KELPA

Kansas English Language Proficiency Assessment

KELPA Technical Manual

October 2024

Contents

Table of Tables

Table of Figures

I. Statewide System of Standards and Assessments

The Kansas English Language Proficiency Assessment (KELPA) is the annual summative assessment for K–12 English learners (ELs) in Kansas, administered each spring. Developed in alignment with the 2018 Kansas Standards for English Learners: Grades K–12 (referred to as the 2018 Standards), KELPA is part of federal legislation concerning elementary and secondary education for ELs. The assessment covers the following grade bands: kindergarten, 1, 2–3, 4–5, 6–8, and 9–12. The target population for KELPA includes all students identified as ELs in grades K–12.

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)—listening, speaking, reading, 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 help students achieve proficiency in both social and academic English, enabling them to meet grade-level academic standards as quickly as possible. The 2018 Standards also informed the design and content of the new KELPA, first administered in 2020. Refer to the *2020 KELPA Technical Manual* (Achievement and Assessment Institute [AAI], 2021a) for more details about the 2018 Standards.

I.2. Test Purposes and Uses

KELPA is an annual summative assessment for K–12 students identified as not proficient in English, regardless of whether they receive English for Speakers of Other Languages (ESOL) services. It is mandated by the Every Student Succeeds Act (ESSA, 2015), which reauthorized the Elementary and Secondary Education Act (ESEA, 1965). In compliance with this law, KELPA results are used to determine the English language proficiency of English Learners (ELs) and to assess their progress in developing listening, speaking, reading, and writing skills 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 indicate 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 necessary English language skills for engaging with grade-level academic content at a level comparable to non-ELs (i.e., level 4—early advanced) in all four domains (listening, speaking, reading, and writing) are considered proficient in English and may exit 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 2018 standards highlight and amplify the critical language, knowledge about it, and skills necessary for ELs to be successful in school.

I.3. Intended Population

The Kansas State Department of Education (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. All students in grades K–12 who are identified as ELs must take KELPA, regardless of whether 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 provided by KSDE.

Some ELs may need accommodations for KELPA. 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*. A summary of accommodations is provided in Chapter V. Inclusion of All Students in this technical manual.

I.4. Overview of Technical Manual Updates

A complete technical manual was created for the first year of operational administration in 2020. During the 2020–2021 school year, an independent alignment study was conducted to document validity evidence for KELPA, as documented in the *2021 KELPA Technical Manual* (AAI, 2021b). This current manual provides updates for the 2024 administration; therefore, only sections with updated information are included in this manual. For a complete description of KELPA, refer to the *2020 KELPA Technical Manual*.

II. Assessment System Operations

This chapter provides updated information about the design, development, administration, and test security of the Kansas English Language Proficiency Assessment (KELPA). For more details (e.g., monitoring test administration), refer to Chapter II. Assessment System Operations in the *2020 KELPA Technical Manual* (Achievement and Assessment Institute [AAI], 2021a).

II.1. Test Design and Development

KELPA, a component of the Kansas Assessment Program (KAP), is a fully computer-based assessment for students in grades 2 through 12. For students in kindergarten and grade 1, the exam is primarily computer-based, but it includes a few writing items that are completed using 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 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, which also provides detailed information about KELPA test blueprints (see Section II.1.1 Test Blueprints), test design (see Section II.1.2 Test Design), and test construction (see 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* includes detailed descriptions of the typical procedures for various 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 both the rubric and the rater-training materials.

II.2.1. Rubric Development

KELPA rubrics developed for the 2020 administration were used in 2021. Refer to Section II.2.4 Rubric Development in the 2020 KELPA Technical Manual for detailed activities of rubric development by phase. To support rater use of the rubrics in kindergarten and grade 1, supplementary documents were added to the rater-training materials to provide additional, more specific guidance on using the writing rubrics. These supplemental documents were also developed in 2020 and used in the 2020 through

2024 administrations. Item-specific rater-training materials finalized in 2022 were used for the 2024 test administration.

II.3. Test Administration and Scoring

The 2024 KELPA testing window was open to students from January 29 through March 8, 2024. Educators were able to enter scores for CR items until March 29, 2024. 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*.

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. DTCs 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*. The provided training webinars, recorded and posted on site, are updated every year. The training slides, frequently asked questions, and responses to these questions are also posted on the DTC Virtual Training site.

The standardized test-administration procedures provided for districts, schools, and teachers are described in the *2023–2024 KELPA Examiner's Manual* (*Examiner's Manual* hereafter). The *Examiner's Manual* also provides guidance and procedures related to the administration of KELPA in 2023–2024. For example, it includes 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 provided in Section II.3.2 Test-Administration Procedures of the *2020 KELPA Technical Manual*.

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*. For more details about security requirements, refer to the 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* provide detailed information about and requirements for test-materials security, test-related data security, security of personally identifiable information, and accommodations-related security.

II.5. Testing Irregularity

During the spring 2024 KELPA test-administration window, KSDE received a total of 20 test resets. A test reset delivers the same test (domain) and wipes the previous responses. This requires approval from the state. Test administrators and coordinators are trained to report test-administration irregularities. During the operational window, monitoring of testing data was conducted by AAI, which oversees and manages the Kite system. ATS conducted data validation daily to monitor system usage and identify testing irregularities. A dashboard of testing activities is available for Educator Portal for administrators

in the field to monitor and record activities for the KAP program. The dashboard records all system usage, including a DTC training log, click history of student responses, test-taking hours, test-status summary, server load, the number of Kite Service Desk (i.e., support for Educator Portal and Student Portal) tickets, and the frequency of test reactivations (which activate a test if needed, without wiping previous responses and does not need KSDE approval).

Examples of testing irregularities include the following:

- Fast test-taking behavior (i.e., students finished a test section in a short amount of time)
    - Typically, only requires a reactivation and not a reset for KELPA.
- Irregular testing time (i.e., a test session started or ended outside of school hours),
    - Recorded by the system and can be viewed in the dashboard in Kite Educator Portal
- Tests reactivated by users (i.e., test administrators) due to student enrollment or demographic data errors
    - Recorded by the system and can be viewed in the dashboard
- Student clicks through a test and then submits without answering any questions.
    - Requires a test reactivation, which can be done by the testing coordinator
- The student was caught cheating.
    - State approval is required for a test reset. The state may require the user to enter a special circumstance code (SC-28). The test proctor will notify the DTC, who will call KSDE for guidance.
    - If a test reset is **not** approved, the DTC will enter the SC code. This makes the test invalid; the student will be recorded as not tested, and it will result in lower participation for the specific school/district.
- The student took someone else's test by accident.
    - Requires a test reset or moving the data to the correct test. Both options require approval from KSDE.
- The educator scored the wrong student and needs the scores moved to the correct student.
    - Requires approval from KSDE to move the scores
- The student's personal need profile (PNP) was not set up correctly before testing.
    - Requires approval from KSDE to reset the test

III. Technical Quality—Validity

Standards for Educational and Psychological Testing defines *validity* as "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). There are five sources of evidence to consider when evaluating test-score validity (APA et al., 2014): (a) test content, (b) response processes, (c) internal test structure, (d) relationships between test scores and other variables, and (e) consequences of testing. The Kansas English Language Proficiency Assessment (KELPA) test forms in 2024 were the same as the operational forms from 2020 and 2023; 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* (Achievement and Assessment Institute [AAI], 2021a). This chapter presents validity evidence collected or evaluated during the 2023–2024 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 rely on the correspondence between items and the 2018 Standards, as well as between the test and test blueprint. The external, independent KELPA alignment study was conducted by the Human Resources Research Organization (HumRRO) with participation of Kansas educators in spring 2021 to examine the extent of alignment among KELPA, the 2018 Standards, and the academic content standards (Sinclair et al., 2021). Chapter III of the *2022 KELPA Technical Manual* documents actions taken in response to the recommendation from the independent alignment study.

III.2. Validity Evidence Based on Relations to Other Variables

The external validity evidence is defined as "evidence based on relationships with other variables provides evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretations" (APA et al., 2014, p. 16). The three types of external evidence are convergent, discriminant, and criterion related (either predictive or concurrent). *Convergent evidence* is provided by relationships among students' performance on different assessments measuring similar constructs. *Discriminant evidence* is provided by relationships among 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, 1951; Messick, 1989).

The external assessments used in this study are the Kansas Assessment Program (KAP) English language arts (ELA) and mathematics assessments, which are administered annually to students in grades 3–8 and 10, 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 English learners' (ELs') proficiency in each KELPA domain may have a different impact on their performance in the grade-level academic tests.

Table III-1 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 .48 (grade 8) to .65 (grade 6); the weakest correlations were observed between ELA and the speaking domain, ranging from .21 (grade 8) to .33 (grade 3). Correlation coefficients between the KAP ELA and KELPA speaking domain across grades were small (less than 0.3 except in grade 3). For relationships between KAP ELA and KELPA listening, reading, and writing, medium to large correlation coefficients were found across grades.

Table III-1. Correlations Between KELPA Domain Scores and KAP English Language Arts (ELA) Scores

Grade	Correlation Between KAP ELA and Domain			
	Listening	Speaking	Reading	Writing
3	0.46	0.33	0.63	0.57
4	0.55	0.29	0.64	0.58
5	0.50	0.23	0.57	0.48
6	0.55	0.27	0.65	0.52
7	0.50	0.22	0.57	0.44
8	0.44	0.21	0.48	0.40
10	0.43	0.25	0.56	0.45

Table III-2 presents correlations between KELPA domain scores and KAP mathematics scores. Compared to the 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 .25 (grade 10) to .54 (grade 3); the weakest correlation was between KAP mathematics and KELPA speaking domain, ranging from .10 (grade 10) to .35 (grade 3). Relationships between KAP mathematics and KELPA were weakest for grade 10.

Table III-2. Correlations Between KELPA Domain Scores and KAP Mathematics Scores by Grade

Grade	Correlation Between KAP mathematics and Domain			
	Listening	Speaking	Reading	Writing
3	0.44	0.35	0.54	0.55
4	0.43	0.26	0.47	0.49
5	0.43	0.22	0.44	0.44
6	0.39	0.21	0.47	0.37
7	0.36	0.21	0.38	0.33
8	0.33	0.25	0.37	0.32
10	0.24	0.10	0.25	0.28

Table III-3 presents correlations between KELPA domain scores and KAP science scores. The strongest correlation was between KAP science and reading scores, ranging from .34 (grade 11) to .52 (grade 5); the weakest correlation was between science and speaking scores, ranging from .10 (grade 11) to .29 (grade 5). Correlations between KAP science and KELPA scores were weakest for grade 11.

Table III-3. Correlations Between KELPA Domain Scores and KAP Science Scores by Grade

Grade	Correlation Between KAP Science and Domain			
	Listening	Speaking	Reading	Writing
5	0.51	0.29	0.52	0.43
8	0.35	0.17	0.39	0.27
11	0.26	0.10	0.34	0.24

Table III-4 presents student performance on KAP ELA, mathematics, and science for proficient KELPA students. More proficient ELs in lower grades scored proficient in KAP ELA and mathematics compared to students at higher grades. For example, 35% of proficient ELs in grade 3 scored at level 3 or level 4 (proficient) in KAP ELA, compared to 8% of proficient ELs in grade 7. Proficient ELs at grade 10 had the lowest performance in KAP ELA and mathematics: 4% of proficient ELs scored at level 3 or level 4 on KAP ELA, and only 8% of proficient ELs scored at level 3 or level 4 on mathematics. For science, proficient ELs at grade 8 had the lowest performance: 8% of proficient ELs scored at level 3 or level 4 on KAP science.

Table III-4. Performance of Proficient English Learners on KAP English Language Arts, Mathematics, and Science Assessments

	Proficient English Learners (ELs)											
	KAP English Language Arts				KAP Mathematics				KAP Science			
Grade	Proficient ELs (*N*)	Level 3 or 4 (proficient) (%)	Level 2 (%)	Level 1 (%)	Proficient ELs (*n*)	Level 3 or 4 (proficient) (%)	Level 2 (%)	Level 1 (%)	Proficient ELs (*N*)	Level 3 or 4 (proficient) (%)	Level 2 (%)	Level 1 (%)
3	468	35	50	15	475	53	37	11	—	—	—	—
4	606	32	61	7	612	30	56	14	—	—	—	—
5	478	21	44	36	487	18	45	36	486	31	46	23
6	191	22	48	30	196	22	44	34	—	—	—	—
7	210	8	47	45	211	8	63	29	—	—	—	—
8	199	5	45	50	201	9	31	60	201	8	26	66
10	245	4	51	44	249	8	26	66	—	—	—	—
11	—	—	—	—	—	—	—	—	129	11	22	67

III.3. Validity Evidence Based on Consequences of Testing

Details about validity evidence based on consequences of testing are described in Section III.5 Validity Evidence Based on Consequences of Testing in the *2020 KELPA Technical Manual*.

IV. Technical Quality—Other

This chapter provides updated evidence related to the technical quality of the Kansas English Language Proficiency Assessment (KELPA) administered in 2024, including reliability-related evidence, a summary of test results, and a description of ongoing program improvement. For technical-quality-related evidence, refer to Section IV.2 Fairness and Accessibility and Section IV.4 Full Performance Continuum in the *2020 KELPA Technical Manual* (Achievement and Assessment Institute [AAI], 2021a). Quality-control steps were elaborated in Section IV.3.5 Quality-Control Checks in the *2020 KELPA Technical Manual*.

IV.1. Reliability

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

> The term *reliability* has been used in two ways in 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 are reported in two ways: reliability coefficients from classical test theory (CTT) and IRT information functions combined with conditional standard error of measurement. CTT reliability coefficients are sample dependent and were updated using the 2024 data. IRT reliability does not change by test sample, only by test form. Because the same test forms were used from 2020 to 2024, IRT-based reliability is not provided in this section. For detailed information about IRT reliability, refer to Section IV.1 Reliability of the *2020 KELPA Technical Manual*. For CTT reliability coefficients, the student-group reliability was also calculated. Indices of classification consistency and accuracy of different domain performance levels, and interrater agreement on speaking and writing constructed-response (CR) items, are also provided in this section of the current manual. For information about the fairness and accessibility of KELPA, refer to Section IV.2 Fairness and Accessibility of the *2020 KELPA Technical Manual*.

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:

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

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 .80 to .97 across most grades or bands and domains. The exceptions were in kindergarten for reading (.71) and writing (.73). Test length and test reliability are closely related, and shorter tests are usually less reliable. Compared to other domains, kindergarten reading and writing tests had lower reliabilities because these tests had the fewest score points among all grade levels. 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	.88	.94	.71	.73
1	.88	.94	.89	.81
2–3	.90	.94	.90	.87
4–5	.90	.96	.85	.86
6–8	.96	.96	.85	.89
9–12	.90	.97	.86	.81

IV.1.1.1. Student-Group Reliability

Reliability estimates were also calculated by the 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 kindergarten in reading and writing, where reliability coefficients for student groups were lower (consistent with the domain-level coefficient alphas). Also, the student-group reliabilities were similar to the overall reliabilities, with most estimates in the .80s to .90s; reading in kindergarten (mostly in the .70 range or lower) and writing in kindergarten (all in the .70 range) and in grade band 9–12 (group with disabilities only, α = .74) had lower reliabilities. The sample size of each student group can be found in Section IV.2.1.1. Test-Enrollment Data.

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								
K	.87	.88	.87	.89	.87	.88	.87	.85
1	.88	.88	.88	.88	.88	.88	.88	.85
2–3	.90	.90	.90	.90	.90	.90	.89	.88
4–5	.90	.91	.91	.90	.91	.90	.88	.89
6–8	.90	.92	.96	.90	.96	.90	.85	.88
9–12	.90	.91	.91	.90	.91	.90	.86	.91
Speaking								
K	.94	.94	.94	.94	.94	.93	.93	.93
1	.94	.94	.94	.94	.94	.93	.93	.93
2–3	.94	.94	.94	.94	.94	.93	.92	.93
4–5	.96	.95	.96	.95	.96	.95	.91	.95
6–8	.96	.96	.96	.96	.96	.95	.93	.96
9–12	.97	.97	.97	.97	.97	.97	.96	.97
Reading								
K	.70	.71	.65	.79	.64	.79	.69	.69
1	.89	.90	.89	.90	.88	.91	.88	.89
2–3	.90	.90	.90	.90	.90	.91	.88	.90
4–5	.84	.86	.85	.85	.85	.85	.81	.84
6–8	.85	.86	.85	.85	.85	.86	.81	.84
9–12	.85	.86	.86	.86	.85	.86	.81	.86
Writing								
K	.73	.73	.71	.77	.71	.78	.72	.73
1	.81	.82	.81	.82	.81	.83	.81	.80
2–3	.87	.87	.87	.87	.87	.87	.85	.86
4–5	.86	.86	.86	.86	.86	.87	.82	.84
6–8	.89	.89	.89	.89	.89	.90	.86	.88
9–12	.82	.80	.81	.82	.81	.83	.74	.81

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

IV.1.2. Classification Consistency and Accuracy

When an assessment uses achievement or proficiency levels to report test results, accuracy, and consistency of classification into different proficiency levels become key indicators of the assessment's quality. 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 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.

Detailed descriptions of the calculation of two indices can be found in Section IV.1.3 Classification Consistency and Accuracy in the *2020 KELPA Technical Manual*. 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 are particularly important for proficiency classification because students must be at level 4 in all four domains to be considered proficient overall. Classification-consistency indices for the KELPA domain tests ranged from .71 to .98 across most cuts and grades or grand bands. Classification-accuracy indices for the KELPA domain tests ranged from .76 to .98 across most 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						
K	.93	.95	.92	.94	.77	.84
1	.93	.95	.89	.92	.82	.87
2	.96	.97	.92	.94	.88	.91
3	.96	.97	.94	.96	.90	.93
4	.95	.97	.94	.96	.85	.90
5	.96	.97	.95	.96	.85	.90
6	.96	.97	.95	.96	.91	.94
7	.96	.97	.95	.97	.92	.95
8	.96	.97	.96	.97	.92	.94
9	.91	.94	.91	.94	.88	.92
10	.93	.95	.91	.94	.89	.92
11	.93	.95	.92	.94	.85	.90
12	.93	.95	.93	.95	.86	.90
Speaking						
K	.93	.95	.91	.93	.81	.85
1	.96	.97	.93	.95	.74	.81
2	.97	.98	.95	.96	.79	.85
3	.97	.98	.96	.97	.78	.85
4	.98	.98	.97	.98	.85	.90
5	.97	.98	.96	.97	.78	.85
6	.96	.97	.95	.97	.84	.88
7	.96	.97	.95	.96	.84	.89
8	.96	.97	.95	.97	.79	.85
9	.96	.97	.95	.97	.91	.94
10	.96	.97	.96	.97	.92	.94
11	.96	.97	.96	.97	.90	.93
12	.97	.98	.96	.97	.85	.90

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						
K	.71	.78	.83	.88	.91	.94
1	.86	.90	.88	.91	.93	.95
2	.86	.90	.89	.92	.90	.93
3	.90	.93	.89	.92	.88	.91
4	.89	.92	.85	.90	.84	.89
5	.89	.92	.84	.89	.82	.87
6	.90	.93	.85	.89	.87	.91
7	.89	.92	.86	.90	.82	.87
8	.90	.93	.86	.90	.81	.85
9	.86	.90	.89	.92	.92	.94
10	.86	.90	.87	.91	.88	.92
11	.87	.91	.86	.90	.87	.90
12	.87	.91	.86	.90	.85	.89
Writing						
K	.84	.88	.78	.84	.87	.91
1	.93	.95	.84	.88	.77	.82
2	.92	.94	.88	.91	.79	.85
3	.92	.95	.88	.92	.76	.81
4	.92	.95	.89	.92	.78	.84
5	.94	.96	.88	.92	.73	.80
6	.93	.95	.89	.92	.79	.84
7	.94	.96	.88	.91	.77	.82
8	.94	.96	.87	.91	.71	.76
9	.85	.90	.81	.86	.81	.87
10	.87	.91	.81	.86	.79	.85
11	.85	.89	.79	.86	.76	.81
12	.86	.90	.77	.84	.71	.76

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

IV.1.3. Interrater-Agreement Study

The interrater-agreement study provides reliability and validity evidence for the educator-scored test items. KELPA CR item scores ranged from 0 to 3 for both speaking and writing. Refer to Table II-13 in the *2020 KELPA Technical Manual* for the number of educator-scored items for speaking and writing by grade or grade band. Within the same grade or grade band in each domain of speaking and writing, holistic rubrics were used to rate CR item responses instead of item-specific rubrics. The rater training provided at local schools and districts, as well as the training materials provided by KSDE, supplied educators with the knowledge and skills needed to apply the rubrics. The scoring accuracy of CR items, which are scored by educators, relies on consistent and appropriate application of the scoring rubrics. Thus, it is important to assess the consistency with which teachers applied the rubrics. The results from

an interrater-agreement study can offer insights for refining training materials and can also reveal the extent to which raters agreed or differed in their evaluations of each constructed-response (CR) item.

IV.1.3.1. Data-Collection Method

An interrater-agreement study of KELPA writing and speaking CR items was conducted during the 2024 KELPA scoring window (January 29–March 29, 2024). Two options were provided to collect second ratings: Kite® Educator Portal scoring interface or a spreadsheet for targeted school districts. The Kite Educator Portal scoring interface was used for individual raters to manually score questions that are not machine scored, and the spreadsheet option was used for school districts to enter information for a roster of students in batches. To allow two scorers to enter scores for the same student response, students selected for second ratings had two scoring tabs in Educator Portal for all CR items. Score of record (used in score reporting, i.e., the first score entered; refer to Section IV.3.1.2 Educator Scoring of the *2020 KELPA Technical Manual* for more information about how scores were entered) for operational scoring remained the same for all students regardless of whether a student was selected for second rating. District Test Coordinators (DTCs) 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. Note that for speaking, in addition to individual or paired/group scoring, educators could also choose deferred scoring (by listening to audio playback) or simultaneous scoring (by sitting next to students during testing).

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

Writing	Speaking	
	Option 1	Option 2
Individual scoring or paired/group scoring	Individual scoring or paired/group scoring	Deferred scoring or simultaneous scoring

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

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

IV.1.3.2. Sampling

A sample of students taking KELPA for the 2024 administration was selected to receive second ratings for their speaking and writing CR items. Samples selected for two ratings were identified at the

beginning of the testing window when all school districts completed KELPA test registration. Each selected student received two ratings per CR item, targeting about 500 students per grade. A random sample included 18% of kindergarten and grade 1 students, 14% of grades 4 and 5, and 11% for all other grades. Table IV-5 shows the numbers of districts, schools, and students selected for the two ratings.

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

Grade or Grade Band	No. of Districts	No. of Schools	No. of Students
K	37	139	579
1	37	139	582
2–3	45	178	753
4–5	43	162	748
6–8	40	91	623
9–12	43	68	789

Data obtained at the end of the window for hand scoring speaking and writing items were used for rater-agreement analyses. Only an exceedingly small percentage (0–1.1%) of responses with two ratings were collected using the paired/group scoring method for both writing and speaking. For speaking responses scored individually, less than 1% 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, as well as 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 Scoring
K	520–522	240–255
1	531–535	272–281
2–3	693–699	356–372
4–5	713–714	334–351
6–8	603–603	266–282
9–12	773–773	283–318

IV.1.3.3. Raters

KELPA constructed responses were scored by qualified educators. DTCs assigned qualified educators within a school district to score KELPA CR items in speaking and writing. Students assigned to receive two ratings were rated by DTC-assigned educators who were different from the raters who rated the primary score. The first and second ratings are considered interchangeable in score quality since scorers were expected to receive the same level of training and be familiar with the scoring rubrics. Refer to Section II.3.1 Test-Administrator and Scorer Training and Section IV.3.1.2 Educator Scoring in the *2020 KELPA Technical Manual* 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 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 each other is referred to as *adjacent agreement*. In general, an exact agreement level of 75% or above is acceptable for most fields, 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 (i.e., pass vs. fail) rating scale, but kappa can also measure agreement when three or more performance levels are used. *Weighted kappa* distinguishes between the number 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 using expected scores and predicted scores, and it measures the agreement between two ratings; the value typically ranges from 0 (i.e., random agreement between raters) to 1 (i.e., complete agreement between raters). When there is less agreement between raters than expected by chance, the value may go below 0. For example, suppose rater A assigns a sample of *n* subjects across *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):

$$\bar{D}_o = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij} v_{ij,} \qquad \text{(Equation IV-2)}$$

$$\bar{D}_e = \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{m} n_{i}.\, n_{.j}\, v_{ij,} \qquad \text{(Equation IV-3)}$$

where n_{ij} denotes the number of subjects assigned to category *i* by rater A and to category *j* by rater B, $n_{i}.$ denotes the total number of subjects assigned to category *i* by rater A, and $n_{.j}$ 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 certain degree of disagreement when a subject is assigned to various categories by the two raters. Quadratic-weighted kappa is then defined by Equation IV-4 (Fleiss & Cohen, 1973):

$$k_w = \frac{\bar{D}_e - \bar{D}_o}{\bar{D}_e} \qquad \text{(Equation IV-4)}$$

It 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 (Cohen, 1968).

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 a grade or grade band—both overall (i.e., mean percentage of

agreement on all responses, regardless of the scoring method applied) and for the individual scoring method—ranged from 60% (grade band 6–8) to 82% (grade 1). The average percentage of exact-plus-adjacent agreement across items within a grade or grade band—both overall and for the individual scoring method—was 95% or above.

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

Grade or Grade Band	Mean Exact Agreement Across Items (%)		Mean Exact-Plus-Adjacent Agreement Across Items (%)	
	Overall	Individual Scoring	Overall	Individual Scoring
K	79	78	96	95
1	82	82	96	96
2–3	78	78	98	98
4–5	72	72	99	98
6–8	60	60	97	96
9–12	61	62	97	97

Table IV-8 summarizes agreement for speaking items. For speaking responses, the average percentage of exact agreement across items within a grade or grade band—for overall (i.e., mean percentage of agreement on all responses regardless of scoring method applied), the individual scoring method, and the combination of individual and deferred scoring methods—ranged from 66% (kindergarten) to 72% (grade band 9–12). The average percentage of exact-plus-adjacent agreement across items within a grade or grade band—for overall, the individual scoring method, and the combination of individual and deferred scoring methods—was 95% or greater.

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 + Deferred	Overall	Individual Scoring	Individual + Deferred
K	66	66	68	95	95	96
1	68	67	69	96	96	97
2–3	71	71	72	97	97	98
4–5	71	70	70	98	98	98
6–8	67	67	66	96	95	97
9–12	72	72	72	96	97	96

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

Table IV-9 shows the classifications of quadratic-weighted kappa values of KELPA CR items. To be consistent with Table IV-5, Table IV-6, Table IV-7, and Table IV-8, the number of items with excellent or good agreement reported in Table IV-9 is based on responses scored using the individual scoring method for writing items 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 writing items in grades 1–5. For both speaking and writing, the data indicate that rater agreement is higher for lower grades, particularly in the "Excellent" category.

Table IV-9. Summary of Quadratic Kappa Classifications

Grade or Grade Band	No. of Items (% of Domain Items)			
	Writing		Speaking	
	Excellent Agreement	Good Agreement	Excellent Agreement	Good Agreement
K	2 (50)	2 (50)	10 (100)	0 (0)
1	4 (100)	0 (0)	9 (90)	1 (10)
2–3	4 (100)	0 (0)	9 (90)	1 (10)
4–5	4 (100)	0 (0)	9 (90)	1 (10)
6–8	0 (0)	3 (100)	7 (70)	3 (30)
9–12	2 (67)	1 (33)	10 (100)	0 (0)

IV.1.3.4.3. Summary

Individual scoring was the dominant scoring method for writing and speaking items in 2024. Individual scoring paired with deferred scoring was the dominant scoring method for speaking. The average percentage of exact agreement between two raters across items within a grade or grade band ranged from 60% to 82% for writing responses and from 66% to 72% for speaking responses. The average percentages of exact-plus-adjacent agreement across items within a grade or grade band were 96% or greater for writing responses and 95% or greater for speaking responses.

Statistics for the quadratic-weighted kappa show that for writing responses, raters had excellent agreement on all items in grades 1 through 5, with 100% of items in these grades achieving excellent agreement. For grade band 6–8, all items were classified as having good agreement. Grade band 9–12 had a mixture of good to excellent agreement, with 67% of items classified as excellent. For speaking responses, excellent agreement was found across all items in kindergarten and grade band 9–12, while grades 1 through 5 had 90% of items with excellent agreement and 10% with good agreement. Grade band 6–8 showed a mix of 70% of items with excellent agreement and 30% with good agreement.

The degree of rater agreements based on agreement rates and quadratic kappa appears to point to similar conclusions: while kindergarten writing and speaking showed strong rater agreements, lower consistency was observed in higher grades, particularly in grade band 6–8 for both domains.

IV.2. Scoring and Scaling

This section provides test-result summaries for the 2024 administration. For information about the procedures for scoring individual items, scoring the test, scaling, and specific quality-control process followed by AAI and 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*.

IV.2.1. Operational Test Results

The number of students who took KELPA in 2024, 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

entire population and for 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. For students registered for the first time in K–12 schools 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 does not pass the screener is considered an EL and will take KELPA in the spring. Students who scored as proficient on KELPA in 2024 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 viewed a listening or reading test, even if they did not answer any questions, are categorized as having taken the domain test. For the writing and speaking tests, students are categorized as having taken the domain test if a teacher has scored the tests, even if students did not answer any items. Students who took at least one domain test received a score report and will be considered to have participated in the test. Table IV-17 in Section IV.2.2.1. Comparison of Enrollment in the current manual presents the number and percentage of enrolled students who were tested in each grade for KELPA administrations from 2022 to 2024. The participation rate or tested rate for 2024 KELPA, computed as number of students tested divided by number of students enrolled. The participation rates for the 10 State Board of Education (SBOE) districts in 2024 are presented in Table IV-10 by grade or grade band. Kansas has 286 school districts that are grouped into 10 SBOE districts.[1] The participation rates (i.e., tested rates) are all close to 100%.

[1] A school district can be represented by multiple SBOE districts because of its geographic location. Religious school districts are not represented by SBOE districts and are therefore excluded from SBOE-related statistics.

Table IV-10. 2024 KELPA Participation Rates by State Board of Education (SBOE) District and Grade or Grade Band

SBOE 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	1,139	100[a]	1216	100[a]	2,408	100[a]	1,953	100[a]	2,259	99	2,474	95
2	1,658	100[a]	1707	100[a]	3,201	100[a]	2,420	100[a]	2,843	100[a]	3,101	96
3	905	99	863	100	1,402	100[a]	968	100[a]	1,132	100[a]	1,239	98
4	1,664	100[a]	1,790	100[a]	3,346	100[a]	2,610	100[a]	3,088	100[a]	3,434	96
5	1,021	100[a]	974	100[a]	1,830	100[a]	1,479	100	1,853	99	2,031	98
6	433	100[a]	435	100[a]	749	100[a]	632	100[a]	682	100[a]	803	97
7	1,083	100[a]	1087	100[a]	2,052	100[a]	1,693	100[a]	2,101	99	2,397	98
8	900	100[a]	860	100[a]	1,738	100[a]	1,467	100[a]	1,824	99	2,151	98
9	159	99	169	100	236	100	224	100	214	100	236	100
10	990	100[a]	965	100[a]	1,890	100[a]	1,593	100[a]	1,970	99	2,284	98

[a] Calculated as 100% with rounding.

For all tested ELs, Table IV-11 shows the percentage of students in each demographic group by grade.[2] The groups include race, ethnicity, disability status, and gender. The percentage of students in each group was generally consistent across grades, except for a higher proportion of American Indian students and a lower proportion of White students in the upper grades. The majority race group was White, with White students comprising around 76% in kindergarten and slightly decreasing to about 74% by grade 12. In contrast, American Indian students increased from 5.5% in kindergarten to 10.7% in grade 12. The majority ethnicity group was Hispanic, with over 80% Hispanic students from grade 3 onward. The percentage of students with disabilities (SWD) increased steadily, from 11% in kindergarten to 17.4% in grade 12. Gender distribution was about equal, with slightly more male students in each grade, ranging from 50.2% to 56.9%.

[2] To protect student privacy, economically disadvantaged status is not shared with ATS. Therefore, this student group is not included in the comparison.

Table IV-11. Percentage of Tested Students by Demographic Characteristic and Grade

Characteristic	Grade												
	K (*N* = 4,795)	1 (*N* = 4,804)	2 (*N* = 4,824)	3 (*N* = 4,024)	4 (*N* = 3,722)	5 (*N* = 3,274)	6 (*N* = 2,917)	7 (*N* = 2,693)	8 (*N* = 2,653)	9 (*N* = 2,983)	10 (*N* = 2,369)	11 (*N* = 2,098)	12 (*N* = 1,403)
Race													
Black	5.2	5.1	4.9	5.8	5.1	4.9	5.1	5.7	5.5	6.1	6.7	6.5	6.8
American Indian	5.5	6.2	6.3	6.7	6.6	8.1	6.4	7.5	8.3	8.4	9.1	10.6	10.7
Asian	11.9	9.4	10.2	7.8	8.6	8.0	6.6	7.1	7.4	6.3	6.8	7.9	7.0
NHPI	1.1	1.5	1.3	0.9	1.5	1.3	1.3	1.5	1.4	1.4	0.8	0.9	0.8
White	76.3	77.8	77.3	78.7	78.2	77.7	80.5	78.3	77.4	77.7	76.5	74.2	74.7
Hispanic													
Yes	76.3	79.3	79.3	81.5	81.8	82.2	83.3	82.2	83.4	83.5	82.7	82.2	82.1
No	23.7	20.7	20.7	18.5	18.2	17.8	16.7	17.8	16.6	16.5	17.3	17.8	17.9
SWD													
Yes	11.0	12.2	12.7	15.0	16.1	18.7	20.8	18.7	18.8	17.1	19.0	15.3	17.4
No	89.0	87.8	87.3	85.0	83.9	81.3	79.2	81.3	81.2	82.9	81.0	84.7	82.6
Gender													
Female	47.0	47.0	49.8	46.6	44.7	46.6	45.2	43.1	44.2	43.4	43.8	44.1	43.6
Male	53.0	53.0	50.2	53.4	55.3	53.4	54.8	56.9	55.8	56.6	56.2	55.9	56.4

Note. NHPI = Native Hawaiian and Pacific Islander; SWD = student 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 scale score 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, the 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, Table IV-12 shows that the distribution of scale scores was negatively skewed in grades K, 1, 5, 7, and 8, and positively skewed in other grades. For the speaking test, as shown in Table IV-13, the distribution of scale scores was negatively skewed in grades 9–12, and positively skewed in grades K–8. Table IV-14 shows that for the reading test, the distribution of scale scores was negatively skewed in grades 5, 9, and 10, and positively skewed in other grades. Lastly, for the writing test, as indicated in Table IV-15, the distribution of scale scores was negatively skewed in grades 3–6, 8, and 12, and positively skewed in other grades.

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

Grade	M	SD	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	539.24	177.25	0	354	421	492	589	695	1,000
1	500.00	151.41	0	313	404	480	573	717	1,000
2	483.79	170.45	0	315	378	453	541	605	1,000
3	557.67	213.65	0	328	435	503	605	1,000	1,000
4	485.76	169.77	0	292	393	458	535	611	1,000
5	522.14	188.17	0	314	411	491	611	611	1,000
6	460.32	123.94	0	311	384	453	510	615	1,000
7	492.45	142.80	0	311	398	478	552	615	1,000
8	525.19	165.34	0	323	432	510	615	725	1,000
9	448.57	149.81	0	303	338	437	506	622	1,000
10	481.40	157.67	0	315	382	477	547	622	1,000
11	498.54	167.24	0	327	394	477	547	622	1,000
12	521.98	182.84	0	327	407	506	622	622	1,000

Note. M = mean; SD = standard deviation; 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 by Grade for Speaking

Grade	M	SD	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	471.84	187.54	0	250	422	515	580	634	1,000
1	505.67	199.85	0	305	436	512	576	640	1,000
2	498.69	191.15	0	319	421	500	550	616	1,000
3	539.32	218.37	0	330	459	531	575	1,000	1,000
4	531.87	237.79	0	315	435	502	577	1,000	1,000
5	550.52	244.81	0	315	435	502	577	1,000	1,000
6	480.43	204.87	0	306	410	485	536	583	1,000
7	503.22	233.51	0	294	400	496	555	1,000	1,000
8	520.93	249.60	0	289.8	410	496	555	1,000	1,000
9	488.98	289.08	0	0	379	477	535	1,000	1,000
10	517.36	290.87	0	0	405	493	556	1,000	1,000
11	539.68	300.97	0	0	411	502	556	1,000	1,000
12	515.57	315.48	0	0	405	493	556	1,000	1,000

Note. M = mean; SD = standard deviation; 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 by Grade for Reading

Grade	M	SD	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	488.19	130.59	0	363	399	463	552	656	1,000
1	473.37	124.86	0	355	393	439	527	648	1,000
2	459.27	124.07	0	330	377	441	516	606	1,000
3	517.12	157.78	0	362	403	498	566	673	1,000
4	463.85	127.65	0	313	373	442	521	602	1,000
5	498.83	142.92	0	329	404	491	557	665	1,000
6	446.61	108.70	0	314	372	443	511	579	1,000
7	474.84	123.89	0	336	390	463	541	628	1,000
8	501.70	135.00	0	336	407	485	579	699	1,000
9	437.18	106.51	0	314	359	424	502	566	1,000
10	463.86	117.73	0	338	377	454	542	594	1,000
11	476.93	114.81	0	338	393	469	542	631	1,000
12	490.83	122.79	0	348.5	393	485	566	631	1,000

Note. M = mean; SD = standard deviation; 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 by Grade for Writing

Grade	M	SD	Min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	Max
K	504.51	152.85	0	342	426	499	590	637	1,000
1	492.27	160.45	0	336	400	464	588	691	1,000
2	453.85	127.48	0	281	368	465	523	622	1,000
3	502.68	139.74	0	328	420	501	580	687	1,000
4	462.46	136.77	0	285	383	479	532	600	1,000
5	501.28	151.10	0	302	418	504	563	649	1,000
6	462.32	143.23	0	299	380	471	525	596	1,000
7	489.71	167.08	0	299	394	496	557	652	1,000
8	518.79	183.01	0	299	410	525	596	652	1,000
9	418.56	141.79	0	268.7	340	429	511	555	1,000
10	445.30	134.58	0	300	375	448	532	585	1,000
11	463.67	139.62	0	321	393	468	532	585	1,000
12	461.36	166.08	0	300	393	490	555	632	1,000

Note. M = mean; SD = standard deviation; 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 (i.e., levels 1 through 4) is shown 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 considered proficient overall. For listening, the percentage of students in level 4 ranged from 30% (grade 1) to 66% (grade 3). For speaking, the percentage of students in level 4 ranged from 21% (kindergarten) to 45% (grade 10). In reading, level 4 ranged from 11% (kindergarten) to 22% (grades 11 and 12). Finally, for writing, the percentage of students in level 4 ranged from 10% (kindergarten) to 33% (grade 9).

Figure IV-1. 2024 Performance-Level Results for Listening

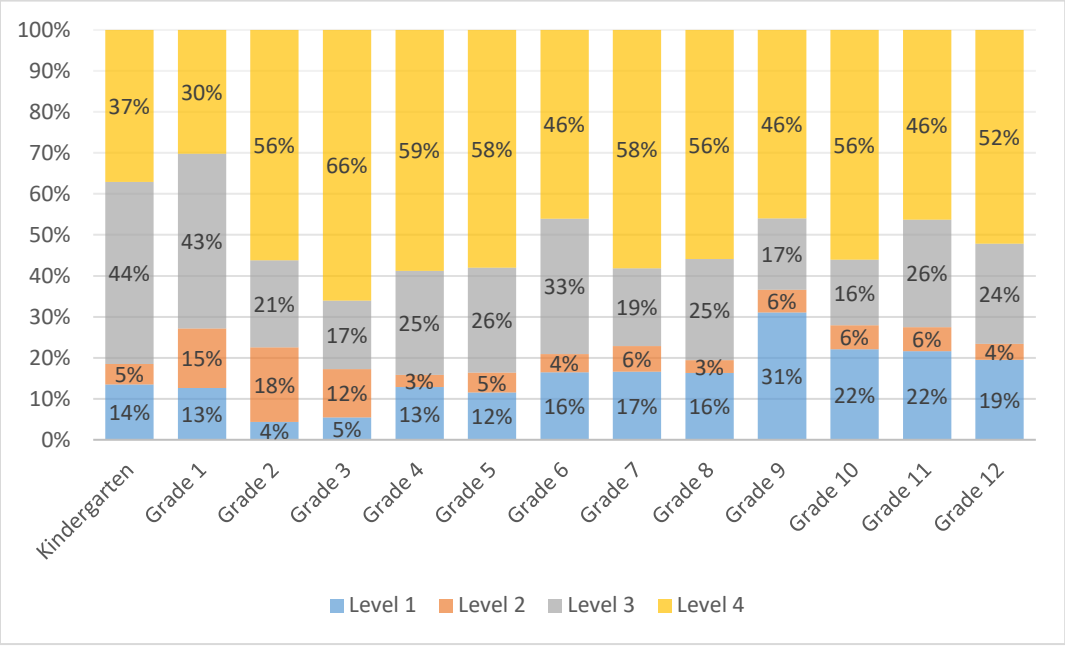


Figure IV-2. 2024 Performance-Level Results for Speaking

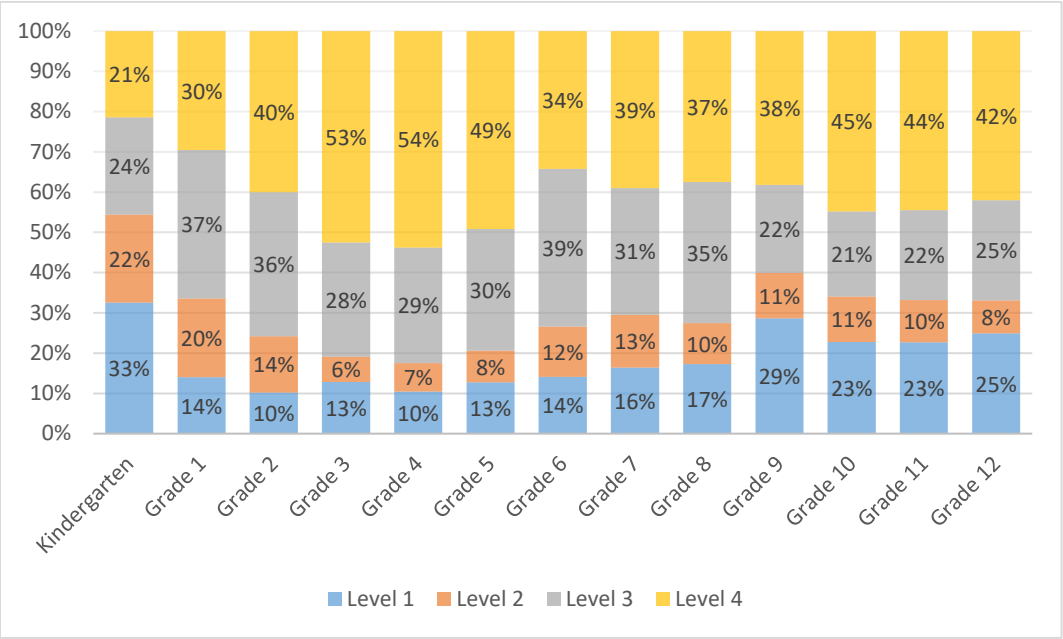


Figure IV-3. 2024 Performance-Level Results for Reading

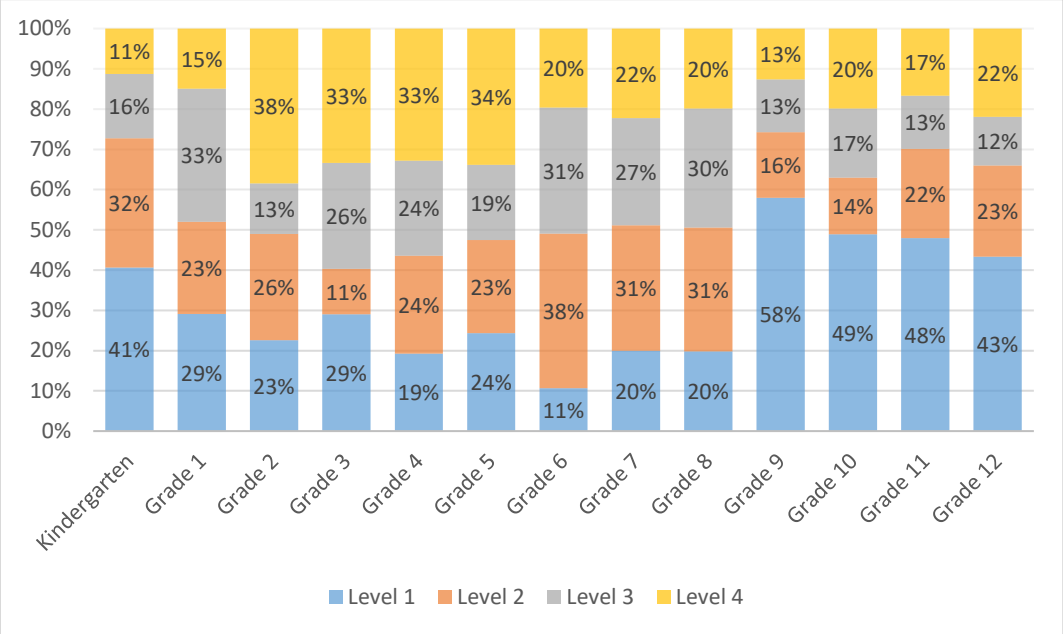
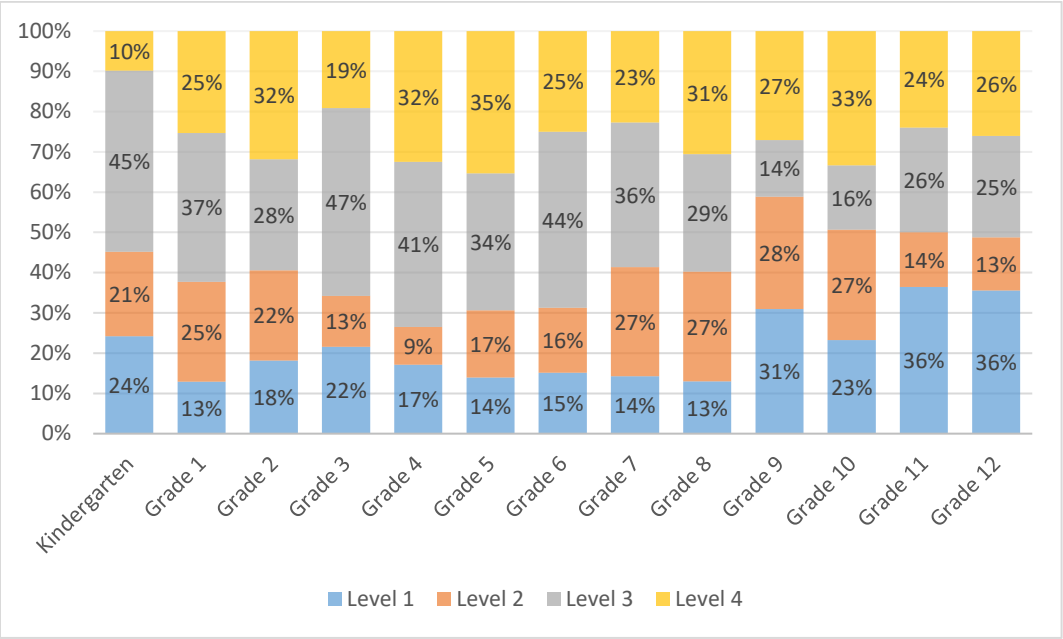
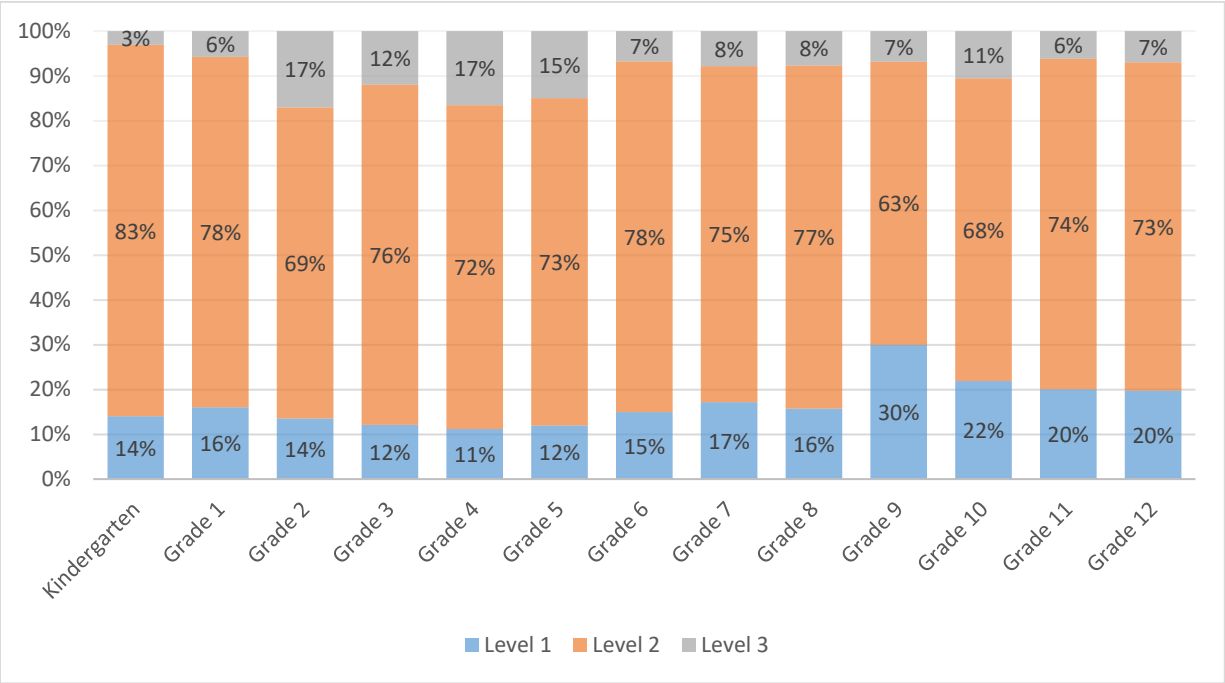


Figure IV-4. 2024 Performance-Level Results for Writing



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 4 (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 not classified as proficiency level 3 or level 1 are at level 2 (i.e., nearly proficient). The overall proficiency levels in 2024 are presented in Figure IV-5. Results indicate that most students were categorized as level 2; the percentages ranged from 63% (grade 9) to 83% (kindergarten). Proficiency rates ranged from 3% (kindergarten) to 17% (grade 4). Kindergarten 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 limited exposure to formal instruction or services for English for speakers of other languages.

Figure IV-5. 2024 Overall Performance-Level Results



IV.2.1.3. Student-Group Test Results

Summaries of average scale scores by demographic groups[3] are presented in Table IV-16, Table IV-17, Table IV-18, and Table IV-19. Asian students had the highest mean scores across most grades and domains. In grade 4, American Indian (AI) students achieved the highest mean scores on the listening test. For the speaking test, Black students scored highest in grade 11, while Native Hawaiian / Pacific Islander (NHPI) and White students shared the top scores in the same grade. Asian students consistently led in reading scores across all grades, and in the writing test, they had the highest mean scores in most grades, except for grade 7, where NHPI students scored the highest.

Across all domains, non-Hispanic students had higher mean scores than Hispanic students in most grades, although their scores were slightly lower in some cases, such as grades 4 and 8 in listening, grades 10 and 11 in reading, and grades 4 and 11 in speaking. Similarly, students without disabilities outperformed students with disabilities across all domains and grades, except in speaking for grades 6–12, listening in grade 9, writing in grade 8, and reading in grades 9 and 11, where their scores were lower or only slightly higher.

Female students outperformed male students in all grades for both the speaking and writing tests. For listening and reading, however, male students had higher mean scores in certain grades, particularly in grades 4, 5, 10, 11, and 12 for listening, and grades 4 and 8 through 12 for reading. These results are consistent with 2023 findings, indicating persistent achievement gaps among demographic groups, despite efforts to develop fair assessments.

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

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

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	521	177	476	140	476	158	535	204	510	174	542	191	460	115	483	121	517	159	457	140	505	173	481	159	503	189
Asian	583	195	531	168	526	192	582	215	486	162	545	203	471	131	518	148	533	153	504	178	515	171	506	170	571	198
Black	526	188	476	141	454	165	535	222	487	192	487	170	438	127	476	138	497	153	442	172	452	140	480	168	524	162
NHPI	534	184	479	138	457	154	578	243	450	143	482	143	454	107	489	98	499	134	425	88	469	209	440	85	506	114
White	533	171	499	151	482	169	559	214	485	171	519	187	461	125	492	145	527	168	444	146	478	154	503	169	521	182
Hispanic																										
Yes	530	172	496	147	477	164	555	213	486	168	521	187	460	123	491	142	527	166	444	144	480	158	496	165	516	182
No	569	192	515	168	511	190	568	217	483	179	526	192	462	129	498	145	518	160	472	174	486	157	508	178	549	183
SWD																										
Yes	492	165	451	148	430	144	507	191	446	143	478	138	442	92	475	116	517	139	449	113	477	135	474	121	498	158
No	545	178	507	151	492	173	567	216	493	174	532	197	465	131	496	148	527	171	448	156	482	163	503	174	527	187
Gender																										
Female	558	182	520	159	494	175	566	214	478	160	521	180	467	124	500	143	526	169	453	150	478	158	497	152	518	171
Male	523	171	482	142	473	166	551	213	492	177	524	195	455	124	487	142	525	163	445	150	484	157	500	179	525	192

Note. M = mean; SD = standard deviation; AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability.

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

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	448	192	476	205	488	201	518	221	547	243	543	228	486	207	500	257	522	254	529	292	553	274	538	300	464	342
Asian	516	180	550	193	529	175	569	198	532	229	569	232	498	169	559	242	535	254	514	281	559	273	514	296	579	315
Black	494	208	534	212	476	193	537	233	502	204	537	261	454	253	503	228	496	224	461	282	483	294	543	322	542	270
NHPI	503	110	456	174	542	191	536	145	522	184	554	226	562	226	568	218	543	212	485	268	502	273	416	366	649	263
White	463	187	499	198	496	192	537	220	532	242	548	248	480	202	496	232	519	251	485	290	512	294	543	300	511	315
Hispanic																										
Yes	457	187	496	198	493	189	535	219	533	241	546	245	477	204	497	233	520	251	486	291	514	294	542	300	504	320
No	519	181	543	201	521	197	560	215	528	225	571	246	496	207	532	235	528	241	502	280	532	276	531	307	569	288
SWD																										
Yes	420	182	448	171	461	155	503	175	505	187	534	204	484	181	510	199	537	218	544	268	537	291	559	279	520	314
No	478	187	514	202	504	195	546	225	537	246	554	253	480	211	502	241	517	256	478	292	513	291	536	305	515	316
Gender																										
Female	485	191	527	203	516	204	554	226	549	247	577	258	495	220	522	249	529	262	501	297	520	293	557	304	539	310
Male	460	184	487	195	481	176	526	211	518	230	528	231	469	191	489	220	515	239	480	282	516	290	526	298	498	319

Note. M = mean; SD = standard deviation; AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability.

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

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	480	120	446	108	449	124	511	159	461	120	504	145	432	90	467	111	492	130	438	93	478	119	472	115	484	128
Asian	561	174	546	153	515	147	543	155	494	143	516	132	464	118	501	136	527	142	481	124	486	118	472	122	521	130
Black	504	157	473	115	456	137	498	145	451	111	466	147	431	108	464	144	487	131	426	119	426	114	451	105	473	113
NHPI	478	133	452	101	459	143	507	125	444	138	507	131	430	110	498	122	470	112	422	86	422	104	443	100	464	97
White	475	116	466	120	452	117	517	160	462	128	497	144	447	109	473	122	501	136	435	105	465	117	481	115	492	122
Hispanic																										
Yes	472	113	461	113	450	117	513	157	462	126	497	143	446	107	473	121	501	135	434	103	465	117	478	115	488	124
No	542	165	522	153	495	144	534	159	474	136	506	145	450	115	486	137	503	136	453	122	458	119	472	116	502	119
SWD																										
Yes	460	134	440	116	413	103	457	123	417	108	449	109	413	83	438	103	469	114	428	80	452	101	455	85	470	111
No	492	130	478	125	466	125	528	161	473	129	510	147	456	113	483	127	509	138	439	111	467	121	481	119	495	125
Gender																										
Female	493	128	475	123	465	127	521	155	459	124	499	138	450	105	482	122	500	129	435	102	457	112	477	108	487	121
Male	484	133	472	127	453	120	514	161	468	130	499	147	444	112	469	125	503	140	439	110	469	122	477	120	494	124

Note. M = mean; SD = standard deviation; AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability.

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

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	491	140	468	145	435	121	490	144	462	133	508	162	451	137	487	148	504	180	432	135	447	130	457	143	454	199
Asian	584	184	573	176	525	137	561	158	507	157	550	174	495	149	528	172	551	200	482	165	488	149	486	149	510	182
Black	528	171	493	175	449	137	488	144	455	128	467	151	446	158	453	146	487	184	395	157	411	126	436	125	445	141
NHPI	483	150	468	132	437	119	527	124	442	136	545	130	483	140	539	154	505	178	391	142	396	163	377	199	492	95
White	490	142	483	156	446	124	498	136	458	135	495	147	462	142	487	168	519	182	415	137	445	134	465	139	459	163
Hispanic																										
Yes	487	139	479	152	442	122	496	135	458	134	497	147	459	141	487	166	517	180	414	136	443	133	463	137	455	166
No	563	179	543	182	498	138	534	155	484	148	523	167	477	151	501	174	529	197	440	165	457	143	468	151	489	166
SWD																										
Yes	450	158	438	147	400	114	457	134	409	119	452	106	425	113	456	136	487	129	422	112	442	101	467	116	445	165
No	511	151	500	161	462	128	511	139	473	138	513	158	472	149	497	173	526	193	418	147	446	141	463	143	465	166
Gender																										
Female	514	154	500	161	462	128	514	140	469	139	520	158	480	143	517	182	536	194	434	154	457	137	487	142	483	172
Male	496	151	485	160	446	127	493	139	457	135	485	143	448	142	469	151	505	172	407	131	437	132	446	135	444	160

Note. M = mean; SD = standard deviation; AI = American Indian; NHPI = Native Hawaiian and Pacific Islander; SWD = student with disability.

IV.2.2. Trend Data

The 2024 KELPA administration was the sixth administration of the new KELPA aligned with the 2018 Standards. The next subsections present changes in enrollment data and performance-level distributions from 2022 to 2024.

IV.2.2.1. Comparison of Enrollment

Enrollment and test participation rates significantly declined across all grades during the 2020–2021 academic year due to the impact of the COVID-19 pandemic but have shown an increase over the past three years (see Table IV-20). In the 2024 administration, a total of 42,882 students were enrolled, and 42,559 students participated in testing, resulting in an overall participation rate of 99%. Participation rates across grades ranged from 94% (grade 12) to 100% (kindergarten through grade 5). Compared to the 2023 administration, the enrollments in 2024 continued to increase for grades 1–5, 11, and 12. On average, the enrollments in 2024 increased by 1% compared to the 2023 administration.

Table IV-20. Number and Percentage of Enrolled and Tested Students by Grade: 2022 Through 2024

Grade	2022				2023				2024			
	No. Enrolled	No. Tested	Participation %	% Enrollment Change (2021 to 2022)	No. Enrolled	No. Tested	Participation %	% Enrollment Change (2022 to 2023)	No. Enrolled	No. Tested	Participation %	% Enrollment Change (2023 to 2024)
K	4,638	4,597	99	8	4,603	4,579	99	–1	4,808	4,795	100	1
1	4,471	4,436	99	1	4,767	4,742	99	7	4,809	4,804	100	1
2	4,376	4,342	99	1	4,445	4,405	99	2	4,826	4,824	100	1
3	3,929	3,884	99	0	3,949	3,915	99	1	4,032	4,024	100	1
4	3,623	3,583	99	2	3,628	3,585	99	0	3,730	3,722	100	1
5	3,114	3,061	98	2	3,194	3,161	99	3	3,277	3,274	100	1
6	2,692	2,639	98	–1	2,770	2,733	99	3	2,932	2,917	99	0
7	2,684	2,619	98	6	2,665	2,627	99	–1	2,713	2,693	99	0
8	2,424	2,387	98	–2	2,596	2,564	99	7	2,669	2,653	99	0
9	2,844	2,736	96	11	2,724	2,683	98	–4	3,040	2,983	98	0
10	2,205	2,093	95	–12	2,487	2,443	98	13	2,409	2,369	98	0
11	2,003	1,878	94	–16	1,973	1,913	97	–1	2,141	2,098	98	1
12	1,823	1,510	83	–13	1,727	1,530	89	–5	1,496	1,403	94	5
Total	**40,826**	**39,765**	**97**	**0**	**41,528**	**40,880**	**98**	**2**	**42,882**	**42,559**	**99**	**1**

Note. Positive values indicate a percentage increase; negative values indicate a percentage decrease.

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 from 2022 through 2024 by domain and grade, with a focus on 2024 compared to 2022. For listening (Figure IV-6), in 2024, the level 4 percentages increased in grades K and 1 and decreased in the remaining grades (2 through 12). For speaking (Figure IV-7), in 2024, the level 4 percentages fluctuated within the range of 2% in grades K, 1 to 7, 10, and 12; grade 11 increased about 5% from 2022; and grades 8 and 9 decreased 4% and 6% respectively. For reading (Figure IV-8), in 2024, the level 4 percentages remained stable within a range of 2% in grades K, 1, 2, 9, 11, and 12, and decreased over 2% in other grades. For writing (Figure IV-9), in 2024, the level 4 percentages remained stable within 2% in grades L, 1, 2, 3, 7, and 8, and decreased in the remaining grades.

Figure IV-6. Comparison of Performance-Level (PL) Results From 2022 Through 2024 for Listening

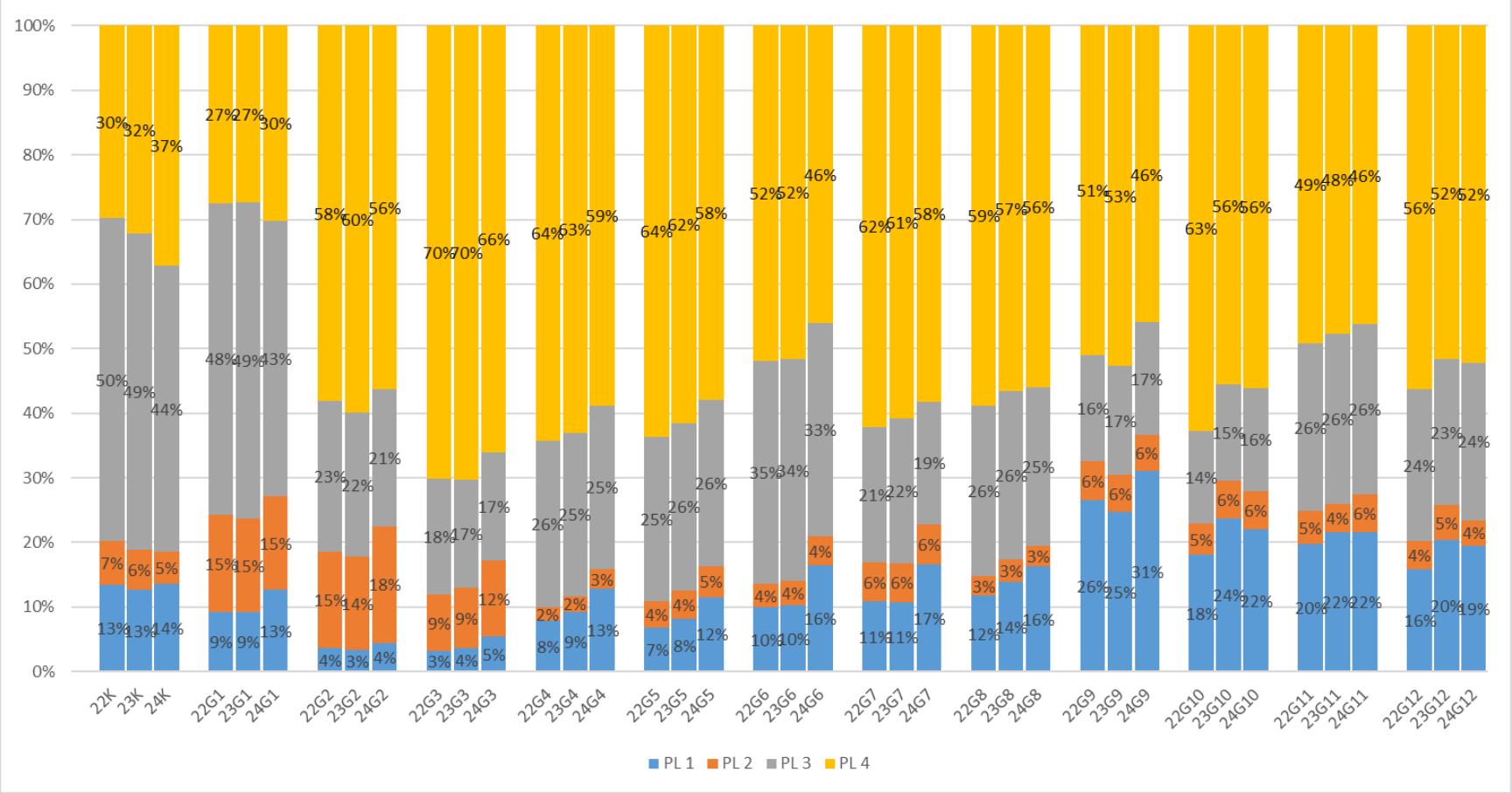


Figure IV-7. Comparison of Performance-Level (PL) Results From 2022 Through 2024 for Speaking

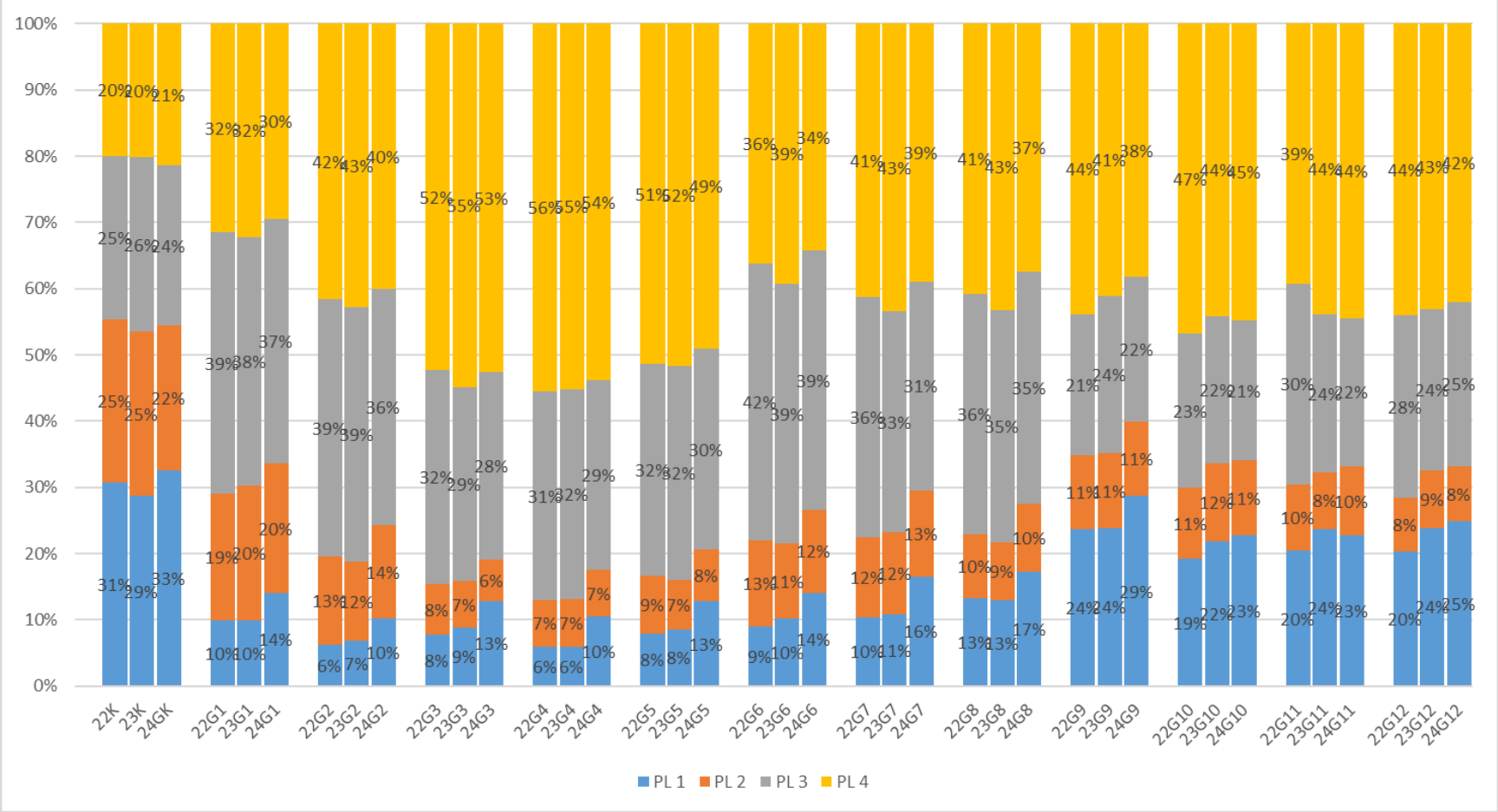


Figure IV-8. Comparison of Performance-Level (PL) Results From 2022 Through 2024 for Reading

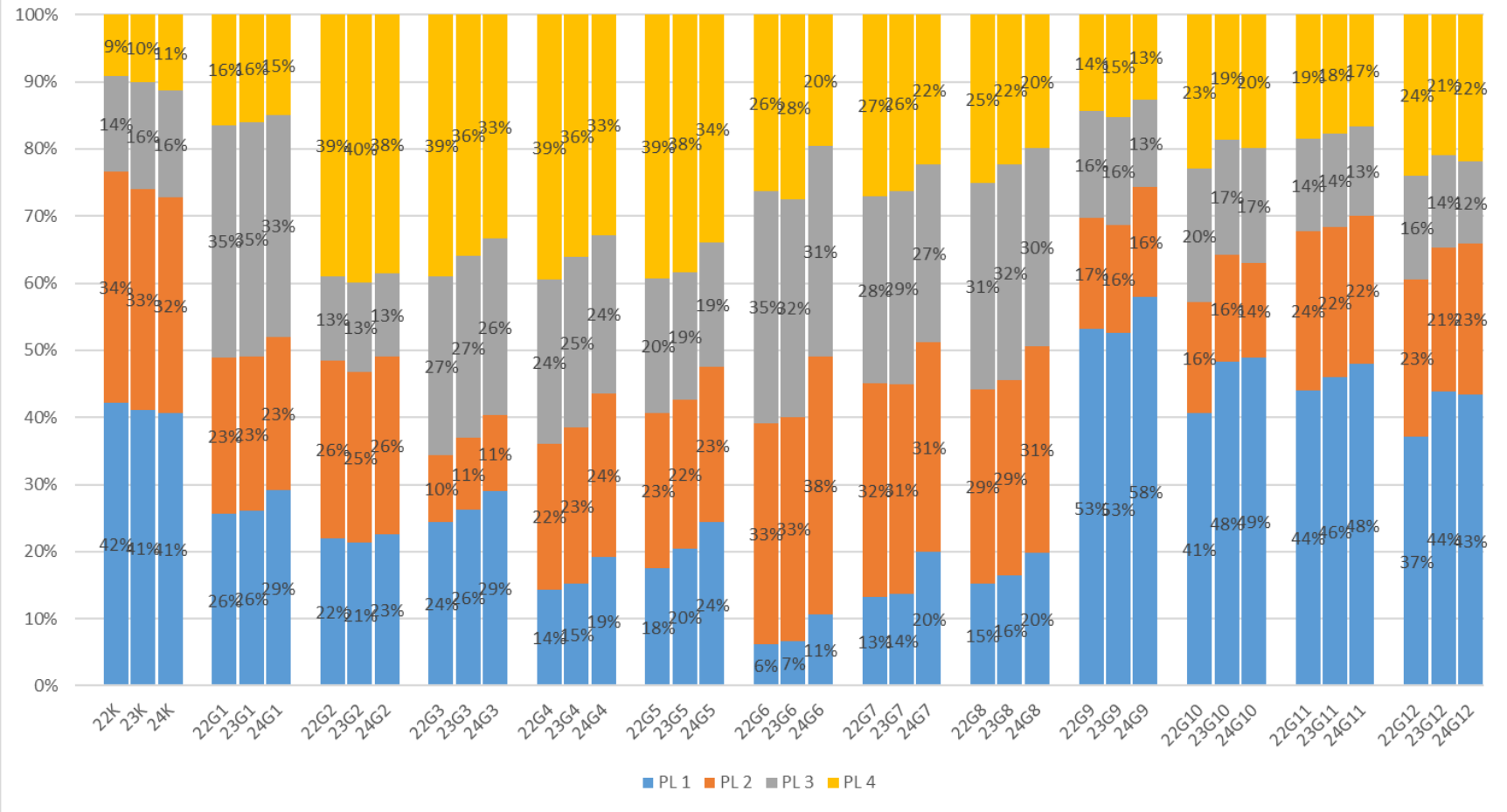


Figure IV-9. Comparison of Performance-Level (PL) Results From 2022 Through 2024 for Writing

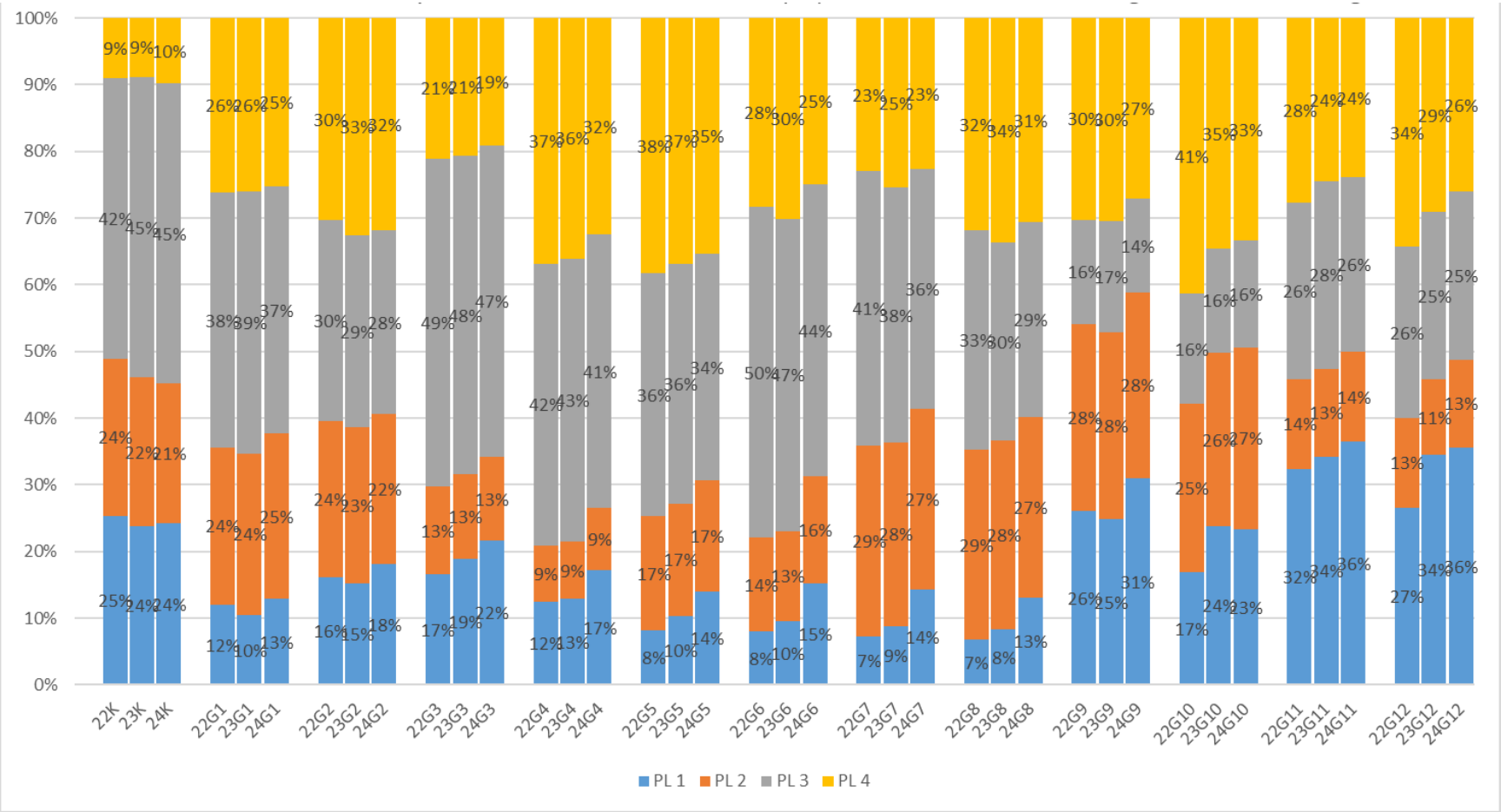
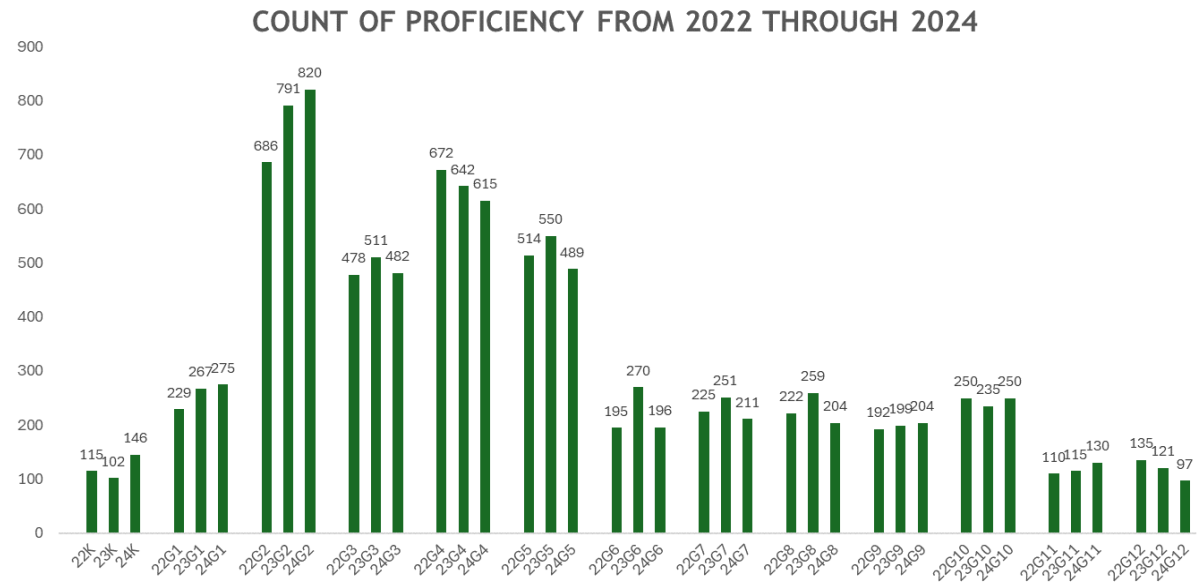


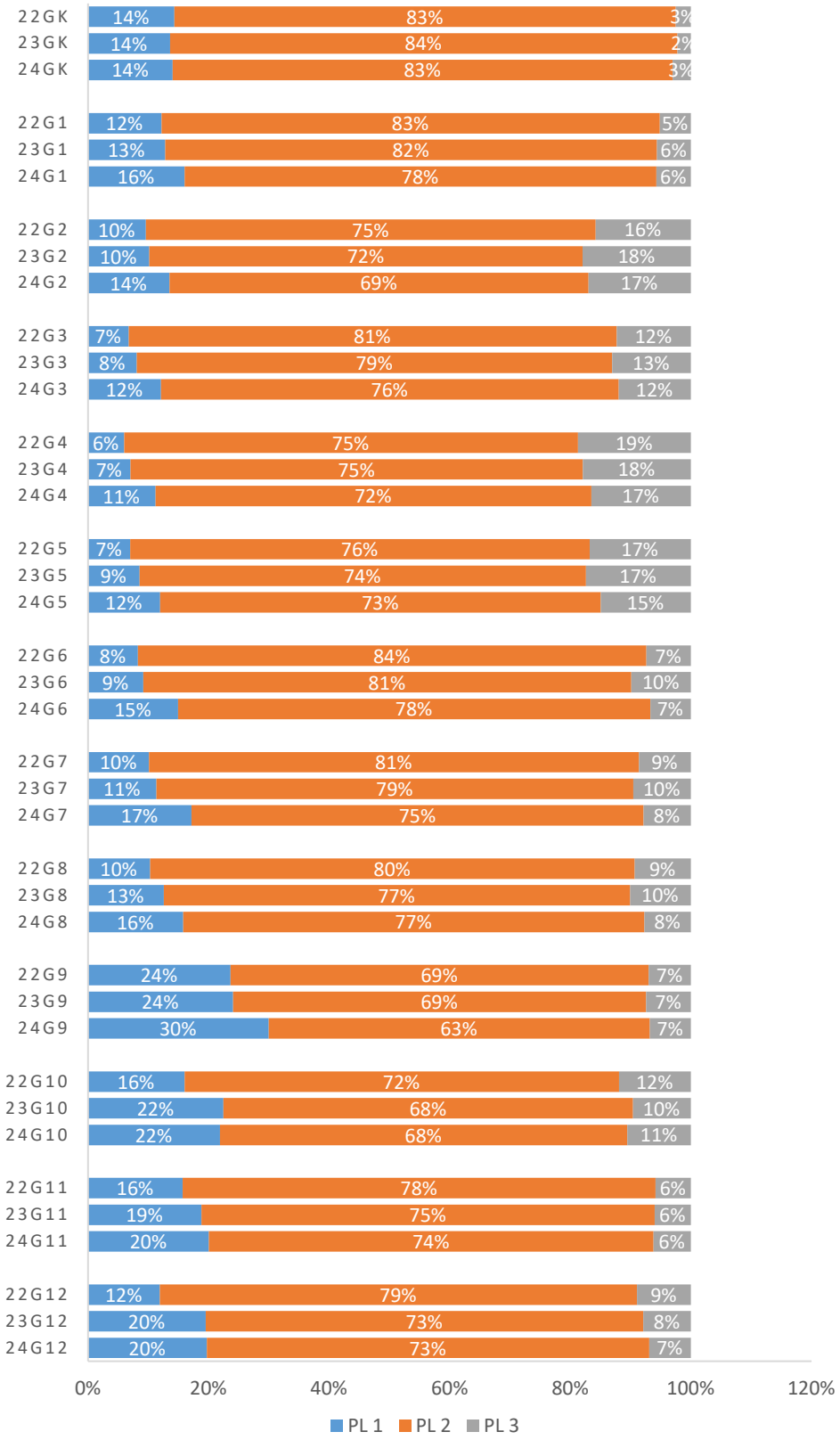
Figure IV-10 presents a bar chart comparing proficiency counts from 2022 through 2024. Proficiency counts in 2024 show a declining trend across grades. The highest count, 820, occurs in grade 2, with a further decline as the grades advance. By the later grades, the proficiency count drops, reaching a minimum of 97 at grade 12. Proficiency in 2024 exhibits a consistent decline in later grades when compared to the previous years.

Figure IV-10. Comparison of Count of Proficiency From 2022 Through 2024



The trend of the overall proficiency rates (PL3) is provided in Figure IV-11. From 2022 to 2024, the overall proficiency rates remained the same for grade 5. For grades 1, 6, 10, and 11, proficiency increased by 1–2%. Grades 2, 4, and 9 showed more noticeable increases in proficiency, with gains of 1–3%. On the other hand, proficiency rates decreased slightly in grades 7 and 12, with drops of 1–2%. Grades K, 3, and 8 had stable proficiency rates, with minimal changes from 2022 to 2024.

Figure IV-11. Comparison of Overall Performance-Level (PL) Results From 2022 Through 2024



IV.3. Full Performance Continuum

The overall performance level of KELPA is a summary of students' performance in the four domains of listening, speaking, reading, and writing. The overall performance encompasses a full spectrum of student performance profiles in each of the four domains. There are 256 profiles for all domain performance, because each domain performance can be 1, 2, 3, or 4. For proficiency level 1, there are 16 profiles because each domain can be either 1 or 2. There is only one profile (4444) for proficiency level 3. Therefore, there are 239 (256 – 16 – 1) profiles for level 2. Table IV-21 shows the number of performance profiles observed for overall proficiency by grade or grade band for proficiency levels 1 and 2. It shows that 16 profiles were observed across all grades and grade bands for students in proficiency level 1. For students in proficiency level 2, the number of profiles observed ranged from 165 to 199, with the highest number observed for students in grades 9–12. The high number of profiles observed compared to all profiles indicates that KELPA results cover a broad spectrum of student performance in the four domains.

Table IV-21. Number of Performance Profiles by Grade or Grade Band

Proficiency Level	Grade or Grade Band	Number of Profiles Observed
1	K	15
	1	16
	2–3	16
	4–5	16
	6–8	16
	9–12	16
2	K	174
	1	180
	2–3	173
	4–5	165
	6–8	172
	9–12	199

Table IV-22 shows the most observed domain performance profiles of students in overall proficiency levels 1 and 2, listed for each grade or grade band. For proficiency level 1, the top four profiles are consistent across grades, with most students having a profile of 1111, indicating level 1 performance in listening, speaking, reading, and writing. Other common profiles, such as 1112 and 1121, indicate minor variations in one domain, typically at level 2. For proficiency level 2, the most observed profile for kindergarten students is 3111, indicating listening skills at level 3 and level 1 in speaking, reading, and writing. For students in grades 2–3, the most common profile is 4443, showing level 4 in listening, speaking, and reading, and level 3 in writing. For older students in grades 6–12, profiles such as 4434 and 4334, reflecting increased performance in listening, speaking, and writing, but slightly lower reading skills, become more prevalent. This variation suggests that as students progress through the grades, the composition of their domain-specific proficiency becomes more differentiated.

Table IV-22. Domain Performance Profiles by Grade or Grade Band

Overall Proficiency	Domain Performance Profiles in Listening, Speaking, Reading, and Writing					
	Grade K	Grade 1	Grades 2–3	Grades 4–5	Grades 6–8	Grades 9–12
Level 1	1111	1111	2111	1111	1111	1111
	1112	1112	1111	1211	1121	1211
	1121	2212	2211	1121	1211	1112
	2111	2111	2222	1112	1112	2111
	1211	2112	2212	2211	1221	1212
Level 2	3111	3333	4443	4443	4434	4434
	4323	4434	4433	4434	4433	4424
	3112	4333	4344	4433	4333	4423
	3223	3323	4343	4344	4443	4443
	4433	4433	4333	4423	3323	4433

V. Inclusion of All Students

This chapter provides a summary of the frequency of accommodations used in the 2024 Kansas English Language Proficiency Assessment (KELPA) administration, as well as 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 in the *2020 KELPA Technical Manual* (Achievement and Assessment Institute [AAI], 2021a).

V.1. Accommodations

All students who are identified as English learners, including those who need accommodations, must take KELPA. 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; refer to The Kansas Accessibility Manual. Accessibility tools, which vary by testing program, are available for all students taking various components of the Kansas assessments in the Kansas Assessment Program [4] (KAP). Without altering the assessment's validity, score interpretation, reliability, or security, assessment accommodations provide equitable access during assessments for students with disabilities. 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.

More details about KELPA accommodations can be found in the KELPA Examiner's Manual, 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]). The Kite Educator Portal Manual for Test Coordinators provides additional information about accommodations for Kite® tools.

V.1.1. Selection of Accommodations

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; refer to the KELPA Examiner's Manual for details. According to the Kite Educator Portal Manual for Test Coordinators, accommodations must be recorded in a PNP or in Access Profile in Educator Portal. To use an accommodation not listed in Tools and Accommodations for the Kansas Assessment Program, the examiner should contact the District Test

[4] 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.

Coordinator, who will send the request to the Kansas State Department of Education (KSDE). Refer to Section V.4.1 Selection of Accommodations in the *2020 KELPA Technical Manual* for guidelines that apply to accommodation selection.

V.1.2. Frequency of Accommodations

In addition to accommodations that are built-in features of the Kite system, test administrators provide some accommodations that are allowed locally for KELPA. 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 availability of test accommodations for individual students. Table V-I presents the number of students who took KELPA in Kansas in 2024 and had PNP accommodations. The summary in the table shows that accommodations were requested for no students in kindergarten, for one student in grade 1, for 14 students in grade band 2–3, for 29 students in grade band 4–5, for 57 students in grade band 6–8, and for 156 students in grade band 9–12. The most frequent accommodation (i.e., 189 students) was auditory calming, which provides relaxing, peaceful background music while a student takes the test. The second- and third-most frequent accommodations (i.e., 30 and 18 students) were whole screen magnification and color contrast, respectively.

Table V-1. Number of Students Using Accommodations by Grade or Grade Band

Grade or Grade Band	No. of Students Using Accommodation								
	ASL	Auditory Calming	Color Contrast	Color Overlay	Masking	Reverse Contrast	Switches	WSM	Total
K	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	1
2–3	0	11	0	0	0	1	0	2	14
4–5	3	16	0	0	0	0	4	6	29
6–8	1	39	8	0	0	1	0	8	57
9–12	2	123	10	3	1	1	3	13	156
Total	6	189	18	3	1	3	7	30	257

Note. ASL = American Sign Language; WSM = whole screen magnification.

V.2. Domain Exemptions

In some situations, students may be exempt from taking a domain test. Special-circumstances codes available in Educator Portal allow school districts to manage test exemptions. Domain exemption requests were reviewed and approved by KSDE. Exempted domains were not 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. Table V-2 shows the number of students exempted from testing by domain for the 2024 administration. Speaking is the domain with the most students exempted from testing, with a total student count of 70 across all grades. Writing had 42 students exempted from testing. The lowest number of students exempted across all grades were in Reading (23) and listening (24).

Table V-2. Number of Students Exempted for Testing by Domain and Grade

Grade	Listening	Speaking	Reading	Writing
K	0	20	0	4
1	0	30	0	0
2	0	0	0	0
3	0	20	0	38
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	23	0
10	0	0	0	0
11	24	0	0	0
12	0	0	0	0
Total	24	70	23	42

VI. Academic Achievement Standards and Reporting

The Kansas English Language Proficiency Assessment (KELPA) standard-setting event was held 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 to set cut scores. The Bookmark standard-setting method (Cizek & Bunch, 2007) was used to establish cut scores. For detailed procedures of the KELPA standard-setting event, as well as information about evaluations of the standard-setting method and event, refer to Chapter VI of the *2020 KELPA Technical Manual* (Achievement and Assessment Institute [AAI], 2021a). Because there were no updates to anything related to standard setting or performance level during the 2023–2024 school year, this chapter briefly updates information about student score reports.

VI.1. Reporting

The 2024 KELPA testing window ended on March 8, 2024, and the scoring window closed on March 29, 2024. KELPA student reports were made available to all school districts on April 30, 2024, and in the Parent Portal on May 07, 2024.

VI.1.1. Student Reports

Performance levels for listening, speaking, reading, and writing were used to determine overall proficiency level, which is defined by the Kansas State Department of Education (KSDE). To be considered proficient (i.e., level 3 on overall proficiency) and eligible to exit the English for speakers of other languages (ESOL) program, students must receive 4s on all domain scores. Students who receive all 1s or 2s on the domain scores are considered not proficient (i.e., level 1 on overall proficiency). Students who do not meet the criteria for either level 1 or level 3 on overall proficiency are considered nearly proficient (i.e., level 2). 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 for the 2022 to 2024 administrations:

> 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.

The 2024 KELPA student report kept the same format and information used in the 2023 student report. Both the overall proficiency level and the domain performance levels are provided in the student report. The overall proficiency levels are derived from student performance in the four domains.

VI.1.2. Interpretive Guides

Descriptions of what students should know and be able to do at each performance level are provided in the reports. Nontechnical language is used to assist readers in interpreting the information in the reports. In addition, the Educator Guide to KELPA Student Score Reports and the Parent Guide to KELPA Student Score Reports (and its Spanish translation) are provided to assist the interpretation of the score reports. 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 English proficiency.

VII. Ongoing Maintenance for KELPA Program

This chapter summarizes the ongoing program improvements and maintenance for the Kansas English Language Proficiency Assessment (KELPA).

VII.1. Updates for the 2024 Administration

Test Development completed a multi-year enhancement effort of the KELPA rater-training materials in 2023, with all rater-training materials for the KELPA Screener and assessment available to educators in winter 2024. For 2024, all speaking and writing prompts for the KELPA Screener, and the current assessment have rater-training materials. The purpose of the updated materials is to support educators in applying rubrics to student responses to speaking and writing prompts, which enhances the validity of the constructed-response item scores.

VII.2. Plans for Future Administration

VII.2.1. Multiple Test Forms

The KELPA program utilizes a pre-equated design, where operational items with known item statistics are used to develop scoring tables prior to test administration. The Kansas State Department of Education (KSDE) plans to use the same methodology to expand the KELPA item pool in the future. Newly developed items will be embedded in the operational test administration for field testing. The operational items will serve as linking items to place the field-test items on the KELPA item response theory (IRT) scale. Items in the expanded item pool will be used to develop new test forms for KELPA.

VII.2.2. Improve Reliability

Classification consistency and accuracy analyses at domain performance cut points provided information about whether KELPA provides accurate and reliable classification around the three performance cut points. The results of classification consistency and accuracy analyses may guide item development to improve classification accuracy at the level 3 and level 4 (proficient) cut points when KSDE expands the KELPA item pool for future administrations. For example, the grade K listening test has a classification consistency of .77 between performance levels 3 and 4 (lower than the classification consistency between levels 1 and 2, and between levels 2 and 3), which indicated the current grade K listening test may need more difficult items to differentiate students at higher ability levels. Similarly, the grade K reading test has a lower classification consistency between performance levels 1 and 2, which indicates the need for easier items to differentiate students at lower ability levels.

VII.2.3. Linguistic Process

The KELPA item pool was developed between 2019 and 2020. Through an external item-review process, Kansas educators reviewed items for item content and fairness, and also informally evaluated the intended linguistic-processing complexity needed for responding to the test items.

In the future expansion of the KELPA item pool, a formal evaluation of linguistic-processing complexity will be included as part of the external item-review process, so that educators who have experience working directly with English learners (ELs) can evaluate whether KELPA items elicit the intended linguistic-processing complexity from those students. The educator-reviewers will use a rubric to

determine the linguistic-processing complexity across the receptive and productive language-processing domains, including reading, listening, speaking, and writing. Reviewer feedback will be solicited and documented for each newly developed item and then analyzed by item-development specialists trained in EL item development. Items will be accepted when the response process is aligned with the intended linguistic process. Some items may be revised to better align with the intended linguistic process.

References

Achievement and Assessment Institute. (2021a). *2020 KELPA technical manual*. Kansas State Department of Education. https://ksassessments.org/sites/default/files/documents/technical-manuals/KELPA_Technical_Manual_2020.pdf

Achievement and Assessment Institute. (2021b). *2021 KELPA technical manual*. Kansas State Department of Education. https://ksassessments.org/sites/default/files/documents/technical-manuals/KELPA_Technical_Manual_2021.pdf

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. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220. https://doi.org/10.1037/h0026256

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

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

Elementary and Secondary Education Act, 20 U.S.C. § 6301 (1965). https://govinfo.gov/content/pkg/COMPS-748/uslm/COMPS-748.xml

Every Student Succeeds Act, 20 U.S.C. § 6301 (2015). https://www.congress.gov/bill/114th-congress/senate-bill/1177

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). ERIC; Westat, Center for Educator Compensation Reform. https://files.eric.ed.gov/fulltext/ED532068.pdf

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.

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