88
12	.87	.91	.84	.89	.82	.88
Writing						
Kindergarten	.76	.84	.68	.76	.83	.90
1	.92	.95	.83	.88	.76	.81
2	.92	.95	.86	.90	.80	.85
3	.93	.95	.89	.92	.75	.79
4	.94	.95	.90	.93	.76	.83
5	.96	.97	.89	.93	.70	.77
6	.96	.97	.90	.93	.76	.82
7	.96	.97	.87	.91	.74	.80
8	.97	.98	.87	.91	.70	.75
9	.91	.94	.81	.87	.78	.84
10	.91	.94	.83	.88	.79	.85
11	.86	.90	.81	.87	.78	.84
12	.88	.91	.82	.88	.77	.84

Note. Categories 1, 2, 3, and 4 represent proficiency levels 1, 2, 3, and 4.

IV.1.3 Interrater Agreement Study

The purpose of the rater-agreement study is to provide reliability and validity evidence for the educator-scored test items. KELPA CR item scores range from 0 to 3 for both speaking and writing. Refer to Table II-13 in the *2020 KELPA Technical Manual* (AAI, 2020a) for the number of educator-scored items for speaking and writing by grade or grade band. Holistic, instead of item-specific, rubrics within the same grade or grade band in each domain of speaking and writing were used to rate CR item responses. The rater training and training materials provided educators with the knowledge and skills needed to apply the rubrics. The scoring accuracy of CR items, which are scored by educators, rely on consistent and appropriate application of the scoring rubrics. Therefore, it is worthwhile to evaluate if teachers were applying the rubrics consistently, which can help identify further improvements to raining materials, and examine how much raters agreed or disagreed with each other on their ratings for each of the CR items.

IV.1.3.1 Data Collection Method

An interrater agreement study on KELPA writing and speaking CR items was conducted during the 2021 KELPA scoring window (February 15–April 20, 2021). Two methods were used to collect second ratings: Kite Educator Portal or a spreadsheet for targeted school districts. The Kite Educator Portal method was used for individual raters to enter the scores and the spreadsheet option was used for school districts to batch enter information for a roster of students. Students selected for second ratings had two scoring tabs in Educator Portal for all CR items to allow two scorers to enter scores for the same student response. Scores of record for operational scoring remained the same (i.e., the first score entered; see Educator Scoring in the *2020 KELPA Technical Manual* [AAI, 2021a] for more information about how scores were entered.). District Test Coordinators were responsible for monitoring the process for collecting second ratings from selected educators in their district. Table IV-4 shows available scoring methods for both first and second raters in speaking and writing.

Table IV-4. Available Scoring Methods for Speaking and Writing

Writing	Speaking	
Individual scoring	Individual scoring	Deferred scoring
		Simultaneous scoring
Paired/group scoring	Paired/group scoring	Deferred scoring
		Simultaneous scoring

In addition to the second scores, information collected through the user interface of Educator Portal also included:

- Scoring method for the first rating: Users may select individual (i.e., scoring items individually) or paired/group (i.e., scoring items in pairs or a small group) scoring.
- Speaking scoring options for the first rating: Users may select simultaneous (i.e., scoring items in the moment that students are responding) or deferred (i.e., scoring items later by listening to the recordings) scoring.
- Designated scorer for the first rating: Default to user logged in; users may change name of scorer if scored by another user.
- Scoring method for the second rating: Users may select individual or paired/group scoring.
- Speaking scoring options for the second rating: Users may select simultaneous or deferred scoring.
- Designated scorer for the second rating: Default to user logged in; users may change name of scorer if scored by another user.

IV.1.3.2 Sampling

A sample of approximately 10% of students taking KELPA in each school district for the 2021 administration was selected to receive second ratings for their speaking and writing CR items. Samples to have two ratings were identified at the very beginning of the testing window when all school districts completed KELPA test registration. School districts with more than 10 EL students at a grade or grade band is eligible to be selected to receive two ratings with a target sample size of approximately 500 students per grade. Random sample of 14% of registered kindergarten and grade-1 students were

selected. Random sample of 11% of registered students in grades 2–12 were selected. Table IV-5 shows the number of districts, number of schools and number of students selected for two ratings.

Table IV-5. Number of Districts, Schools, and Students Selected for Two Ratings

Grade or grade band	No. of districts	No. of schools	No. of students
Kindergarten	45	209	535
1	46	210	541
2–3	58	275	1,054
4–5	55	254	845
6–8	59	160	986
9–12	69	113	1,372

Data obtained at the end of the window for hand scoring speaking and writing items were used for rater-agreement analyses. Valid item level scores are used for analyses [3]. Only a very small percentage (0%–4%) of responses with two ratings were collected using the paired/group scoring method for both writing and speaking. For speaking responses scored individually, 0%–3% of these responses were simultaneously scored. Sample sizes, both for paired/group scoring in writing and speaking and simultaneous scoring for speaking, were not sufficient to make meaningful statistical inferences. Therefore, Table IV-6 shows the number of student responses per item using the individual scoring method for writing and the number of student responses per item using the combination of individual and deferred scoring methods for speaking.

Table IV-6. Number of Students With Two Ratings by Domain and Grade or Grade Band

Grade or grade band	Number of student responses per item	
	Writing: Individual scoring	Speaking: Combination of individual and deferred scorings
Kindergarten	393–395	317–323
1	433–435	368–375
2–3	630–642	506–515
4–5	539–540	438–449
6–8	620–623	541–549
9–12	663–664	596–606

IV.1.3.3 Raters

KELPA constructed responses are scored by qualified educators. District test coordinators (DTC) assigned qualified educators within a school district to score KELPA constructed-response items in speaking and writing. Students assigned to receive two ratings were rated by DTC assigned educators that are different from raters who are responsible in providing rating for the primary score. There were no difference in training and assigning educators for second ratings from those who provided first ratings. Refer to Section II.3.1 Test-Administrator and Scorer Training and Section IV.3.1.2 Educator

[3] Any blank response with a score of 0 was excluded from the data for analyses to avoid inflating rater agreements.

Scoring in the *2020 KELPA Technical Manual* (AAI, 2021a) for details about rater training and assignment.

IV.1.3.4 Interrater Agreement

IV.1.3.4.1 Methods

Agreement measures how frequently two raters assign the exact same rating (Graham et al., 2012). The percentage of items on which raters agree exactly is referred to as *exact agreement*; the percentage of items on which raters agree either exactly or within one point of one another is referred to as *adjacent agreement*. An exact agreement level of 75% or above is acceptable, and exact plus adjacent agreements should be 90% or above (Graham et al., 2012). Kappa originally measured the agreement between two raters on a two-level (e.g., pass vs fail) rating scale but can also measure the agreement when three or more performance levels are used. Weighted kappa distinguishes between the numbers of ratings falling within one performance level and the numbers of ratings that differ by two or more performance levels (Graham et al., 2012). The quadratic-weighted kappa is calculated between the expected scores and the predicted scores and measures the agreement between two ratings; the value typically ranges from 0 (random agreement between raters) to 1 (complete agreement between raters). When there is less agreement between raters than expected by chance, the value may go below 0. Suppose rater A assigns a sample of *n* subjects across the *m* categories of a categorical scale and suppose rater B independently does the same thing. Equation IV-2 shows how the mean observed degree of disagreement is calculated and Equation IV-3 shows how the mean degree of disagreement expected by chance (i.e., expected if A and B assign subjects randomly in accordance with their respective base rates) is calculated (Fleiss & Cohen, 1973).

$$\overline{D}_o = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij} v_{ij}, \tag{IV-2}$$

$$\overline{D}_e = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{i\cdot} n_{\cdot j}\, v_{ij}\,, \tag{IV-3}$$

where n_{ij} denotes the number of subjects assigned to category *i* by rater A and to category *j* by rater B; $n_{i\cdot}$ denotes the total number of subjects assigned to category *i* by rater A and $n_{\cdot j}$ denotes the total number of subjects assigned to category *j* by rater B; v_{ij} denotes the disagreement weight associated with categories *i* and *j*.

When v_{ij} = 0, it reflects no disagreement when a subject is assigned to category *i* by both raters; when v_{ij} > 0, for $i \neq j$, it reflects some degree of disagreement when a subject is assigned to different categories by the two raters. Weighted kappa is then defined by Equation IV-4 (Fleiss & Cohen, 1973):

$$\mathrm{k}_w = \frac{\overline{\mathrm{D}}_e - \overline{D}_o}{\overline{D}_e}. \tag{IV-4}$$

Kappa is a special case of weighted kappa when v_{ij} = 1 for all $i \neq j$. The quadratic weight emphasizes the importance of near disagreement and drops quickly when there are two or more category differences. A kappa value greater than .75 indicates excellent agreement, a value less than .40 indicates poor agreement, and any value between .40 and .75 indicates good agreement ("Weighed kappa in R," n.d.).

IV.1.3.4.2 Results

Table IV-7 summarizes rater agreement for writing items. For writing responses, the average percentage of exact agreement across items within grade or grade band—both overall (i.e., mean percentage of agreement on all responses no matter what scoring method was applied) and for the individual scoring method—ranged from 60% (grade band 6–8) to 84% (kindergarten and grade 1). The average percentage of exact plus adjacent agreement across items within grade or grade band—both overall and for the individual scoring method—was 96% or above.

Table IV-7. Rater Agreement on Writing Items Scored Using the Individual Scoring Method

Grade or grade band	Mean exact agreement across items (%)		Mean exact plus adjacent agreement across items (%)	
	Overall	Individual scoring	Overall	Individual scoring
K	84	84	96	96
1	84	83	98	98
2–3	76	76	98	98
4–5	70	69	97	97
6–8	60	60	98	98
9–12	64	64	97	97

Table IV-8 summarizes agreement for speaking items. For speaking responses, the average percentage of exact agreement across items within grade or grade band—for overall (i.e., mean percentage of agreement on all responses no matter what scoring method was applied), the individual scoring method ,and the combination of individual and deferred scoring method—ranged from 61% (kindergarten) to 71% (grade band 2–3). The average percentage of exact plus adjacent agreement across items within grade or grade band—for overall, the individual scoring method, and the combination of individual and deferred scoring methods—was 94% or above.

Table IV-8. Rater Agreement on Speaking Items

Grade or grade band	Mean exact agreement across items (%)			Sum of mean exact plus adjacent agreement across items (%)		
	Overall	Individual scoring	Individual x deferred	Overall	Individual scoring	Individual x deferred
K	63	61	63	95	95	96
1	67	67	68	96	96	97
2–3	71	71	71	98	98	98
4–5	70	70	70	97	97	97
6–8	66	66	66	97	97	97
9–12	69	69	69	94	94	94

Note. Individual x deferred = combination of individual and deferred scoring methods.

Table IV-9 shows the classifications of quadratic-weighted kappa values of KELPA CR items. To keep consistent with Table IV-5, Table IV-6, and Table IV-7, the number of items for excellent or good agreement reported in Table IV-8 is based on responses scored using the individual scoring method for writing item and the combination of individual and deferred scoring methods for speaking items. Quadratic kappa results show that all items had good to excellent agreement. Excellent agreement was

found for responses to kindergarten and grades 1–3 writing items. For both speaking and writing, lower grades (i.e., kindergarten through grade 3) had better agreement than higher grades. The only exception was that all grade 9–12 speaking items had excellent agreement.

Table IV-9. Summary of Quadratic Kappa Classifications

	No. of items (% of domain items)			
	Writing		Speaking	
Grade or grade band	Excellent agreement	Good agreement	Excellent agreement	Good agreement
K	2 (100)	0 (0)	4 (40)	6 (60)
1	4 (100)	0 (0)	6 (60)	4 (40)
2–3	4 (100)	0 (0)	6 (60)	4 (40)
4–5	1 (25)	3 (75)	1 (10)	9 (90)
6–8	0 (0)	3 (100)	3 (33)	6 (67)
9–12	1 (33)	2 (67)	10 (100)	0 (0)

IV.1.3.4.3 Summary

Individual scoring was the dominant scoring method for both writing and speaking items in 2021. Deferred scoring was the dominant scoring method for speaking. Likely, because of the COVID-19, only a small proportion of responses were scored in pairs or a small group. To summarize, the average percentage of exact agreement between two raters across items within grade or grade band ranged from 60% to 84% for writing responses and from 61% to 71% for speaking responses. The average percentage of exact plus adjacent agreement across items within grade or grade band was 96% or above for writing responses and 94% or above for speaking responses. Statistics of the quadratic-weighted kappa show that, for writing responses, raters had excellent agreement on lower grades items (kindergarten through grade band 2–3) and a mixture of good to excellent agreement on upper grades items; for speaking responses, raters had a mixture of good to excellent agreement on items from kindergarten through grade band 6–8. The exception is that raters had excellent agreement on all the grade band 9–12 speaking items.

IV.2 Scoring and Scaling

This section provides test-result summaries for 2021 administration. For information about the procedures of scoring individual items, scoring the test as a whole, scaling, and specific quality-control process followed by AAI and the Agile Technology Solutions to ensure the accuracy of scoring results, refer to Section IV.3.5 Quality-Control Checks of the *2020 KELPA Technical Manual* (AAI, 2021a).

IV.2.1 Operational Test Results

The number of students who took KELPA in 2021, 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.2.1.1 Test Enrollment Data

All students who are identified as ELs must take KELPA[4]. 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 Kansas State Department of Education (KSDE)-approved EL screener to determine whether KELPA is required. A potential EL who did not pass the screener is considered an EL and will take KELPA in the spring. Students who scored as Proficient on KELPA in 2021 are not required to take KELPA again in the next school year.

KELPA was administered in the four domains: listening, speaking, reading, and writing. Students who took the tests were in grades K–12. Students who have viewed a listening or reading test, even if they did not answer any questions, are categorized as having taken the domain test. Students who either viewed a speaking or writing test or not, but whose tests were scored by teachers, are categorized as having taken the domain test, even if they did not answer any items. Students who took at least one domain test received a score report. Table IV-20 in Section IV.2.2.1 Comparison of Enrollment Rates in this current manual presents the number and percentage of enrolled students who were tested in each grade for the 2020 and 2021 KELPA administrations.

The participation rates for the 10 State Board of Education (SBOE) districts in 2021 are presented in Table IV-10 by grade or grade band. Kansas has 286 school districts, which were separated into 10 SBOE districts. The tested rates ranged from 55% (SBOE district 8 in grade band 9–12) to 99% (SBOE District 5 in multiple grades and grade bands and SBOE district 6 in grade band 4–5). The tested rates were lower in grade band 9–12 across all SBOE districts than in other grades and grade bands. The two largest school districts are the Kansas City public school district (part of board district 1, whose average tested rate was 90% across grades and grade bands) and the Wichita public school district (part of district 7, whose average tested rate was 87% across grades and grade bands). Both are from SBOE districts that had very high participation rates in elementary school but decreased participation rates in middle and high schools. The decreased participation rates in higher grades in these two SBOE districts are consistent with the dramatic enrollment drop from 2020 to 2021 in grades 8–11 reported in Table IV-20, indicating that the two largest school districts experienced a significant impact on both enrollment and participation rates.

[4] During the 2021 administration, students were allowed to opt out of KELPA because of the pandemic.

Table IV-10. KELPA Participation Rates by Grade or Grade Band and Board District in 2021

Board district	Kindergarten		Grade 1		Grade band 2–3		Grade band 4–5		Grade band 6–8		Grade band 9–12	
	Enrolled students (*n*)	Tested students (%)	Enrolled students (*n*)	Tested students (%)	Enrolled students (*n*)	Tested students (%)	Enrolled students (*n*)	Tested students (%)	Enrolled students (*n*)	Tested students (%)	Enrolled students (*n*)	Tested students (%)
1	1,187	95	1,254	94	2,364	95	1,834	95	2,204	89	2,615	72
2	728	89	728	91	1,283	92	920	93	943	92	1,204	84
3	664	88	668	91	1,171	91	849	91	860	92	1,070	82
4	203	88	210	87	431	84	339	82	423	75	537	65
5	1,060	99	1,033	99	1948	99	1,690	98	2043	99	2,529	95
6	193	97	187	97	309	98	243	99	239	97	231	97
7	988	95	1,055	95	1,899	95	1,559	94	1,887	82	2,639	59
8	787	94	822	95	1,542	94	1,301	93	1,564	80	2,295	55
9	148	99	160	98	315	98	202	98	239	97	252	93
10	912	95	955	95	1,773	94	1,451	93	1,753	81	2,495	58

For all tested ELs, Table IV-11 shows the percentage of students in each demographic group by grade[5]. The groups include race, ethnicity, disability status, and gender. The percentage of students in each student group was 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.

[5]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. Percentage of Tested Students by Demographic Group and Grade

Characteristic	Grade (%)												
	K (*n* = 4,090)	1 (*n* = 4,212)	2 (*n* = 4,119)	3 (*n* = 3,730)	4 (*n* = 3,359)	5 (*n* = 2,889)	6 (*n* = 2,452)	7 (*n* = 2,310)	8 (*n* = 2,207)	9 (*n* = 2,092)	10 (*n* = 1,996)	11 (*n* = 1,780)	12 (*n* = 1,361)
Race													
Black	4.8	4.6	4.3	4.4	4.6	5.0	4.6	4.3	6.0	4.4	4.4	5.1	4.8
American Indian	6.1	6.6	7.9	7.1	7.7	8.8	9.8	10.9	11.4	13	16.5	18.3	21.0
Asian	10.3	10.5	10.5	9.8	8.6	8.1	6.4	7.2	6.3	6.4	7.0	7.9	9.5
NHPI	1.0	1.4	1.4	1.2	1.2	1.2	1.1	1.1	0.9	0.5	1.1	0.9	1.0
White	77.9	77	76	77.5	77.9	76.9	78.1	76.6	75.4	75.7	70.9	67.8	63.7
Hispanic													
Yes	79.6	79.2	80.1	81.8	82.9	83.3	84.9	84.3	83.7	85.7	86.2	85.1	81.8
No	20.4	20.8	19.9	18.2	17.1	16.7	15.1	15.7	16.3	14.3	13.8	14.9	18.2
SWD													
Yes	11.3	10.6	12.8	14.8	16.2	18.9	21.6	21.4	18.1	18.5	14.4	15.5	13.1
No	88.7	89.4	87.2	85.2	83.8	81.1	78.4	78.6	81.9	81.5	85.6	84.5	86.9
Gender													
Female	47.1	47.8	47.7	47.0	44.6	44.0	44.0	42.1	43.8	42.3	41.7	44.9	44.6
Male	52.9	52.2	52.3	53.0	55.4	56.0	56.0	57.9	56.2	57.7	58.3	55.1	55.4

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

IV.2.1.2 Test Results for All Students

Summaries of scale scores by grade and domain are presented in Table IV-12, Table IV-13, Table IV-14, and Table IV-15. As the tables show, the minimum and maximum values were within the lowest obtainable scale score (LOSS) (i.e., 0) and the highest obtainable scale score (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 distribution of scale scores was negatively skewed in grades 3–5, 7, and 10, and it was positively skewed in other grades; the distribution was symmetric in grade 9. For the speaking test, the distribution of scale scores was symmetric or nearly symmetric in grades 1 and 8; distributions for other grades were skewed negatively. For the reading test, the distribution of scale scores was negatively skewed in grades 7–8 and 11–12 and positively skewed in other grades. For the writing test, the distribution of scale scores was approximately symmetric in grades 7–8; positively skewed in grades 1, 4, 6, 10, and 12; and negatively skewed in other grades.

Table IV-12. Scale-Score Descriptive Statistics for Listening by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	526.19	168.45	0	354	421	492	589	695	1,000
1	492.89	131.36	0	335	410	470	552	648	1,000
2	489.32	166.43	0	328	391	453	541	605	1,000
3	576.50	208.04	0	378	453	541	605	1,000	1,000
4	500.01	165.99	0	337	411	491	535	611	1,000
5	550.34	193.46	0	362	432	535	611	1,000	1,000
6	468.98	112.38	0	347	414	453	510	615	1,000
7	507.63	133.26	0	358	432	510	552	725	1,000
8	536.92	156.93	0	358	453	510	615	725	1,000
9	494.75	147.45	0	338	407	477	547	622	1,000
10	511.48	160.26	0	338	421	506	547	622	1,000
11	536.72	176.72	0	360	437	506	622	622	1,000
12	541.07	179.71	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-13. Scale-Score Descriptive Statistics for Speaking by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	490.63	146.03	0	345	434	514	563	630	1,000
1	518.52	162.00	0	365	447	511	575	638	1,000
2	509.68	165.39	0	364	434	500	550	616	1,000
3	554.07	189.26	0	398	459	531	575	1,000	1,000
4	532.19	198.92	0	366	435	502	542	1,000	1,000
5	558.76	214.84	0	366	460	520	577	1,000	1,000
6	502.25	188.69	0	349	429	490	533	582	1,000
7	532.80	218.59	0	346	441	503	552	1,000	1,000
8	546.41	222.34	0	357	453	517	582	1,000	1,000
9	522.52	256.30	0	331	429	502	556	1,000	1,000
10	541.48	268.68	0	331	429	502	556	1,000	1,000
11	553.07	273.47	0	345	439	502	556	1,000	1,000
12	564.97	285.19	0	331	448	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-14. Scale-Score Descriptive Statistics for Reading by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	493.52	135.05	0	363	399	463	552	656	1,000
1	479.28	124.81	0	362	393	439	548	648	1,000
2	462.82	125.17	0	334	379	441	515	603	1,000
3	534.25	160.74	0	365	429	515	603	671	1,000
4	479.70	122.68	131	344	388	465	557	602	1,000
5	518.31	137.55	0	358	422	491	602	665	1,000
6	475.59	112.32	0	355	407	463	541	628	1,000
7	508.40	127.21	0	355	424	511	579	699	1,000
8	539.00	138.05	0	372	443	541	628	699	1,000
9	474.68	106.53	0	359	409	469	542	594	1,000
10	497.42	119.89	150	359	409	485	566	631	1,000
11	510.01	125.09	0	359	424	502	566	631	1,000
12	525.99	133.64	0	377	439	521	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-15. Scale-Score Descriptive Statistics for Writing by Grade

Grade	*M*	*SD*	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	503.76	183.54	0	310	407	502	568	676	1,000
1	494.94	163.64	0	321	400	464	588	691	1,000
2	463.13	124.75	0	314	381	465	523	622	1,000
3	520.84	135.37	0	355	449	523	580	687	1,000
4	482.99	121.21	85	335	418	479	563	600	1,000
5	521.75	131.66	0	367	457	532	600	649	1,000
6	482.62	123.98	0	340	410	471	557	596	1,000
7	512.53	150.72	0	353	428	496	557	652	1,000
8	546.56	169.52	0	366	448	525	596	652	1,000
9	472.97	111.88	0	337	407	486	530	585	1,000
10	488.02	121.05	0	337	421	486	554	632	1,000
11	507.52	129.42	0	355	425	508	554	632	1,000
12	525.20	137.40	0	372	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[6] (i.e., Levels 1 through 4) is provided by domain and grade in Figure IV-1, Figure IV-2, Figure IV-3, and Figure IV-4. Students must obtain 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 and grade 1) to 71% (grade 3) across grades for listening, from 20% (kindergarten) to 52% (grade 4) across grades for speaking, from 12% (kindergarten) to 40% (grade 2) across grades for reading, and from 10% (kindergarten) to 49% (grade 10) across grades for writing.

[6] Refer to Section IV.2 Achievement Standard Setting of the 2020 KELPA Technical Manual for the KELPA performance level setting process.

Figure IV-1. Performance-Level Results for Listening

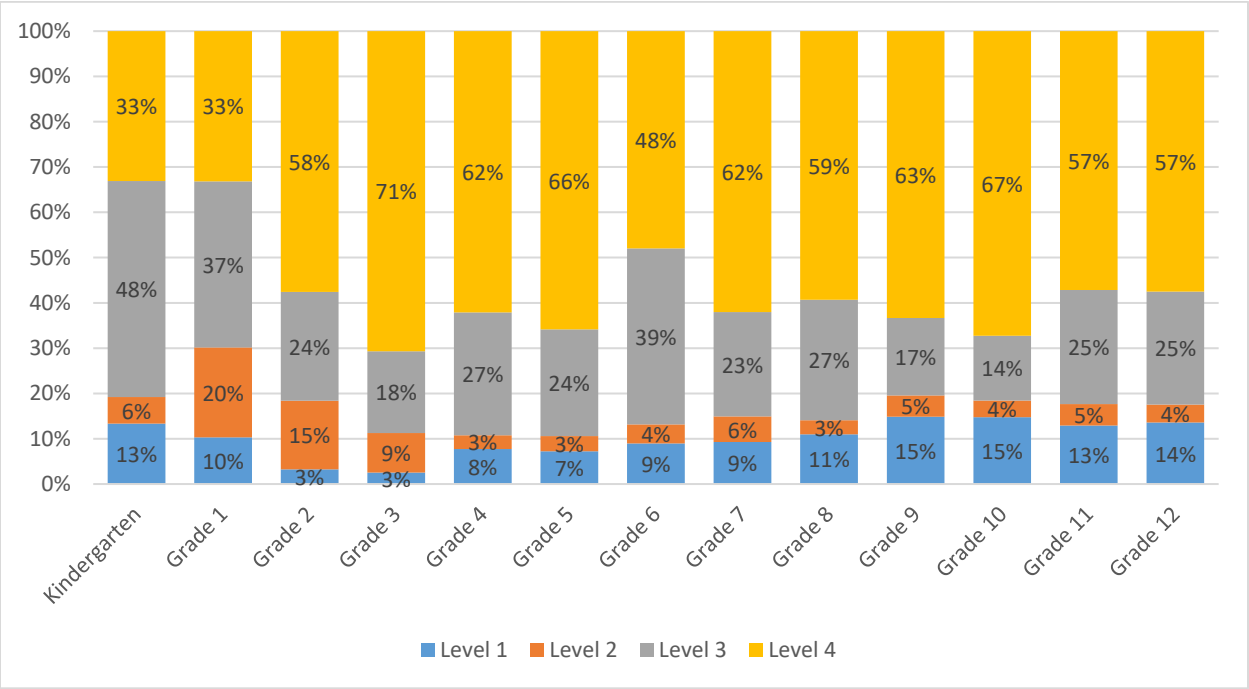


Figure IV-2. Performance-Level Results for Speaking

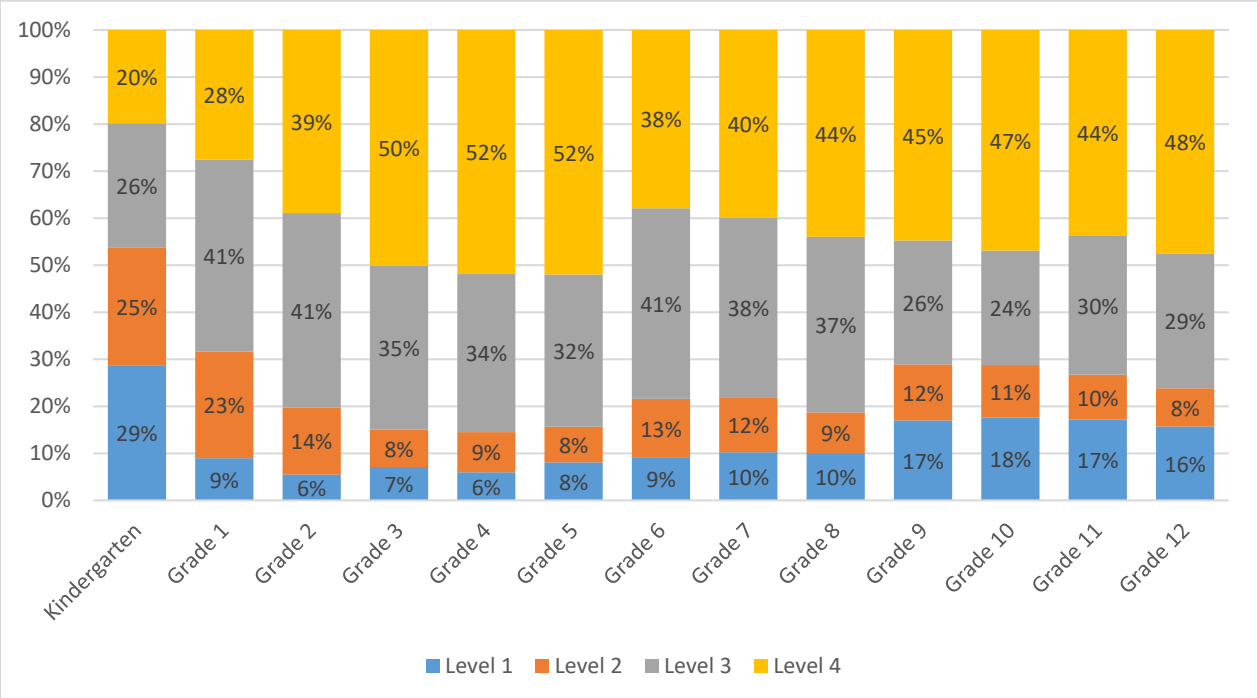


Figure IV-3. Performance-Level Results for Reading

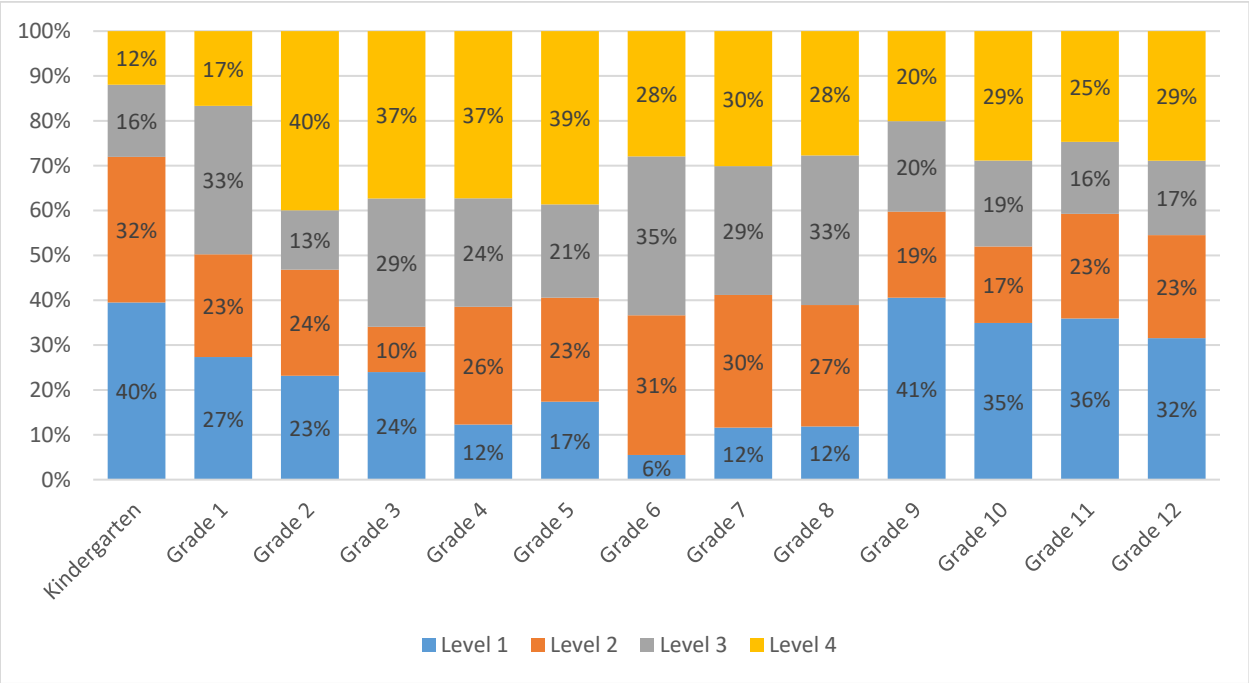
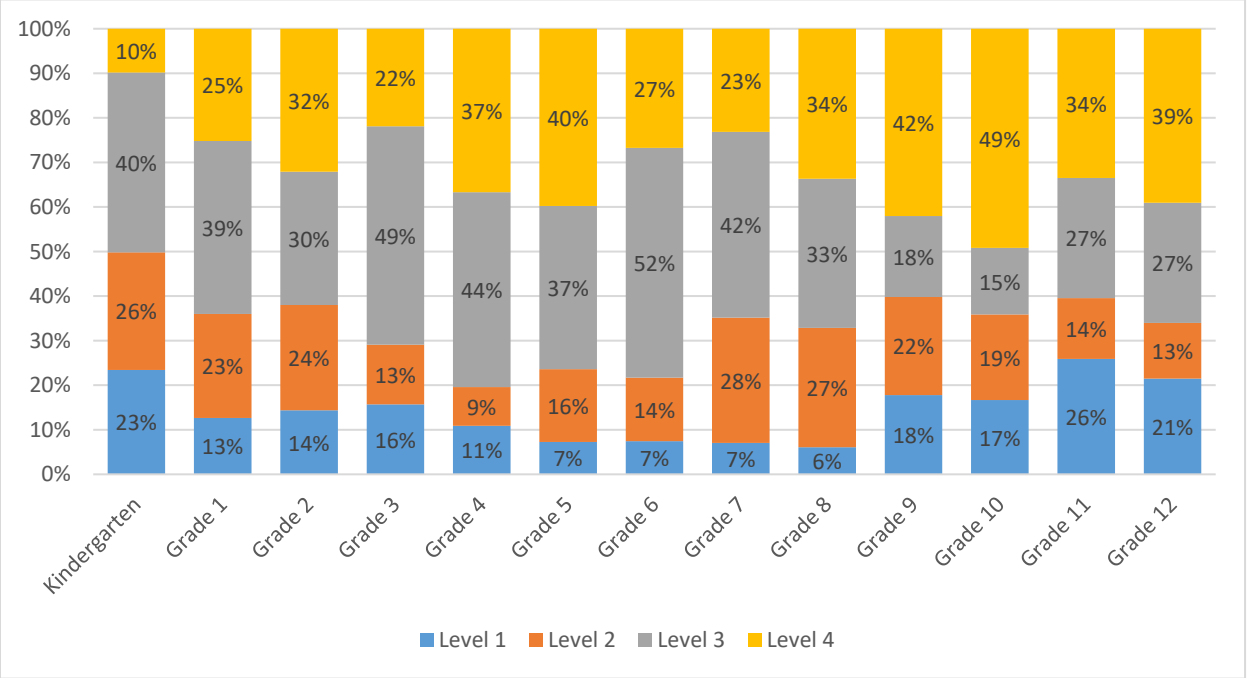


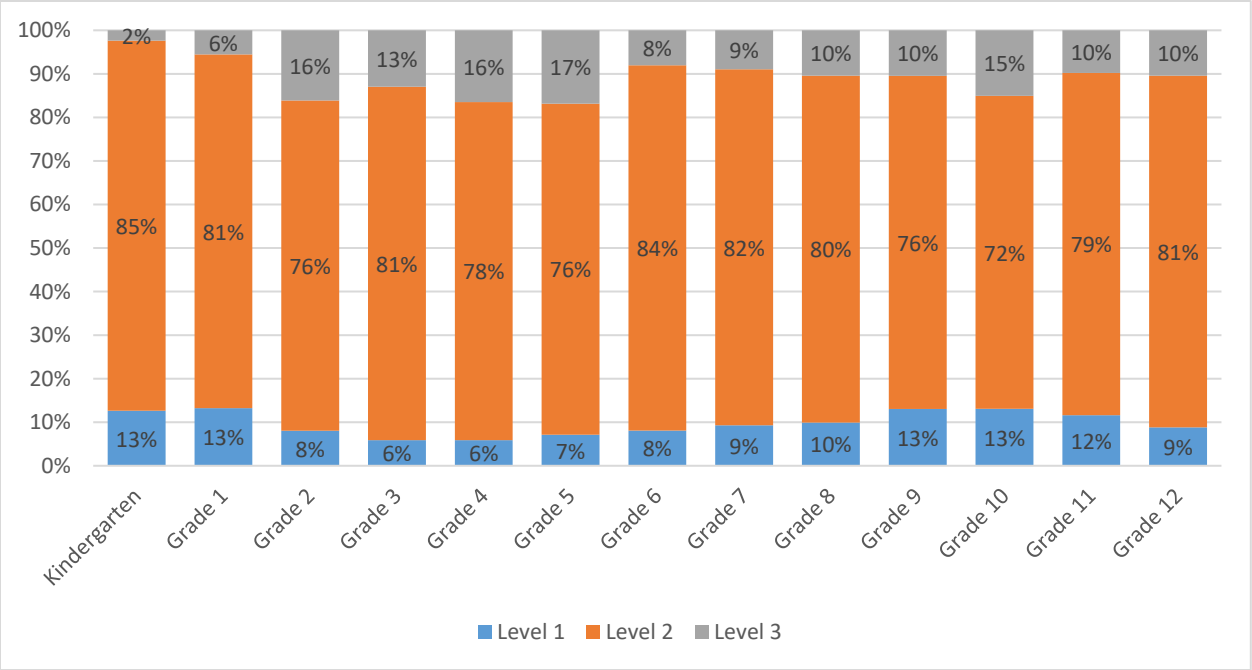
Figure IV-4. Performance-Level Results for Writing



The overall proficiency levels are determined from 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 is 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 in 2021 are presented in Figure IV-5. Results indicate that most students were categorized as Level 2; the percentages ranged from 72% (grade 10) to 85% (kindergarten). Overall, the proficiency rates ranged from 2% (kindergarten) to 17% (grade 5). Kindergarten and grade 1 had lower percentages of students in Level 3, compared to other grades which is expected and consistent with results in previous years given that students in early grades have had little exposure to formal instruction or ESOL services.

Figure IV-5. Overall Performance-Level Results



IV.2.1.3 Student-Group Test Results

Summaries of average scale scores by demographic groups [7] are presented in Table IV-16, Table IV-17, Table IV-18, and Table IV-19. For group sample sizes, refer to Table IV-11 . In most grades and domains, Asian students had the highest mean scores. However, NHPI students had the highest mean score for the listening test in grades 9, 10, and 11. For the speaking test, Black students had the highest mean score in kindergarten, White students had the highest mean score in grade 4, and NHPI students had the highest mean score in grades 10 and 12. For the reading test, NHPI students had the highest mean score in grades 9 and 10, and White students had the highest mean score in grade 11. NHPI students had the highest mean score for the grade-9 writing test. Across all domains, the mean scores of non-Hispanic students were higher than those of Hispanic students in most grades and were slightly lower in some grades (i.e., grades 8–9 in listening, grade 4 in speaking, grades 8 and 10–12 in reading, and grades 8 and 10 in writing). Across all domains and grades, the mean scores of students without a disability were slightly higher than those of students with a disability. For speaking and writing tests, the mean scores of female students were higher than those of male students in all grades. For the listening test, the mean scores of female students were higher than those of male students in most grades, except for grades 4

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

and 5. For the writing test, the mean scores of male students were slightly higher than those of female students in most grades. These findings are similar to 2020 findings. Even when a test is carefully constructed with many considerations on fairness, differences may exit among student groups as a result of achievement gaps. Trend data comparing both the overall test results and results in each domain from 2020 to 2021 are provided in the following subsection to monitor any changes across years.

Table IV-16. Demographic Group Scale-Score Descriptive Statistics for Listening by Grade

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	513	162	478	125	473	145	579	211	491	173	545	206	463	116	481	127	531	170	486	138	514	160	534	172	531	168
Asian	536	179	510	138	512	191	608	223	527	194	561	206	488	108	546	143	558	152	505	129	509	142	547	174	576	182
Black	518	177	469	132	475	167	564	229	468	159	533	196	458	122	480	150	476	146	451	134	491	187	528	195	507	174
NHPI	507	131	444	124	446	160	496	165	484	179	535	221	424	82	461	118	502	161	595	216	551	211	569	236	557	209
White	525	166	493	129	489	165	573	204	498	160	552	191	468	112	509	130	539	155	496	149	511	158	537	176	542	183
Hispanic																										
Yes	523	166	490	128	485	161	572	203	498	164	550	192	468	111	504	130	539	157	495	148	511	160	535	173	537	178
No	539	179	505	145	507	187	597	227	510	176	552	200	477	118	526	147	527	155	493	146	516	164	549	195	558	189
SWD																										
Yes	469	157	438	128	427	148	508	178	447	149	496	165	434	92	474	104	499	124	464	122	463	110	489	142	483	152
No	534	169	499	130	498	167	588	211	510	167	563	197	479	116	517	139	545	162	502	152	520	166	546	181	550	182
Gender																										
Female	540	170	511	134	502	174	584	211	494	153	538	181	479	113	509	133	539	158	502	151	512	157	550	181	551	179
Male	514	167	476	127	478	158	570	206	505	175	560	203	461	112	507	134	535	156	490	144	511	163	526	172	533	180

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

Table IV-17. Demographic Group Scale-Score Descriptive Statistics for Speaking by Grade

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	467	155	479	158	494	174	538	191	496	185	523	209	468	174	489	206	515	219	534	265	508	246	530	267	515	307
Asian	490	173	540	184	529	200	569	205	523	197	573	233	543	202	548	203	568	239	560	242	596	281	584	286	599	237
Black	507	136	527	187	526	188	533	186	521	228	550	218	493	181	533	225	504	186	478	211	474	269	485	248	563	287
NHPI	464	192	467	138	498	154	493	127	514	159	495	170	437	180	499	244	547	181	545	164	624	250	558	337	614	304
White	491	142	520	158	508	160	555	188	537	198	563	214	505	191	536	220	552	225	518	256	547	271	560	273	574	281
Hispanic																										
Yes	487	143	514	157	504	157	553	187	533	198	559	212	500	190	527	219	545	223	522	260	537	266	552	272	558	292
No	505	158	534	180	533	193	558	198	527	205	560	227	513	179	563	215	554	221	527	236	570	281	561	281	595	254
SWD																										
Yes	421	168	448	143	443	148	501	166	484	173	514	183	478	158	514	174	545	204	494	235	531	219	536	228	526	262
No	499	141	527	162	519	166	563	191	542	202	569	221	509	196	538	229	547	226	529	261	543	276	556	281	571	288
Gender																										
Female	506	147	546	172	526	175	577	202	553	210	578	228	514	205	547	227	555	233	540	259	560	274	566	272	586	282
Male	477	144	494	149	495	155	534	175	515	188	544	203	494	174	523	212	540	213	510	254	528	264	543	274	548	287

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

Table IV-18. Demographic Group Scale-Score Descriptive Statistics for Reading by Grade

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	472	120	458	111	446	110	534	148	472	123	515	147	467	102	487	120	537	153	474	107	505	127	511	127	521	117
Asian	567	182	560	155	509	155	584	174	503	140	544	162	504	127	559	136	562	115	497	122	500	120	512	127	537	137
Black	509	156	488	133	467	122	507	151	478	148	494	141	461	131	472	119	486	149	439	114	451	118	468	130	445	104
NHPI	500	99	458	94	447	117	518	166	460	90	493	140	448	112	473	111	543	159	566	108	525	126	439	88	490	100
White	484	123	470	117	458	122	529	159	478	119	517	132	475	111	509	127	539	135	474	104	496	117	513	121	531	136
Hispanic																										
Yes	479	120	465	113	455	118	527	156	476	120	515	133	474	109	506	126	540	137	474	104	498	120	511	122	529	133
No	550	171	534	150	496	145	566	176	496	134	537	158	487	129	523	131	533	142	479	120	493	119	505	140	512	135
SWD																										
Yes	459	133	437	109	409	104	460	129	419	103	456	110	422	88	459	107	485	106	442	96	456	105	466	95	472	110
No	498	135	484	126	471	126	547	162	491	123	533	139	490	114	522	129	551	141	482	107	504	121	518	128	534	135
Gender																										
Female	497	133	484	125	469	131	540	163	479	119	514	135	481	108	508	122	540	139	474	98	492	112	509	117	524	128
Male	491	137	475	124	457	119	529	159	480	125	522	139	472	115	509	131	538	137	476	112	502	125	511	132	527	138

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

Table IV-19. Demographic Group Scale-Score Descriptive Statistics for Writing by Grade

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	474	175	487	172	454	116	524	142	465	122	521	153	474	106	489	147	535	180	471	107	495	124	507	128	519	118
Asian	598	227	578	196	520	147	585	161	516	136	548	146	501	118	549	149	580	166	513	126	505	128	533	137	579	172
Black	529	177	503	197	459	131	496	134	459	133	502	150	447	109	471	127	483	156	410	96	423	100	455	122	472	117
NHPI	449	165	478	168	466	128	509	163	480	82	504	172	441	110	501	186	523	161	569	120	502	111	502	102	495	100
White	492	175	485	153	457	121	513	129	481	117	520	124	485	128	515	152	549	168	472	109	488	120	508	129	523	139
Hispanic																										
Yes	486	170	480	151	453	119	513	130	478	118	519	126	482	123	508	149	547	168	472	109	489	121	507	128	521	133
No	575	215	553	195	503	139	556	154	505	134	537	157	489	132	535	158	542	175	478	127	484	120	511	138	543	156
SWD																										
Yes	449	183	422	135	397	121	451	125	417	103	457	112	433	93	466	114	499	129	447	92	457	101	470	106	461	111
No	511	183	504	165	473	122	533	134	496	120	537	132	496	128	525	157	557	176	479	115	493	123	515	132	535	139
Gender																										
Female	511	179	509	169	470	124	530	138	499	123	537	137	503	133	533	163	570	186	493	115	508	129	532	134	557	147
Male	498	187	482	158	457	126	513	133	470	118	509	126	467	114	498	140	529	153	458	107	474	113	488	122	500	124

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

IV.2.2 Trend Data

The 2021 KELPA administration was the second administration of the new KELPA aligned with the 2018 Standards. The subsection presents changes in enrollment data and performance-level distributions from 2020 to 2021.

IV.2.2.1 Comparison of Enrollment Rates

Because of the impact of the COVID-19 pandemic on the 2020–2021 academic school year, the enrollment and test participation rates decreased in each grade from 2020 to 2021 (see Table IV-20). For the 2021 administration, 40,834 students were enrolled and 36,597 students tested; the overall tested rate was 90%. The tested rates across grades ranged from 65% (grade 12) to 95% (grades 1–5). Compared to the 2020 administration, both total enrollment and participation rates for all grades decreased in the 2021 administration, likely because students were allowed to opt out of testing because of the COVID-19 pandemic[8]. The total enrollment rate dropped by 8% from 2020 to 2021; the largest decreases were in grades 9, 10, and 11 (18%, 20%, and 16%, respectively).

Table IV-20. Number and Percentage of Enrolled and Tested Students by Grade: 2020 vs. 2021

Grade	2020			2021			Enrollment drop %
	No. enrolled	No. tested	Participation %	No. enrolled	No. tested	Participation %	
K	4,614	4,522	98	4,305	4,090	95	7
1	4,619	4,573	99	4,434	4,212	95	4
2	4,734	4,734	100	4,336	4,119	95	8
3	4,051	4,051	100	3,926	3,730	95	3
4	3,829	3,791	99	3,536	3,359	95	8
5	3,242	3,210	99	3,041	2,889	95	6
6	2,809	2,809	100	2,724	2,452	90	3
7	2,663	2,636	99	2,538	2,310	91	5
8	2,755	2,727	99	2,480	2,207	89	10
9	3,110	3,079	99	2,551	2,092	82	18
10	3,129	3,066	98	2,495	1,996	80	20
11	2,830	2,773	98	2,373	1,780	75	16
12	2,179	2,092	96	2,094	1,361	65	4
Total	44,564	44,063	99	40,834	36,597	90	8

IV.2.2.2 Comparison of Performance-Level Results

Figure IV-6, Figure IV-7, Figure IV-8, and Figure IV-9 show the proportion of students in each performance level in 2020 and 2021 by domain and grade. From 2020 to 2021, for listening, the Level 4 percentages stayed the same in kindergarten and grade 5 but decreased in most grades. For speaking, the Level 4 percentages stayed the same in kindergarten, increased slightly in grades 1, 2, 3, 5, and 12, and decreased slightly in other grades from 2020 to 2021. For reading, the Level 4 percentages stayed the same or approximately the same in grades 8 and 12 and decreased slightly in the other grades. For

[8] A special circumstances code called SC19 was established in 2021 to capture COVID exemption.

writing, the Level 4 percentages decreased in all grades except grade 12, where it increased by 1%. In most grades and domains, the percent of students in level 4 decreased from 2020 to 2021.

Figure IV-6. Comparison of 2020 and 2021 Performance-Level (PL) Results for Listening

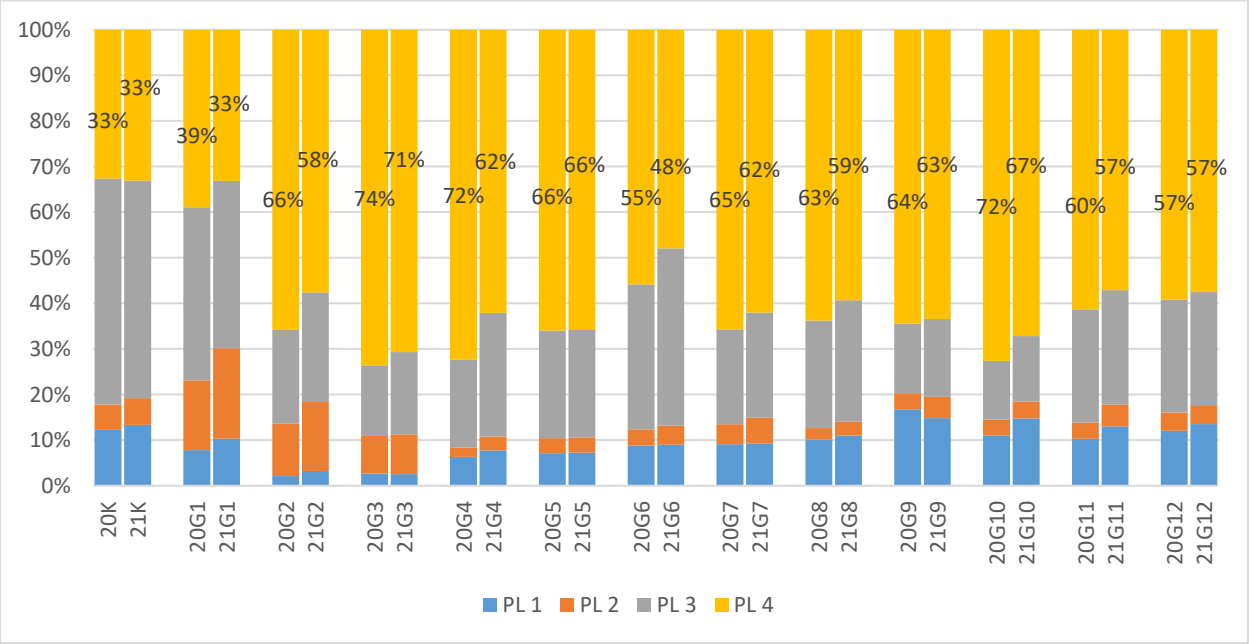


Figure IV-7. Comparison of 2020 and 2021 Performance-Level (PL) Results for Speaking

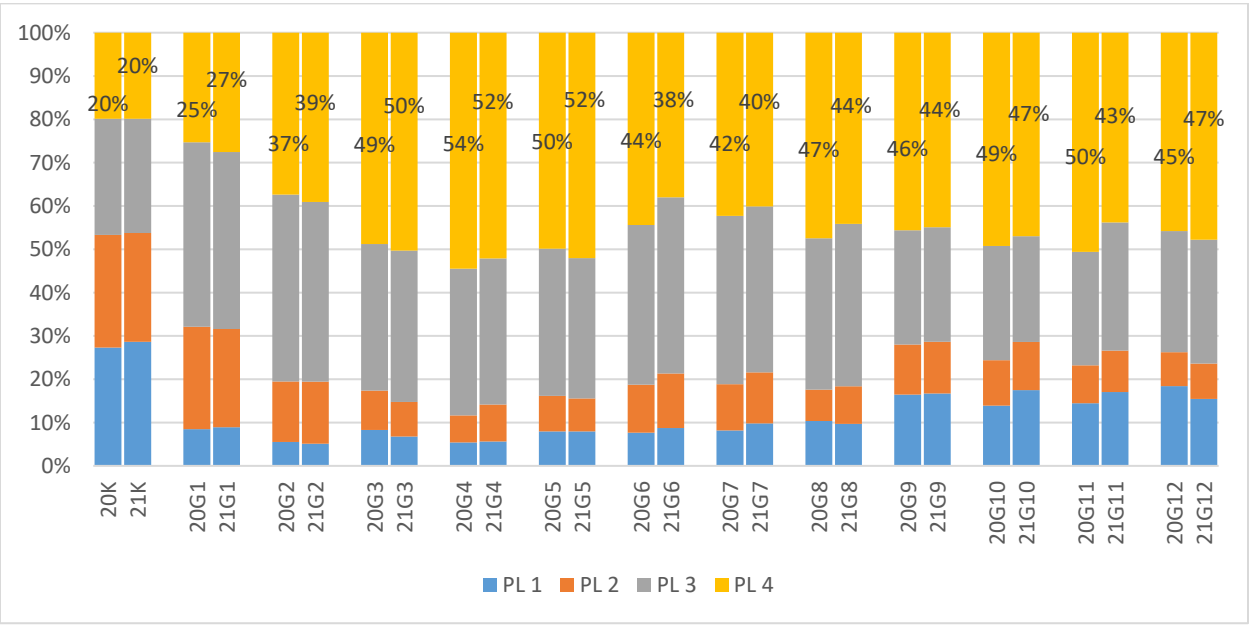


Figure IV-8. Comparison of 2020 and 2021 Performance-Level (PL) Results for Reading

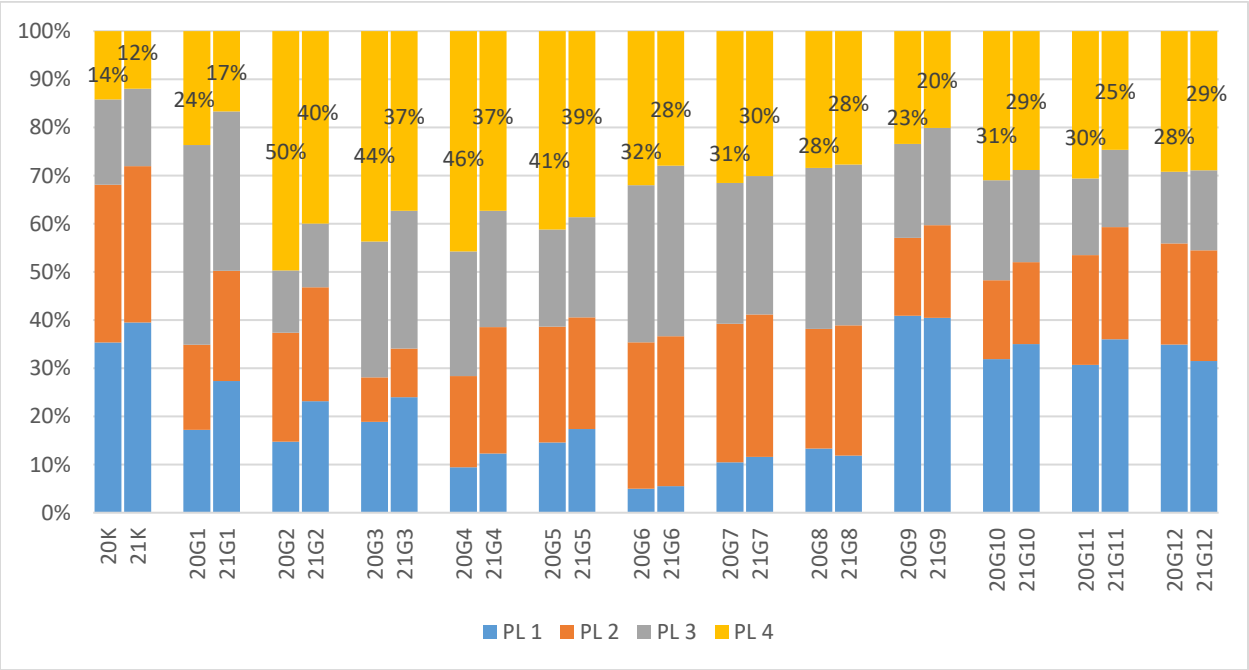
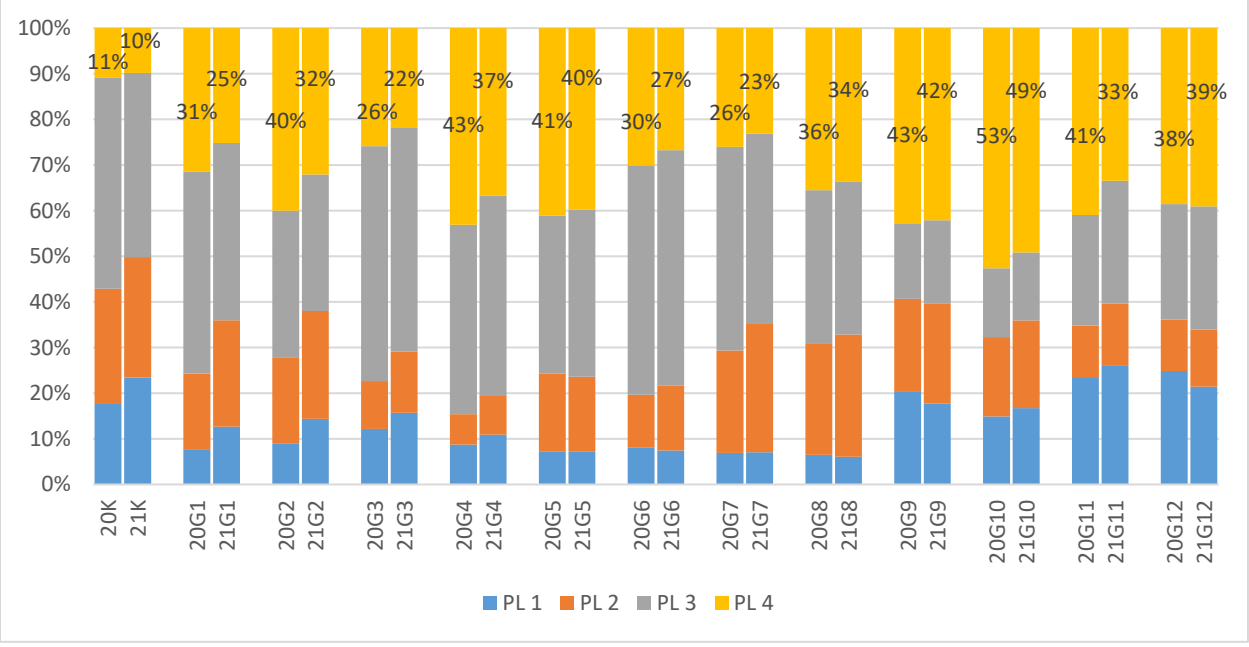
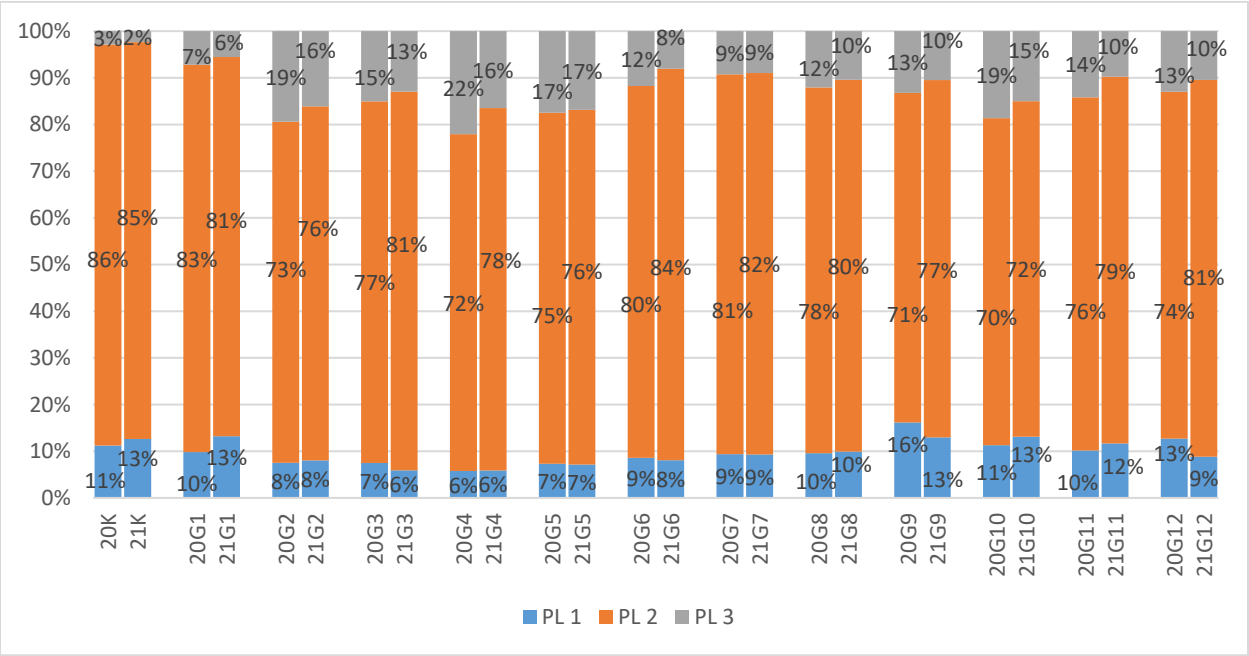


Figure IV-9. Comparison of 2020 and 2021 Performance-Level (PL) Results for Writing



The trend of the overall proficiency rates is provided in Figure IV-10. From 2020 to 2021, the overall proficiency rates stayed the same for grades 5 and 7; for other grades, there was a decrease in proficiency rates that ranged from 1% to 6%.

Figure IV-10. Comparison of 2020 and 2021 Overall Proficiency-Level (PL) Results



IV.3 Ongoing Program Improvement

This section summarizes the ongoing improvements for KELPA. Three upcoming initiatives intended to contribute to the validity evidence for KELPA are described.

IV.3.1 Enhanced Rater-Training Materials Development

The KELPA rater-training materials were redone for the 2021 administration to provide new prompts and exemplar student responses to one operational CR item per grade or grade band in speaking and writing. From that point, these materials are being expanded to cover all CR items by the 2023 administration. The purpose of the updated materials is to provide training materials supporting educators in applying rubrics to specific prompts. For detailed information, refer to Section II.2.2 Development of Rater-Training Materials of the current manual.

IV.3.2 Constructed-Response Score-Validation Study

Upon completion of the rater-training material enhancement effort in 2023, a CR score-validation study is planned. The objective of the CR score-validation study is understand the quality and accuracy of the locally derived CR scores by comparing them to scores obtained from a group trained scoring experts from both the Center for Accessible Teaching, Learning, and Assessment Systems (ATLAS) and KSDE. An in-person workshop with six expert panels will be convened to score approximately 300 student responses for each operational KELPA CR item. The expert panel will be grouped by grade or grade band and content (speaking and writing). Sample of student responses will be randomly selected from data collected during 2013 operational administration with appropriate representation of each score point from 0 to 3. At the in-person workshop, the panels will go through ATLAS lead training for their specific grade or grade band using the rater-training materials made available during the 2023 KELPA administration. The training will include ATLAS lead review of the anchor set of student responses for

each score point, a discussion of the training set of student responses and a final validation exercise to ensure the panelists are well calibrated. The panelists will then proceed with scoring the pool of student responses selected for the study. At the conclusion of the study, scores obtained from expert panel (considered as a proxy of true score) will be compared with scores of record obtained during operational test administration. A very close relationship between expert scores and field scores indicates accuracy of scores obtained from the field.

IV.3.3 Domain-Score Exemption

In some situations, students may be exempt from taking a domain test. Special circumstances codes available in Educator Portal, which allow school districts to manage test exemptions, will be enhanced to include KELPA domain exemptions for the 2021–2022 administration. Domain exemption requests will be reviewed and approved by KSDE. Exempted domains will not be included in the determination of overall proficiency. For example, students who are deaf or hard of hearing may be exempted from the listening test. For these students, overall proficiency will be determined by speaking, reading, and writing domain performance, and students will be considered proficient overall if they score at Level 4 in the speaking, reading, and writing domains.

IV.3.4 Incident Response Manual

A KAP system-wide Incident Response manual, which is applicable to KELPA program, will be utilized for the 2021-2022 test administration and beyond. The purpose of the Incident Response Manual is to guide investigative efforts for AAI staff when presented with a potential KAP/KELPA testing incident. This response plan outlines the steps for managing and addressing any item/test level incidents in order to remedy the effects and properly document relevant information.

V. Inclusion of All Students

This chapter provides an updated summary of the frequency of accommodation requests in 2021 KELPA administration and information about domain exemption in KELPA administration. For more detailed information about the accessibility framework in Kansas assessments, accessibility supports, available accommodations on KELPA, and the guidelines and procedures for selecting accommodations on KELPA, refer to Sections V.1 through V.3 of Chapter V in the *2020 KELPA Technical Manual* (AAI, 2021a).

V.1 Accommodations

All students who are identified as English learners (ELs), including those who need accommodations, must take KELPA. As described in *The Kansas Accessibility Manual*, a three-tiered accessibility framework (i.e., Tier 1: universal features for all students, Tier 2: designated features for some students, Tier 3: accommodations) is applied in Kansas state assessments. Accessibility tools are available for all students taking various components of the Kansas assessments; the tools available to students vary by testing programs under the Kansas Assessment Program [9] (KAP). 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. Refer to Section V.4.1 Selection of Accommodations in the *2020 KELPA Technical Manual* (AAI, 2021a) for guidelines that are applied to accommodation selection.

The *2020–2021 KELPA Examiner's Manual* provides more details about KELPA accommodations, including an overview, prohibited practices, and recording accommodations used during testing (i.e., most testing accommodations should be entered into the student's Personal Needs Profile [PNP]). Additional information about accommodations or Kite® tools can be found in the *Kite Educator Portal Manual for Test Coordinators*.

V.1.1 Selection of Accommodations

According to the *2020–2021 KELPA Examiner's Manual*, individualized education programs (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 in 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 not listed in *Tools and Accommodations for the Kansas Assessment Program (KAP)*, the examiner should contact the District Test Coordinator, who will send the request to the Kansas State Department of Education (KSDE). If the accommodation requested for a student changes the construct being tested, the test will not be valid for the student. Refer to Section V.4.1 Selection of Accommodations in the *2020 KELPA Technical Manual* (AAI, 2021a) for guidelines that are applied to every available accommodation on KELPA.

V.1.2 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. Any nonstandard accommodation requests and approvals are handled by KSDE. 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

[9] The Kansas Assessment Program provides general education assessments (i.e., assessments on English language arts, mathematics, and science), alternate assessments, career and technical education assessments, and KELPA.

availability of test accommodations for individual students. Table V-1presents the number of students who took KELPA in Kansas in 2021 and had PNP accommodation requests [10] for each accommodation. The summary in the table shows one accommodation request for kindergarten (i.e., whole screen magnification) and one for grade 1 (i.e., switches), 22 requests for grade band 2–3, and more than 100 requests for grade bands 4–5, 6–8, and 9–12. The most frequently requested accommodation was auditory calming, which provides relaxing, peaceful background music while a student takes the test. The second most frequently requested accommodation was whole screen magnification.

Table V-1. 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	1
1	0	0	0	0	0	1	0
2–3	11	2	0	0	0	1	5
4–5	172	4	2	1	0	0	4
6–8	153	4	3	1	0	0	13
9–12	98	4	0	0	1	4	35

Note. WSM = whole screen magnification.

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

VI. Academic Achievement Standards and Reporting

The KELPA standard-setting event occurred virtually in October 2020. 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. The Bookmark standard-setting method (Cizek & Bunch, 2007) was used to establish cut scores. For detailed procedures regarding the KELPA standard-setting event as well as information about evaluations of standard-setting method and event and other related information, refer to the *2020 KELPA Standard-Setting Technical Report* (AAI, 2021b) and Chapter VI. Academic Achievement Standards and Report of the *2020 KELPA Technical Manual* (AAI, 2021a). Because there were no updates to anything related to standard setting or performance level during 2020–2021 school year, this chapter briefly provides the updates about the student score report.

VI.1 Reporting

The 2021 KELPA testing window ended on March 31, 2021, and the scoring window closed on April 20, 2021. The KELPA student reports were made available to print and distribute on May 6, 2021. KELPA score reports are to students 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.

VI.1.1 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. 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. In 2021, in response to the COVID-19 pandemic and in consultation with KSDE and the Kansas Technical Advisory Committee, the following text was added to the top of the student report:

> When interpreting student progress toward proficiency on the KELPA, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance.

A sample 2021 KELPA Student Report is provided in 0.

VI.1.2 Interpretive Guides

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 (and its Spanish translation) are provided to assist the interpretation of the score reports. They 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 EL proficiency.

References

Achievement & Assessment Institute. (2021a). *2020 KELPA technical manual*. Kansas State Department of Education. https://ksassessments.org/sites/default/files/documents/2020_KELPA_Technical_Manual.pdf

Achievement & Assessment Institute. (2021b). *2020 KELPA standard-setting technical report*. Kansas State Department of Education.

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

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.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement*, (pp. 443–507). Washington DC: American Council on Education.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings* (ED532068). Westat, Center for Educator Compensation Reform; ERIC. https://files.eric.ed.gov/fulltext/ED532068.pdf

Johnson, D. (2005). *Aligning ELP assessments to ELP standards.* Pearson Education, Inc.

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

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

Rosenblum, I. (2021, February 22). [Letter to Chief State School Officers]. United States Department of Education, Office of Elementary and Secondary Education. https://www2.ed.gov/policy/elsec/guid/stateletters/dcl-assessments-and-acct-022221.pdf

Sinclair, A. L., Paulsen, J., & Thacker, A. (2021). *Kansas English Language Proficiency Assessment: Alignment study*. The Human Resources Research Organization.

Weighted kappa in R: For two ordinal variables. (n.d.). Inter-rater reliability measures in R [Course materials]; Datanovia. https://www.datanovia.com/en/lessons/weighted-kappa-in-r-for-two-ordinal-variables/

Appendix A. 2021 KELPA Teacher Survey

2021 KELPA Teacher Survey Questions

I. Demographics

1. Although you may serve many roles in your district, please select the one role that best describes your position as it relates to the Kansas English Language Proficiency Assessment (KELPA).
   - Building Test Coordinator (BTC)
   - Building User (BU)
   - Curriculum Director, Curriculum Coordinator
   - District or Building Administrator
   - District Test Coordinator (DTC)
   - District User (DU)
   - Program Director, Program Coordinator
   - Teacher (i.e., Classroom, Title 1, Special Education, EL) who administered KELPA
   - Teacher (i.e., Classroom, Title 1, Special Education, EL) who did **NOT** administer KELPA
   - Technology Director, Technology Coordinator
   - Support Staff

2. If your role in KELPA is test administrator, for which grade/grade band did you administer KELPA this year? [Please select all that apply.]
   - Grade K
   - Grade 1
   - Grade 2–3
   - Grade 4–5
   - Grade 6–8
   - Grade 9–12

3. Please indicate your number of years of K–12 educational experience in each of the following areas.
   English language arts _____
   Mathematics _____
   Science _____
   English learners _____

II. Technology

The following questions are about your use of Kite® Educator Portal and Student Portal.

1. Educator Portal is used to manage data for KELPA. Please rate how easy or hard it was to do the following in Educator Portal this year.

	Very hard	Somewhat hard	Somewhat easy	Very easy	Not applicable
Navigate the site.					
Enter Personal Needs and Preferences (PNP). Enter profile and First Contact information.					
Manage student data (e.g., rosters).					
Manage my account.					
Manage tests.					
Upload student responses for K-1 writing.					
Upload batch student scores.					
Enter scoring method for constructed-response items.					
Assign raters (as a DTC).					
Enter scoring option for speaking items.					

2. Kite Student Portal is used to deliver tests to students. Please rate how easy or hard it was for your students to do the following in Kite Student Portal this year.

	Very hard	Somewhat hard	Somewhat easy	Very easy	Not applicable
Enter the platform (logging in, selecting a test).					
Navigate within a test.					
Record a speaking response.					
Submit a completed test.					
Take the tests on laptops.					
Take the tests on Chromebooks.					
Take the tests on desktops.					
Take the tests on iPads.					

3. Please indicate your level of agreement or disagreement with the given statement.

	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
The Technology Practice Test familiarized students and teachers with the procedures for answering different types of technology-enhanced items.					

4. Please indicate your level of agreement or disagreement with each given statement.

	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
Items on the KELPA Practice Tests familiarize students and teachers with the assessment format.					
Items on the KELPA Practice Tests familiarize students and teachers with procedures for responding to different types of KELPA items.					

5. Please provide any additional feedback about your use of Kite Educator and Student Portal. [*Open-ended response*]

III. Scoring

1. The training materials were helpful in applying rubrics for scoring students' responses to speaking items.
    - strongly disagree
    - disagree
    - agree
    - strongly agree
    - not applicable

2. The training materials were helpful in applying rubrics for scoring students' responses to writing items.
    - strongly disagree
    - disagree
    - agree
    - strongly agree
    - not applicable

3. The length of the state scoring window was sufficient.
    - strongly disagree
    - disagree
    - agree
    - strongly agree
    - not applicable

4. Please rate the following statements about KELPA rater training workshops provided in your local school district.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The rater training helped me understand the scoring rubrics.					
The rater training helped me know how to use the scoring rubrics.					
The rater training provided useful information for my role as a rater.					
The rater training was well organized.					
The rater training materials were easy to use.					
The rater training materials helped me to score responses confidently.					
The amount of time used for rater training was about right.					

IV. Test-Administration Experience

1. Please rate the following statements about test administration for **listening** domain.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The domain test length was appropriate for corresponding grade levels.					
The test instructions were clear.					
The test instructions were helpful to students.					

2. Please rate the following statements about test administration for **speaking** domain.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The domain test length was appropriate for corresponding grade levels.					
The test instructions were clear.					
The test instructions were helpful to students.					

3. Please rate the following statements about test administration for **reading** domain.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The domain test length was appropriate for corresponding grade levels.					
The test instructions were clear.					
The test instructions were helpful to students.					

4. Please rate the following statements about test administration for **writing** domain.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The domain test length was appropriate for corresponding grade levels.					
The test instructions were clear.					
The test instructions were helpful to students.					

5. Please rate the following statements about your administration experience in general.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
I was confident in my ability to administer KELPA.					
The required test administrator training prepared me for the responsibilities of a test administrator.					
The District/Building Test Coordinator training sessions provided across the state were helpful.					

6. Please provide any suggestions for things that would help improve your ability to administer KELPA.
   [*Open-ended response*]

V. Student Experience

1. The content of KELPA measured important English language proficiency knowledge, skills and abilities.
   - strongly disagree
   - disagree
   - agree
   - strongly agree
   - not applicable

2. My student(s) had access to all necessary accessibility supports in order to participate in the assessment.
   - strongly disagree
   - disagree
   - agree
   - strongly agree
   - not applicable

VI. Resources

1. Please rate the following statements about KELPA support materials.

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The 2020–2021 KELPA Examiner's Manual was useful and helpful.					
KAP Practice Test Guide for Educators 2020–2021 was useful and helpful.					
KELPA Test Administration and Scoring Directions for Speaking files were helpful.					
KELPA Test Administration and Scoring Directions for Writing files were helpful.					

Appendix B. Summary Results of Teachers' Responses to Survey Questions [11]

Table B-1. Responses About Teacher's Role Relating to KELPA (N = 146)

Role	*n*	%
Building Test Coordinator	35	24
Building user	8	5
Curriculum director/coordinator	2	1
District or building administrator	7	5
District Test Coordinator	7	5
Program director/coordinator	7	5
Support staff	2	1
Teacher who administered KELPA	74	51
Teacher who did not administer KELPA	4	3

Table B-2. Distribution of Test Administrators by Grade or Grade Band (N = 146)

Grade or grade band	%
Kindergarten	18
1	19
2–3	19
4–5	18
6–8	15
9–12	11

Table B-3. Educators' Professional Experience in Years (N = 146)

Years	Experience with (%)			
	English language arts	Mathematics	Science	English learners
0–2	25	35	45	10
3–5	6	9	11	13
6–9	14	12	8	20
10 or more	55	45	36	58

[11] Percentages in the tables may not sum to 100% because of rounding.

Table B-4. Educators' Responses About User Experience of Kite Educator Portal (N = 146)

	Very hard	Somewhat hard	Somewhat easy	Very easy	Not applicable
Navigate the site	4%	16%	46%	34%	0%
Enter Personal Needs and Preferences Enter profile and first contact information	2%	3%	26%	12%	56%
Manage student data (e.g., rosters)	5%	12%	36%	16%	31%
Manage my account	2%	7%	48%	38%	5%
Upload student responses for grades K & 1 writing	0%	7%	14%	21%	58%
Upload batch student scores	4%	5%	8%	10%	73%
Enter scoring method for constructed-response items	7%	14%	35%	34%	10%
Assign raters (as a DTC)	3%	5%	10%	8%	74%
Enter scoring option for speaking items	8%	14%	37%	38%	3%

Note. DTC = district test coordinator

Table B-5. Educators' Responses About User Experience of Kite Student Portal (N = 146)

	Very hard	Somewhat hard	Somewhat easy	Very easy	Not applicable
Enter the platform (logging in, selecting a test)	1%	8%	32%	59%	1%
Navigate within a test	1%	11%	36%	51%	1%
Record a speaking response	3%	12%	42%	40%	3%
Submit a completed test	0%	2%	23%	73%	2%
Take tests on laptops	1%	7%	15%	21%	57%
Take tests on Chromebooks	1%	8%	23%	27%	41%
Take tests on desktops	0%	1%	6%	11%	82%
Take tests on iPads	1%	4%	10%	14%	71%

Table B-6. Educators' Responses About KELPA Technology Practice Test (N = 146)

Technology practice test . . .	Strongly disagree	Disagree	Undecided	Agree	Strongly agree	Not applicable
Familiarized students and teachers with the procedures for answering different types of technology-enhanced items	0%	4%	14%	47%	14%	21%

Table B-7. Educators' Responses About KELPA Practice Tests (N = 146)

Items on KELPA practice tests . . .	Strongly disagree	Disagree	Undecided	Agree	Strongly agree	Not applicable
Familiarized students and teachers with the assessment format	1%	6%	23%	60%	12%	0%
Familiarized students and teachers with the procedures for responding to different types of KELPA items	1%	5%	23%	59%	12%	0%

Table B-8. Educators' Responses About KELPA Scoring (N = 146)

	Strongly disagree	Disagree	Agree	Strongly agree	Not applicable
The training materials were helpful in applying rubrics for scoring students' responses to **speaking** items.	2%	8%	67%	18%	5%
The training materials were helpful in applying rubrics for scoring students' responses to **writing** items.	2%	4%	65%	28%	1%
The length of the state scoring window was sufficient.	1%	10%	66%	16%	5%

Table B-9. Educators' Responses About KELPA Rater-Training Workshops in School District (N = 146)

The rater training . . .	Strongly disagree	Disagree	Agree	Strongly agree	Not applicable
Helped me understand the scoring rubrics.	4%	3%	12%	62%	18%
Helped me know how to use the scoring rubrics.	3%	3%	14%	62%	18%
Provided useful information for my role as a rater.	3%	2%	14%	63%	18%
Was well organized.	3%	3%	13%	65%	16%
Materials were easy to use.	3%	3%	13%	67%	14%
Materials helped me to score responses confidently.	4%	5%	18%	59%	14%
Had about right amount of time.	3%	2%	14%	63%	18%

Table B-10. Educators' Responses About KELPA Test Administration for Domain Tests (N = 146)

Domain		Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
Listening	The domain test length was appropriate for corresponding grade levels.	3%	8%	18%	68%	2%
	The test instructions were clear.	1%	4%	12%	81%	1%
	The test instructions were helpful to students.	2%	8%	27%	62%	1%
Speaking	The domain test length was appropriate for corresponding grade levels.	8%	4%	18%	68%	2%
	The test instructions were clear.	1%	10%	19%	68%	1%
	The test instructions were helpful to students.	3%	9%	28%	58%	2%
Reading	The domain test length was appropriate for corresponding grade levels.	5%	8%	18%	66%	3%
	The test instructions were clear.	1%	2%	16%	79%	2%
	The test instructions were helpful to students.	2%	5%	23%	67%	3%
Writing	The domain test length was appropriate for corresponding grade levels.	3%	3%	10%	82%	1%
	The test instructions were clear.	1%	3%	19%	75%	1%
	The test instructions were helpful to students.	3%	5%	25%	65%	2%

Table B-11. Educators' Responses About KELPA Administration Experience in General (N = 146)

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
I was confident in my ability to administer KELPA.	1%	2%	5%	87%	4%
The required test administrator training prepared me for the responsibilities of a test administrator.	2%	2%	16%	72%	8%
The District/Building Test Coordinator training sessions provided across the state were helpful.	1%	4%	12%	60%	23%

Table B-12. Educators' Responses About Student Experiences (N = 146)

	Strongly disagree	Disagree	Agree	Strongly agree	Not applicable
The content of KELPA measured important English language proficiency knowledge, skills and abilities.	2%	16%	73%	8%	1%
In general, ELs classified as Proficient based on the KELPA are able to fully access grade-level academic content.	0%	15%	69%	15%	1%
In general, ELs classified as Not Proficient based on the KELPA are not able to fully access grade-level academic content without the use of ESOL.	4%	27%	54%	13%	2%
My student(s) had access to all necessary accessibility supports in order to participate in the assessment.	1%	2%	69%	26%	3%

Note. EL = English learner; ESOL = English for speakers of other languages.

Table B-13. Educators' Responses About KELPA Support Materials (N = 146)

	Disagree	Somewhat disagree	Somewhat agree	Agree	Not applicable
The 2020–2021 KELPA Examiner's Manual was useful and helpful.	1%	3%	16%	77%	3%
The KAP Practice Test Guide for Educators 2020–2021 was useful and helpful.	1%	2%	12%	64%	21%
The KELPA Test Administration and Scoring Directions for **speaking** files were helpful.	1%	5%	16%	75%	3%
The KELPA Test Administration and Scoring Directions for **writing** files were helpful.	1%	5%	20%	72%	2%

Appendix C. Response to 2021 External Evaluation of KELPA Alignment Study

October 2021

Purpose

An alignment study to review the Kansas English Language Proficiency Assessment (KELPA) test forms was conducted in spring 2021. This response to the study includes a KELPA overview, the results of the study, and a response to each claim. The contents reflect discussions among the Kansas Department of Education (KSDE), the Technical Advisory Committee, and the Achievement and Assessment Institute (AAI) at the University of Kansas.

KELPA Overview

KELPA is a yearly summative assessment designed to measure English language proficiency (ELP) of students identified as English learners (ELs) in four domains: listening, speaking, reading, and writing. Identified ELs in grades K–12 take KELPA, and the assessed grades or grade bands include kindergarten, 1, 2–3, 4–5, 6–8 and 9–12. KELPA administration is mostly computer based, except for a small number of paper-based writing items for students in kindergarten and grade 1. KELPA contains a mix of educator-scored constructed-response (CR) items and machine-scored items. Results from KELPA in each domain are used to determine an overall proficiency score. Students who score at Level 4 in each domain are considered proficient overall. These students demonstrate the English language skills necessary to engage with grade-level academic content and are thus eligible to exit English language support services.

The 2018 Kansas English Language Proficiency (KELP) Standards for English Learners are based on the 2017 Kansas Standards for English Language Arts. Overall, twenty-one 2018 KELP standards are used across grades and domains that vary slightly according to grade-appropriate expectations. The 2018 KELP standards served as the basis for the design and content of KELPA. The Standards reflect progressions of the grade-level standards within the four domains and outline the crucial language skills that ELs need to thrive academically. The performance-level rubric outlines the stages of language acquisition and continua of social language, receptive language, and expressive language. It includes levels from beginner (Level 0) to mastery (Level 6); Level 4 (proficient) indicates a student is able to access grade-level content without support and demonstrates language skills equivalent to non-EL peers.

The KELPA test blueprint is organized by clusters that are composed of smaller groups of similar standards. Clusters in listening and speaking include Comprehension & Collaboration, Presentation of Knowledge & Ideas, and Language in Speaking & Listening. Clusters in the reading domain include Reading Foundations, Language in Reading, Discourse Comprehension, and Craft and Structure. Writing domain clusters include Language in Writing and Production of Writing. The test blueprint indicates a score-point range for each cluster by grade or grade band; computer-scored items are worth 1 point, and educator-scored items in speaking and writing are worth 3 points.

In spring 2021, Human Resources Research Organization (HumRRO) conducted an external alignment study for KELPA using ratings from panels of experts. The evaluated claims and associated criteria are outlined in Table 1.

Table 1. Claims and Criteria Investigated by the Alignment Study

Claim	Criterion
1. KELPA items are aligned to 2018 KELP standards.	There are no flagged items (i.e., Not Aligned; any flagged items should be reviewed by content experts).
2. KELPA items represent 2018 KELP standards.	At least 50% of domain-specific standards are represented on domain-level tests.
3. KELPA meets test blueprints, representing a balanced assessment.	Panel data indicate that KELPA domain-level tests meet blueprint specifications.
4. KELPA domain-level tests are reliable.	Not determined by the alignment study, but Cronbach's alpha coefficients from the KELPA Technical Manual are reported to support reliability evidence related to alignment. Cronbach's alpha coefficients of .70 and above are generally considered indications of acceptable internal-consistency reliability (Cortina, 1993).
5. KELPA includes items representing a range of LDLs.	All LDLs should be represented on each domain. More than 50% of items are at LDL 2 or higher.
6. Language proficiency requirements of the academic standards are addressed by the 2018 KELP standards.	Language proficiency requirements of at least 70% of the academic standards within a given content area are rated at Level 4 (Proficient) or lower on the Kansas Standards for English Learners Performance Level Rubric.

Note. Adapted from Sinclair et al., 2021

The alignment workshop contained two parts: the items-to-standards activity and the standards-correspondence activity. The items-to-standards activity addressed Claims 1–3 and 5. Panelists evaluated the alignment between KELPA items and 2018 KELP standards, as well as the items' linguistic difficulty level (LDL), which could range from Level 1 to Level 3 (Johnson, 2005). Panelists independently matched items to standards and rated the LDL before discussing their rationales with other panelists. After the first discussion, panelists were shown the item metadata containing the standard and LDL assigned to the item. Panelists then engaged in further group discussion before reaching final consensus. The standards-correspondence activity addressed Claim 6 and involved panelists' evaluating the alignment between the language proficiency expectations in the 2018 KELP standards and Kansas's academic content standards. This activity involved using the Kansas Standards for English Learners Performance Level Rubric to determine the level of language proficiency a student needs to acquire to demonstrate achievement in the academic knowledge and skills reflected in the standards in English language arts, mathematics, and science. The Kansas Standards for English Learners Performance Level Rubrics is a holistic description of EL performance in five broad categories: stage of language acquisition, using language to communicate in social contexts, using language to construct meaning (reading and listening), using language to convey ideas (speaking and writing), and using language to engage in grade-level content.

Overall results from the items-to-standard alignment activity are represented in Table 2.

Table 2. Grade and Grade-Band Tests by Domain Meeting Alignment Criteria for Claims 1–5

Grade or grade band	Domain	Claim 1 *	Claim 2	Claim 3	Claim 4	Claim 5
		KELPA items are aligned to 2018 KELP standards.	KELPA items represent 2018 KELP standards.	KELPA meets test blueprint, representing a balanced assessment.	KELPA domain-level tests are reliable.	KELPA domain-level tests include a range of LDLs.
Kindergarten	Listening	✓	✓	×	✓	✓
	Speaking	✓	✓	×	✓	×
	Reading	✓	✓	×	✓	✓
	Writing	✓	✓	✓	✓	✓
1	Listening	✓	✓	✓	✓	✓
	Speaking	✓	✓	×	✓	×
	Reading	✓	✓	×	✓	×
	Writing	✓	✓	✓	✓	✓
2–3	Listening	✓	×	×	✓	✓
	Speaking	✓	✓	✓	✓	✓
	Reading	✓	✓	×	✓	✓
	Writing	✓	✓	×	✓	✓
4–5	Listening	✓	✓	✓	✓	✓
	Speaking	✓	✓	×	✓	✓
	Reading	✓	×	✓	✓	✓
	Writing	✓	✓	×	✓	✓
6–8	Listening	✓	✓	✓	✓	✓
	Speaking	✓	✓	✓	✓	×
	Reading	✓	✓	✓	✓	✓
	Writing	✓	✓	✓	✓	×
9–12	Listening	✓	✓	✓	✓	✓
	Speaking	✓	✓	✓	✓	×
	Reading	✓	✓	×	✓	×
	Writing	✓	✓	✓	✓	×

Note. * A check mark in the table indicates the claim was met. Adapted from Sinclair et al., 2021.

Table 3 shows the results for the standards-correspondence activity. The criterion for Claim 6 was met in all grades and subjects except grade-1 mathematics.

Table 3. Grade and Grade-Band Tests Meeting Claim 6 Criterion by Academic Content Area

Grade or grade band	Claim 6 [*]: Language proficiency requirements of the academic content standards are addressed by 2018 KELP standards.		
	English language arts	Mathematics	Science
Kindergarten	✓	✓	✓
1	✓	×	✓
2–3	✓	✓	✓
4–5	✓	✓	✓
6–8	✓	✓	✓
9–12	✓	✓	✓

Note. [*] A check mark in the table indicates the claim was met. Adapted from Sinclair et al., 2021.

Items-to-Standard Alignment Activity Results (Claims 1–5)

Claim 1: KELPA items are aligned to 2018 KELP standards.

Results

Panelists determined the content alignment between KELPA items and 2018 KELP standards by identifying the knowledge, skills, and abilities (KSAs) each item represented and matching it to the standard that most closely represented those KSAs. CR items could be linked to a second standard (item metadata currently provide for only one aligned standard per item). Panelists marked each item as Fully Aligned, Partially Aligned, or Not Aligned.

The HumRRO study reported that panelists did not rate any items as Not Aligned, so the criterion for Claim 1 was met. While most items were rated as Fully Aligned, a few were rated as Partially Aligned: grade-1 reading (32%); grade band 2–3 writing (26.3%); grade band 6–8 speaking (22%); and grade band 9–12 listening (41.7%), speaking (30%), and reading (60.9%). The panel for grade band 9–12 was mostly made up of English language arts (ELA) classroom teachers. HumRRO noted this panel might have had more difficulty than other panels in distinguishing ELP concepts from ELA concepts. Although all items were fully or partially aligned, there were cases where panelists identified the alignment according to different standards than what item writers identified in the item metadata. Formatively, HumRRO recommended the metadata be reviewed for items when there was a lower than a 70% match between item writers' identification of standards (metadata) and panelist ratings on a domain test.

Response

Per HumRRO's recommendation and to ensure that the item metadata most accurately represent the intended link to the 2018 KELP standards, AAI staff will review the panelists' standards ratings of the item metadata for all items that were not rated Fully Aligned.

In addition, given that the standards metadata were informed by panels of trained educator that reviewed items during the item-development process and as recommended by HumRRO, careful consideration will be given as to whether the standard selected by the panel represents stronger alignment than the standard identified in the metadata, or whether a dual alignment (i.e., providing both primary and secondary alignments) for the item is warranted. After reviewing recommendations

with KSDE, AAI will update item metadata as needed, which will more accurately represent blueprint coverage.

Claim 2: KELPA items represent 2018 KELP standards.

Results

HumRRO's reported results describe the number of domain-specific standards (based on the number of standards in the test blueprint) compared to the number of standards that panelists linked to items on domain-level tests, using primary and secondary alignment for the CR items and primary alignment for machine-scored items. The majority of domain-specific standards were represented by items on most domain tests. The criterion for Claim 2 was met for all grades or grade bands and domains, with the exception of grade band 2–3 listening (43%) and grade band 4–5 reading (40%).

Response

The HumRRO alignment study was designed to evaluate coverage of all domain-specific standards and did not consider the fact that blueprints were developed from clusters of standards. In October 2018, a panel of 14 Kansas educators of ELs met with staff from KSDE and AAI to discuss the new 2018 KELP standards and their impact on assessment. Educators were asked to consider whether some standards or performance levels provide more valuable information than others, whether any standards subsume any others, or whether the standards can be grouped into clusters. Those discussions informed KELPA's 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 KELP standards; however, clusters of standards (i.e., strands) are included in the 2017 Kansas English language arts standards from which the 2018 KELP standards were adapted. *Clusters* are small groups of similar standards that help organize individual standards for ELP. 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. Therefore, the test blueprints were constructed according to score-point ranges at the cluster level and not at the standard level, informed by the discussion noted above.

To help further evaluate Claim 2 results in light of the KELPA blueprint structure, AAI used the HumRRO alignment-study data in a follow-up analysis to evaluate whether KELPA items represent clusters of standards as defined by the test blueprint. This follow-up analysis evaluates whether the two domain tests that did not meet the HumRRO criterion introduced inconsistencies in blueprint coverage. No items represent the clusters shown in Table 4, consistent with the test blueprints. All standards tested through CR rubrics have secondary ratings supporting them except for kindergarten speaking.

Table 4. KELPA Items Representing Clusters in Blueprints

Domain	Cluster	Grade band	No. of items aligned to cluster	Blueprint range (in score points)
Listening	Language in Listening & Speaking	2–3	0	0–8
Reading	Craft & Structure	4–5	0	0–3

The criterion for Claim 2 was "At least 50% of domain-specific standards are represented by items on domain-level tests." This criterion was established by HumRRO and does not reflect the process used to establish KELPA blueprints. The final blueprints associated score-point ranges for the clusters and not individual standards. *Because the standards without corresponding items are expected given the blueprint structure, no further corrective action is needed.* However, to ensure cluster-level content coverage is maintained after the metadata are revised in response to Claim 1 findings, we will replicate the analysis. Unmeasured standards could also be discussed if KSDE decides to revise the KELPA blueprint in the future.

Claim 3: KELPA meets test blueprints, representing a balanced assessment.

Results

The HumRRO study reported that results for Claim 3 were mixed. Data from the panelists' linking of items to standards were used to identify the number of score points per cluster. This number was compared to the range of score points indicated on the test blueprint for each cluster. If the panel score points fell below or above the specified range for a cluster in a grade or grade band and domain, the criterion for Claim 3 was not met. Secondary linkages were included only if panel score points from primary linkages did not fall within the specified range.

In kindergarten, only the writing domain met the criterion. This may be because there are fewer kindergarten items than in other grades but similar score-point ranges in the blueprint. In grade 1 and grade bands 2–3 and 4–5, two of the four domains met the criterion. In grade band 6–8, all criteria were met in all four domains. In grade band 9–12, the criterion was met for all domains except reading.

HumRRO also reported that the number of 2018 KELP standards that panelists linked to items generally aligned with the number of standards in the metadata.

Table 5 and Table 6 summarize blueprint-coverage inconsistencies by domains in listening and reading and in speaking and writing, respectively, as described in the HumRRO report.

Table 5. Blueprint Inconsistencies for Listening and Reading Domains

Domain	Grade or grade band	Cluster	Difference from blueprint range
Listening	Kindergarten	Comprehension & Collaboration	-2
		Language in Speaking & Listening	+1
	2–3	Comprehension & Collaboration	+1
Reading	Kindergarten	Reading Foundations	+4
		Language in Reading	-2
	1	Reading Foundations	+1
		Language in Reading	-2
	2–3	Reading Foundations	-1
		Discourse Comprehension	+2
	9–12	Reading Foundations	-1

Table 6. Blueprint Inconsistencies for Speaking and Writing Domains

Domain	Grade or grade band	Cluster	Difference from blueprint range using primary link only *
Speaking	1	Comprehension & Collaboration	-6
		Presentation of Knowledge & Ideas	+9
	2–3	Presentation of Knowledge & Ideas	+3
	4–5	Comprehension & Collaboration	-3
		Presentation of Knowledge & Ideas	+6
Writing	2–3	Language in Writing	+13
	4–5	Language in Writing	+3

Note. * All speaking domain items are constructed-response (CR) items, and the writing domain has a mix of machine-scored and CR items. For CR items, the metadata included a primary linkage to a single standard, and the study allowed for a secondary linkage to a secondary KELP standard. To account for this secondary linkage, the study calculated a post-hoc percentage for each cluster, relative to the score points in the domain. A precise point difference from the blueprint range should not be calculated in speaking or writing until a secondary alignment is assigned.

Response

After the creation of operational test blueprints, an operational field-test blueprint was constructed to include additional items as overages for the 2020 operational field test. The overage items were included to ensure sufficient operational items were available for scoring purposes. To strengthen measurement power, after the 2020 operational field-test administration, the overage items were retained for psychometric purposes.

Clusters that exceeded the blueprint ranges were caused by the retained overage items, and *no further corrective action is needed*.

For clusters that fell below the blueprint score-point ranges, blueprint coverage will be reanalyzed after the follow-up steps from Claim 1 are complete; that process may result in some realignment of items, and blueprint coverage may be affected. Blueprints may need to be adjusted to account for the potential dual alignments of individual items to primary and secondary standards. Additionally, items may be replaced or added to the test forms if needed. These efforts in blueprint adjustment may include reducing the number of items in one cluster and increasing the number of items in another cluster within the same domain test, to confirm all clusters meet the minimum score-point thresholds while ensuring test length and testing time do not exceed KSDE limits for test design.

Claim 4: KELPA domain-level tests are reliable.

Results

Claim 4 was not assessed in the alignment study. However, the 2020 KELPA Technical Manual provided evidence that the criterion for acceptable reliability was met for all grades and grade bands across domains.

Response

All areas met the criteria for Claim 4: KELPA domain-level tests are reliable. *No action is required.*

Claim 5: KELPA includes items representing a range of linguistic difficulty levels (LDLs).

Results

Panelists identified the LDL of each item. The two-part criterion for this claim was:

1. All LDLs are represented on each domain.
2. More than 50% of the items are at LDL 2 or higher.

In eight of 24 grade or domain tests, panels identified no items at LDL 1. In those eight grade or domain tests, part of the criteria for Claim 5 was not met. However, the highest percentage of LDL 1 items was 28%, meaning more than 50% of the items in each grade–domain combination were rated at LDL 2 or higher (which meets a part of the criterion for Claim 5).

HumRRO also reported the comparison between the panel LDL ratings and the metadata LDL ratings. Overall, panelists' ratings matched the metadata ratings, although there were some divergences. For example, panel ratings for 40% of speaking items in grade 1 and grade band 2–3 were rated at a lower LDL than the metadata, while panel ratings for 40% of grade-1 listening items, grade band 2–3 writing items, and grade band 9–12 reading items were rated higher than the metadata.

Response

HumRRO identified eight out of 24 domain or grade-band tests that lacked LDL-1 items and therefore did not meet one of the two criteria for Claim 5. LDLs were not used in creating the 2018 KELP standards or KELPA test blueprints, nor were item writers trained to write to different LDLs during the development of KELPA.

The HumRRO alignment results provide useful formative information on the extent to which KELPA items span a range of complexity levels. All domain-level items spanned at least two LDLs. Because there were no a priori specifications for intended distributions of items across LDLs, AAI conducted follow-up analyses comparing empirical analyses of item difficulty and panelists' rated LDL to explore potentially reasonable ranges of LDL distributions. While level-1 items provide some access for students with low English proficiency, those items tend to be easier and do not provide sufficient information regarding students' abilities in the score range closer to proficiency. To balance the test-length need requested by the field (i.e., shorter test length), thus efficiently using student testing time, and to maximize test precision at the most important level-4 cut score, it is not prudent to include many items that provide very little information about students. Therefore, *no further corrective action will be taken for the operational assessment at this time*. *However, if additional content is developed for KELPA, LDL training and criteria will be incorporated into the test-development process to conform to the desired LDL distributions.*

Standards-Correspondence Activity Results (Claim 6)

Claim 6: Language proficiency requirements of the academic standards are addressed by the 2018 KELP standards.

Results

The criterion for Claim 6 is "Language proficiency requirements of at least 70% of the academic content standards within a given academic content area are rated at Level 4 (Proficient) or lower on the Kansas Standards for English Learners Performance Level Rubric." HumRRO reported that panelists rated at least 70% of the grade-level academic content standards in ELA, mathematics, and science as requiring a level of language proficiency of Level 4 or lower. The one exception was in grade-1 mathematics. The criterion was not met for grade-1 mathematics because 42% of the standards were rated as requiring Level 5 (Mastery).

Response

The a priori criterion used a logical assumption about what constitutes reasonable access to academic content taught in English, rather than a criterion derived from 2018 KELP standards development or empirical data.

Panelists provided ratings to map the grade-level, academic content standard to the Kansas Standards for English Learners Performance Level Rubric without engaging in more in-depth conversations about information provided in the 2018 KELP standards, such as grade-level vocabulary and classroom supports for academic discourse. As a preliminary follow-up step, AAI staff more closely examined the 2018 KELP standards and the grade-1 mathematics standards, and speculated that the language demand for mathematics standards might be met by Level 4 students. KSDE and AAI staff will collaborate to define the expected relationship between 2018 KELP standards and the mathematics academic standards. Once this criterion is established, a panel of Kansas educators will be convened to further examine the grade-1 mathematics standards and the correspondence between these standards and performance levels of ELs.

Summary of Next Steps and Timeline

Spring 2022—Winter 2022/2023

- Evaluate the current metadata and alignment-study panelist ratings of items to standards for all items that did not receive a Fully Aligned rating and update metadata as needed.
- Analyze cluster-level content coverage of blueprints if any metadata is revised.
- Update test blueprints to account for secondary alignment to items and reflect proportion of score points by cluster.
- Establish criterion to review grade-1 mathematics standards and correspondence between standards and performance levels of ELs.
- Convene educators to review language demands of grade-1 mathematics standards and correspondence between these standards and performance levels of ELs.

Winter 2022/2023—Fall 2023

- Adjust blueprints and test forms to account for the potential dual alignment of individual items to primary and secondary standards.

Subsequent Years (as needed)

- Determine whether to include any unmeasured standards if test blueprints are revised.
- Include training and criterion to establish LDL distributions if additional test-development occurs.

References

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98.

Sinclair, A. L., Paulsen, J., & Thacker, A. (2021). *Kansas English Language Proficiency Assessment: Alignment study*. The Human Resources Research Organization.

Appendix D. Sample 2021 KELPA Student Report

STUDENT REPORT:
GRADE: 4 / STATE ID:
SCHOOL:
DISTRICT:

2020–2021
KELPA
Kansas English Language Proficiency Assessment

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.

When interpreting student progress toward proficiency on the KELPA, please take into consideration how the conditions for learning, which may have been disrupted by the pandemic, may influence performance.

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.

Domain Performance Levels

	Domain Score				Progress Toward Proficiency
Year	Speaking	Writing	Listening	Reading	
2020	1	1	1	1	
2021	2	1	3	2	Satisfactory Progress

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.