# Evidence Guide

## Interim Assessment
## English Language Arts
## Grades 3-8

v1.0

# Gateway 1

Criterion 1.1: 1.1.a, 1.1.b

Criterion 1.2: 1.2.a, 1.2.b, 1.2.c, 1.2.d, 1.2.e, 1.2.f, 1.2.g

Criterion 1.3: 1.3.a, 1.3.b, 1.3.c

# Gateway 2

Criterion 2.1: 2.1.a, 2.1.b, 2.1.c, 2.1.d

Criterion 2.2: 2.2.a, 2.2.b, 2.2.c, 2.2.d

Criterion 2.3: 2.3.a, 2.3.b, 2.3.c, 2.3.d

Criterion 2.4: 2.4.a, 2.4.b, 2.4.c, 2.4.d

# Gateway 3

Criterion 3.1: 3.1.a, 3.1.b, 3.1.c

Criterion 3.2: 3.2.a, 3.2.b, 3.2.c

Criterion 3.3: 3.3.a, 3.3.b, 3.3.c

Criterion 3.4: 3.4.a, 3.4.b, 3.4.c

## Preamble

Since 2015, EdReports has published over 900 grade- and course-level reviews of core instructional materials. These reports have empowered over 1100 school districts, serving more than 13 million students, in their selection of quality curricular materials.

During that time, many districts asked EdReports about the alignment between interim assessment products and college- and career-ready (CCR) standards. They noted that while educators used interim assessment results to adapt curriculum and adjust instruction, there was a lack of evidence showing alignment between assessment products and CCR standards. EdReports responded to these inquiries by designing the Interim Assessment (IA) Review Criteria. Similar to the instructional materials criteria, the IA Review Criteria are based on the focused principles of instructional shifts/innovations essential to college and career readiness and the Common Core State Standards (CCSS).

Applying a similar framework of standards alignment based on the major instructional innovations will allow for the broadest use of the EdReports interim assessment reviews. Moreover, the familiarity of the EdReports process will ensure the hundreds of districts that have grown to understand our definition of standards alignment will see consistency across all reports. And while the CCSS are not the only example of CCR standards, EdReports recognizes the CCSS as the most widely-known and used educational standards nationwide. Therefore, in reviewing interim assessments, EdReports is asking for test events and assessment design specifications that would be used in states adhering to a version of CCSS.

We realize that measuring progress on learning standards is not the same as providing core materials to teach it. Providing evidence on the alignment of an assessment product—particularly one that is computer adaptive—brings unique challenges. The Implementation Guide Preamble lays out how EdReports has designed the reviews to allow for the myriad of assessment claims and designs to be understood and recognized in our reports.

Specifically for ELA, EdReports structured Criterion 1.1 and 1.2 to focus on the three instructional shifts/innovations within the design, development, and operationalizing of literacy assessments in order to measure students' performance in these areas: use of texts that are challenging enough to meet the demands of college and careers; development of text-dependent questions and tasks that require close reading and analysis; and building content knowledge through reading rich informational texts.

Complex Texts and Academic Language:

- High-quality, grade-appropriate complex texts are imperative to ascertain accurate measures of reading proficiencies that mirror classroom and standards' expectations. Within these texts should be a focus on academic vocabulary, i.e., words used in multiple disciplines and linked to contextual understanding to achieve reading comprehension.
    - Together, Indicators 1.1.a and 1.1.b evaluate whether the intended assessment design assures students are assessed using high-quality passages, including academic vocabulary in context that reflect the complexity and quality of texts called for in CCR standards.

- Indicator 1.2.a evaluates whether the actual passages provided on the tests align to the assessment design and to CCR expectations for high-quality passages.
- Indicators 1.2.b, 1.2.c, and 1.2.d evaluate if items intended to assess students' understanding of vocabulary are designed to focus appropriately on academic words.

Text-Dependent Questions:

- High-quality assessment items for ELA should be written in such a way that students rely on the text for their responses, demonstrating the skills for comprehension they have developed as a part of high-quality instruction.
  - Because each ELA domain has a distinct set of standards, they are reviewed as separate indicators: reading (1.2.d), writing (1.2.e), speaking and listening (1.2.f), and foundational skills (1.2.g). However, each indicator measuring an ELA domain is grounded in the expectation that item stimuli will "require students to use or provide textual evidence," "write using evidence," "evaluate text information," or "use textual context."
  - Additionally, EdReports recognizes that districts may prefer to measure some domains—particularly writing, speaking and listening, and foundational skills—closer to the point of instruction and reserve assessment of these standards for the classroom teacher. Therefore, the indicators for these individual domains will only be reviewed if the test claims to measure the associated standards.

Building Knowledge Through Rich Nonfiction Texts:

- The skills needed to grapple with complex text in content areas such as science and history demand students be exposed to an increasing proportion of nonfiction texts as they progress through the grades. To that end:
  - Indicators 1.1.a, 1.1.b, and 1.2.c have been crafted to ensure that the balance of literary and informational texts is grade appropriate to meet CCR standards.

# Gateway 1: Alignment, Fairness, & Accessibility

## Criterion 1.1

### Test Development Alignment

Assessment design specifications align to the expectations of college- and career-ready (CCR) standards.

## What is the purpose of this Criterion?

The development of any quality assessment requires not only a comprehensive grasp of the content to be assessed but also a strategic plan for fair and equitable assessment design and delivery. This criterion focuses on the interim assessment's adherence to the conventions and processes of high-quality assessment design.

## Research Connection

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects
- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards
- Supplemental Information for Appendix A of the Common Core State Standards for English Language Arts and Literacy: New Research on Text Complexity
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (Eds.). (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.
- Shinn, E., & Ofiesh, N. S. (2012). Cognitive diversity and the design of classroom tests for all learners. *Journal of Postsecondary Education and Disability*, 25(3), 232-255.

## Scoring:

| Meets Expectations | Partially Meets Expectations | Does Not Meet Expectations |
|---|---|---|
| • 8 points | • 6 points | • <6 points |

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.1 | Assessment design specifications align to the expectations of college- and career-ready (CCR) standards. |
| Indicator 1.1.a | Assessment design specifications provide clear expectations and detailed guidance to support the development of high-quality, CCR standards-aligned materials.<br>• Assessment rationale explains the design of the assessment, the benefits of the assessment, and a research foundation grounding the assessment process.<br>• Item development documentation is sufficiently robust to support the writing and review of items measuring CCR standards.<br>• Across all item types, assessment design specifications provide clear scoring information and/or rubrics to evaluate students' levels of understanding with respect to CCR standards being measured.<br>• Passage selection documentation details a process for the review and selection of texts based on the expectations of CCR standards.<br>• Item development documentation includes a description of processes used to ensure items are content-accurate and without technical or editorial flaws. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• The development documentation provides a clear rationale regarding the design of the assessment, the benefits of the assessment to learners or other stakeholders, and a foundation of research grounding the assessment or assessment process.<br>• Item development documentation and writing guidelines provide clear direction in the writing of high-quality CCR items.<br>• Item development documentation provides clear direction for a review process resulting in high-quality CCR items | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• The development documentation provides a rationale regarding the design or benefits of the assessment and may or may not provide research grounding the assessment or process.<br>• Item development documentation and writing guidelines provides guidance in the writing of high-quality CCR items, but is not specific or comprehensive enough to ensure all items are high quality.<br>• Item development documentation provides direction for a review process resulting in high-quality CCR items. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• The development documentation lacks any rationale regarding the design, benefits, or research grounding the assessment or process.<br>• Item development documentation and writing guidelines lack guidance in the writing of high-quality CCR items.<br>• Item development documentation lacks guidance for a review process resulting in high-quality CCR items.<br>• Scoring specifications and criteria are either not present or not aligned.<br>• Guidelines supporting passage selection are either not present |

| | | |
|---|---|---|
| • Scoring specifications and criteria are standards aligned, clearly communicated, and adequate to ensure inter-rater reliability.<br>• Guidelines supporting passage selection are detailed and clearly reflective of CCR standards.<br>• There is evidence of processes used to ensure content accuracy, technical quality, and editorial precision of the items. | • Scoring specifications and criteria are present and aligned but not specific enough to ensure consistent score reporting or inter-rater reliability.<br>• Guidelines supporting passage selection are general in design and reflective of CCR standards.<br>• There is some evidence of processes to ensure content and editorial accuracy but the processes lack the specificity necessary to ensure consistency and technical quality. | or are not reflective of CCR standards.<br>• Assessment documentation lacks review processes and/or guidance to ensure content accuracy, technical quality, and editorial accuracy of items. |

## About this indicator:

What is the purpose of this Indicator?

Planning is essential to the development of high-quality assessments. High level planning may begin with construct maps and assessment frameworks to inform test specifications, assessment blueprints, and other guiding documents. These assessment design specifications also outline and prescribe processes for item creation, guidance to item writers, and systems for process checks and balances. Typically, the public never sees these documents. The purpose of this indicator is to focus attention on the overall quality of test development documentation as provided by the assessment vendor to ensure the design of high-quality, standards-aligned testing materials: item writer guidelines, rubric design, text selection processes, and technical processes for review, revision/correction, and ultimate approval. Further, this indicator evaluates whether the assessment design documentation attends to the standards' adoption of a three-part model for measuring text complexity, noting text complexity is not only related to quantitative measures, but also driven by factors related to the text structure, conventionality, layers of meaning, and knowledge demands of the text as well as considerations of how well the text and the reader will connect. Additionally, this indicator evaluates whether adequate processes are in place to ensure content accuracy, technical quality and editorial precision of the assessment items.

Resources:

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards
  - The Standards Approach to Text Complexity, pp. 4-10
- Supplemental Information for Appendix A of the Common Core State Standards for English Language Arts and Literacy: New Research on Text Complexity (pp. 4-10)
  - II. New Findings Regarding the Quantitative Dimension of Text Complexity
  - III. New Tools for Evaluating the Qualitative Dimension of Text Complexity
  - V. The Issue of Text Quality and Coherence in Text Selection
  - VI. Key Considerations in Implementing Text Complexity
  - VII. The Model in Action: Sample Annotated Reading Text
- CCSSO Criteria for High-Quality Assessments
  - B.1. Assessing student reading and writing achievement in both ELA and literacy
  - B.2. Focusing on complexity of texts

    - A.4. Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years.
    - A.6. Ensuring transparency of test design and expectations
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
    - Part 1: ELA and Literacy Curricula, Grades 3-5; ELA Curricula, Grades 6–12
        - I. Key Criteria for Text Selection (p.1-5)
        - II. Key Criteria for Questions and Tasks (p. 6-8)
    - Part 2: History/Social Studies, Science, and Technical Subjects Literacy Curricula, Grades 6–12
        - I.2.A. Curricula provide texts that are valuable sources of information. (p.15)
        - I.2.B. Curricula include opportunities to combine quantitative information derived from charts and other visual formats and media with information derived from text. (p.15-16)
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
    - Introduction, p. 1-3
    - Key Criteria for Reading Foundations, p. 3-5
    - Key Criteria for Text Selections, p. 5-7
    - Key Criteria for Questions and Tasks, p. 7-9
- *Standards for Educational and Psychological Testing*.
    - Chapter 4: "Test Design and Development" (pgs. 75-84)
        - Cluster 1: "Standards for Test Specifications" (pgs. 85-87)

## Indicator 1.1.a Guiding Questions:

- Do the assessment design specifications provide a clear rationale justifying or explaining the design of the assessment, the benefits of the assessment to learners or other stakeholders, and a foundation of research grounding the assessment or assessment process?
- Is item development documentation sufficiently robust to support the writing of items to measure CCR standards?
- Is the item development documentation sufficiently robust to support the review of items to measure CCR standards?
- As necessary, do assessment design specifications provide clear, standards-aligned scoring information and/or rubrics to evaluate students' levels of proficiency with respect to the targeted CCR standards?
- Does assessment development documentation detail a rigorous process for the review and selection of test passages aligned to the expectations of CCR standards?
- Does item development documentation include a description of processes used to ensure items are content-accurate and without technical or editorial flaws?

### Evidence Collection

Assessment Rationale

- Locate a statement of purpose for the assessment.
- Review explanations justifying the assessment and/or citations validating the assessment's use and benefits to students as well as stakeholders.

Item Development and Review Guidance

- Review guidelines and processes for the production of high-quality items consistently aligned to CCR standards.
- Review assessment's means for measuring cognitive complexity relating to item development.
- Review processes for ensuring a range of CCR standards is assessed.
- Identify level of specificity in item writer training materials.

- Review technical processes used for item review and approval.
- Review documentation regarding the field testing or piloting of proposed assessment items.

Scoring Information, Rubric Design

- Review scoring guides for level of detail in point values attributable to individual items as well as item groups related to targeted standards.
- Review guidelines for the development and design of rubrics and exemplars, noting relationship between expectations and targeted standards.
- Review constructed-response rubrics, exemplars, and annotated student samples.

Text Selection Guidelines

- Review complexity framework guide/reports/documentation for quantitative and qualitative measures associated with passage selection including means for measuring text complexity of multimedia stimuli.
- Review rationales for text/passage selection regarding:
  - balance of text types,
  - alignment to targeted CCR standards,
  - public domain, permissioned passages, and/or commissioned passages,
  - appropriateness of text topics for common student interests at the assessment grade level(s).
- If commissioned passages are included, note and review specificity of writing guidance and alignment to CCR standards.
- If published passages are included, note and review specificity of editing guidance and alignment to CCR standards.

Content and Technical Accuracy

- Note processes to ensure content accuracy of the items.
- Note processes to ensure technical quality and editorial accuracy of the items.
- If Criterion 2.3 is evaluated, note documentation and processes ensuring technical quality of the reported sub-score(s).

## Cluster Meeting Discussion

Assessment Rationale

- Does the assessment rationale explain how students and stakeholders will benefit from having used the assessment?
- Does the assessment rationale provide a foundation of research that grounds the assessment?
- Does the rationale or justification for using the assessment include a history of the assessment's value over time?
- Does the rationale ground the use of the assessment in a larger context, e.g., across schools, districts, states, or nationally?

Item Development and Review Documentation

- What guidance and support does item development documentation provide for the creation of an item bank or item pool that ensures alignment to the full range and intent of targeted college- and career-ready standards?
- How robust are item writing materials in supporting the development of high-quality selected response item stems that require relevant text evidence in the response as prescribed by targeted CCR standards?
  - Are well-grounded rationales for designing text-dependent items and requiring relevant text evidence, explicit and/or implicit, in response to the stimulus included in the documentation?
  - Are clear guidelines for the construction of two-part evidence-based selected responses included?

    - Are sample items incorporating CCR standards provided to illustrate the design of evidence-based selected responses, two-part evidence-based selected response or other innovative/effective designs?
    - Does the documentation address the concept of cognitive complexity and describe the process used to measure cognitive complexity at an item level?
    - Are item writer checklists included?
- How robust are item writing materials in supporting the development of technology-enhanced stems that require relevant text evidence in the response as prescribed by targeted CCR standards?
    - Are guidelines for the construction of technology-enhanced evidence-based selected response and evidence-based constructed response prompts clear and supported with examples that incorporate CCR standards?
    - Are item writer checklists included?
- How robust are item writing materials in supporting the development of constructed-response prompts that require relevant text evidence in the response as prescribed by targeted CCR standards?
    - Are clear guidelines for the development of evidence-based constructed response prompts included?
    - Are sample materials incorporating CCR standards to illustrate the design of evidence-based constructed response prompts included (e.g., prompts, exemplar drafts, and annotated student samples)?
    - Are item writer checklists included?
- How thorough is the review process in ensuring high-quality test items on test events?
- What materials are weak or missing, if any?

Scoring Information, Rubric Design

- How do scoring matrices and/or guides note the point value attributed to items, the proportional distribution of points for item groups, or items targeting individual or clustered standards?
- If constructed-response items are included in the assessments, are the scoring materials adequate to ensure consistency of scores regardless of who is doing the scoring?

Text Selection Guidelines

- What guidance does the assessment documentation provide regarding the distribution of literary and informational texts that is grounded in CCSS?
    - Grade 4: 50% literary; 50% informational
    - Grade 8: 45% literary; 55% informational
    - Grade 12: 30% literary; 70% informational
- What guidance does the assessment documentation provide regarding the range of distribution among text types (e.g., poetry, drama, narrative fiction, speeches, multimedia, literary nonfiction, historical, scientific, and/or technical texts)?
- What guidance does the assessment development documentation provide test designers in selecting engaging, well-crafted texts for the assessment including guidance in the selection of multimedia stimuli?
- What guidance does the assessment development documentation provide test designers in the selection of texts representing diverse cultures and/or periods and include texts of diverse media and formats?
- What information does the assessment documentation provide with regards to whether text is public domain, permissioned passages, and/or commissioned passages?
- What guidance does the assessment documentation give for determining the appropriateness of text topics for the given grade level(s)?
- Which text complexity measurement tools are identified for both conventional passages and multimedia stimuli?
    - Is at least one tool for measuring complexity a research-based instrument?
- How consistently are both quantitative and qualitative measures of text complexity applied in the selection of assessment passages?

    - Are reported quantitative metrics of assessment texts the product of a reliable computer program?
    - Does the qualitative tool consider text structure, language conventions, knowledge demands, and levels of meaning?
    - Which means for measuring text complexity of multimedia stimuli provided?
- Do selection guidelines document caveats for inclusion of off-grade level texts in grade-level assessment?
    - Are annotations or explanations of sample texts provided to indicate where texts may appear to be off-grade level and justify inclusion of such texts in the assessment process?

Content and Technical Accuracy

- What information does the documentation provide to support content accuracy in regard to item development and review?
- What information does the documentation provide to support technical quality, and editorial accuracy in development and review of the items?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.1 | Assessment design specifications align to the expectations of college- and career-ready (CCR) standards. |
| Indicator 1.1.b | Test blueprints and/or assessment design specifications reflect an appropriate distribution of content and related score points, item types, and cognitive demand within test events.<br>● The expected distribution of test content within and across ELA domains is defined within the test blueprint and/or assessment design specifications and reflects the emphasis established in CCR standards.<br>● The expected distribution of score points among ELA domains and subdomains is defined within the test blueprint and/or assessment design specifications and reflects the emphasis suggested in CCR standards.<br>● The expected type and range of item types are reflected in the test blueprint and/or assessment design specifications and are appropriate to address the expectations of CCR standards.<br>● The suggested ranges of cognitive demand are reflected in the test blueprint and/or assessment design specifications and are sufficient to measure the depth of CCR standards. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● The distribution of ELA content outlined within test blueprints and/or assessment design specifications strongly reflects the emphasis established by targeted CCR standards.<br>● The test blueprints and/or assessment design specifications include sufficient guidance to ensure the type and range of items are appropriate to address the expectations of CCR standards.<br>● For assessments evaluating overall ELA proficiencies<br>○ The assessment attributes 15-20% of vocabulary scoring points to the overall literacy assessment score. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● The distribution of ELA content outlined within test blueprints and/or assessment design specifications reflects the emphasis established by targeted CCR standards.<br>● The test blueprints and/or assessment design specifications partially provide sufficient guidance to ensure test items are appropriate in range and/or type to address the expectations of CCR standards.<br>● For assessments evaluating overall ELA proficiencies<br>○ The assessment attributes less than 15% of vocabulary scoring points to the overall literacy assessment score. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● The distribution of ELA content outlined within test blueprints and/or assessment design specifications does not reflect the emphasis established by targeted CCR standards.<br>● The test blueprints and/or assessment design specifications do not provide sufficient guidance to ensure test items are appropriate in range and/or type to address the expectations of CCR standards.<br>● The expected distribution of score points among ELA domains and subdomains as defined within the test blueprints and/or assessment design specifications do not |

| | | |
|---|---|---|
| ○ The assessment attributes 15-20% of the overall score to points earned in targeted CCR aligned writing tasks.<br>○ The assessment attributes 8-10% of standard English conventions scoring points to the overall literacy assessment score.<br>● The test blueprints and/or assessment design specifications address a means by which to measure cognitive demand, and further, the suggested ranges of cognitive demand are reflected in the test blueprint and sufficient to measure the depth of CCR standards. | ○ The assessment attributes less than 15% of the overall score to points earned in targeted CCR writing tasks.<br>○ The assessment attributes less than 8% of standard English conventions scoring points to the overall literacy assessment score.<br>● The test blueprints and/or assessment design specifications address a means by which to measure cognitive demand; however, the application is inconsistent with theory or practice and therefore not sufficient to measure the cognitive complexity of the standards. | reflect the emphasis suggested by CCR standards.<br>● Test blueprints and/or assessment design specifications do not provide a matrix, taxonomy, or classification for measuring cognitive demand within test events. |

## About this indicator:

What is the purpose of this Indicator?

Assessment design specifications focus the goals of the assessment description or framework on precise standards or learning goals and specific cognitive levels. To ensure consistency between and among test events, blueprints further specify the number of items and item types required for valid and reliable measurement. The purpose of this indicator is to evaluate whether the test blueprint and/or assessment design specifications align to CCR standards, including the distribution of test items representing the targeted standards and their associated score points. Additionally, this indicator evaluates whether the blueprint and/or assessment design specifications provides for a range of items across a clearly stated measure of cognitive demand to ensure the depth and the breadth of CCR standards are met.

Resources:

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects
- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects: Appendix A Research Supporting Key Elements of the Standards
  - Writing, p. 23-25
  - Vocabulary, p. 32
  - Three Tiers of Words, p. 33-35
  - Foundational Skills, p. 17-22
- CCSSO Criteria for High-Quality Assessments
  - B.1. Assessing student reading and writing achievement in both ELA and literacy
  - B.3. Requiring students to read closely and use evidence from texts
  - B.4. Requiring a range of cognitive demand
  - B.5. Assessing writing
  - B.6. Emphasizing language skills
  - B.7. Assessing research and inquiry

- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - Introduction
  - Part 1. Section II. Key Criteria for Questions and Tasks
  - Part 1. Section III. Key Criteria for Academic Vocabulary
  - Part 2. Section II. Key Criteria for Questions and Tasks
  - Part 2. Section III. Academic (and Domain Specific) Vocabulary
  - Part 1. Section IV. Key Criteria for Writing to Sources and Research
  - Part 1. Section V.4. Materials embrace the most significant grammar and language conventions.
  - Part 2. Section IV. Key Criteria for Writing to Sources and Research
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
  - Key Criteria for Reading Foundations, p. 3-5
  - Key Criteria for Text Selections, p. 5-7
  - Key Criteria for Questions and Tasks, p. 7-9
- *Standards for Educational and Psychological Testing*.
  - Chapter 4: "Test Design and Development" (pgs. 75-84)
    - Cluster 2: "Standards for Item Development and Review" (pgs. 87-90)
    - Cluster 3: "Standards for Developing Test Administration and Scoring Procedures and Materials" (90-93)
  - Chapter 6, "Test Administration, Scoring, Reporting, and Interpretation" (pgs. 111-113)
    - Cluster 2: "Test Scoring" (pg. 118)

## Indicator 1.1.b Guiding Questions:

- Is the expected distribution of test content within and across ELA domains defined within the test blueprint and/or assessment design specifications?
- Does the expected distribution of test content reflect the emphasis established in CCR standards?
- Is the expected distribution of score points among ELA domains defined within the test blueprint and/or assessment design specifications?
- Is the expected distribution of score points among ELA subdomains defined within the test blueprint and/or assessment design specifications?
  - If the assessment measures vocabulary as a subdomain of an overall literacy score, does the assessment place sufficient emphasis on vocabulary with a significant percentage of the score points devoted to vocabulary skills?
  - If the assessment measures language as a subdomain of an overall literacy score, is sufficient emphasis placed on language skills with a significant percentage of the score points devoted to these skills?
    - Are the majority of score points for items meeting the language standards found in written responses or editing items that reflect authentic writing applications?
  - If the assessment measures writing as a subdomain of an overall literacy score, is there sufficient emphasis on writing and a significant percentage of score points targeting writing standards?
- Does the expected distribution of score points reflect the emphasis suggested in CCR standards?
- Are the expected type and range of item types reflected in the test blueprint and/or assessment design specifications?
- Are the type and range of items appropriate to address the expectations of targeted CCR standards?
- Are suggested ranges of cognitive demand reflected in the test blueprint and or assessment design specifications?
- Are the suggested ranges of cognitive demand sufficient to measure the depth of CCR standards?

## Evidence Collection

Test Content

- Review the blueprint and/or assessment design specifications defining the ELA content to be assessed.

Scoring information

- Review the blueprint and/or assessment design specifications for distribution of score points within and between content areas, domains and subdomains; consider the correlation of scoring to content emphasis suggested by CCR standards.
- Review scoring information for all items (e.g., MC, MS, TE, and CR).
- Review scoring rationales for point distribution among item types (e.g., MC, MS, TE, and CR.

Type and Range of Items

- Review the blueprint and/or assessment design specifications to identify item types and the range of item types represented in the assessment.
- Review the range of item types to determine correlation between item type and content to be assessed (i.e., method-measure match).
    - Examine assessment to determine if students are actually writing to earn scoring points in the writing standards or if there may be selected response items attributed to writing standards.

Cognitive Demand

- Review the protocol for applying the taxonomy or classification system for measuring cognitive demand to the assessment; confirm a consistent means of measuring cognitive demand is applied across all domains and grade levels.

## Cluster Meeting Discussion

Test Content

- Does the blueprint and/or assessment design specifications make clear the ELA content to be assessed in each test event?
- How well is the content outlined within the blueprint and/or assessment design specifications aligned to CCR standards?
    - Does the blueprint and/or assessment design specifications emphasize the importance of vocabulary by clearly targeting a range of vocabulary standards (e.g., CCSS RI.4, RL.4, L.4, L.5, L.6)?
    - Does the blueprint and/or assessment design specifications provide a range of items designed to measure the application of standard English conventions (e.g., capitalization, punctuation, and spelling)?
    - If writing is to be measured, does the assessment ensure test takers are composing written responses in alignment with CCR standards?
        - What methods of measurement are used to assess writing, e.g., short answer, extended response, performance tasks?

Scoring

- What is the distribution of points among the various item types?
- Are all items weighted the same? If not, how are point differences explained? Do point differences make sense?
- As applicable, do the blueprint and/or assessment design specifications ensure the item scoring is representative of the emphasis placed on content by CCR standards?
    - clearly indicate writing scores as a significant input in determining overall literacy scores?

- clearly indicate the score points targeting vocabulary and vocabulary standards (e.g, CCSS RI.4, RL.4, L.4, L.5, L.6)?
- clearly indicate the number of items and/or related score points targeting standards related to language knowledge and conventions (e.g., CCSS L.1, L.2, and L.3)?
  - clearly indicate the majority of score points for items meeting the language standards are found in written responses or editing items that reflect authentic writing applications?
  - provide well-grounded rationales for evaluating common English language usage and convention errors in writing samples and/or performance tasks?
  - establish clear expectations for evaluating common English language usage and convention errors in forced response items?
- What is the distribution of points derived from writing tasks (if applicable) in relation to other literacy domains (e.g., reading, vocabulary, language skills)?

Type and Range of Items

- Does the blueprint and/or assessment design specifications make clear the distribution of item types to be represented on each event?
- How well does the distribution of item types as indicated in the blueprint and/or assessment design specifications match CCR standards/targets to be assessed?

Cognitive Demand

- Does the documentation provide a clear rationale for the selection of a cognitive complexity matrix?
- How/where does the documentation describe an appropriate and consistent means to apply the cognitive complexity matrix across/among test events?
- How well are the levels of cognitive demand ascribed to items within the assessment and/or the assessment overall aligned to the expectations of targeted CCR standards?

# Gateway 1: Alignment, Fairness, & Accessibility

## Criterion 1.2

### Item and Form Alignment

Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards.

**Note: These indicators may be "N/C" if they are not claimed by the publisher to be present in the assessment.*

## What is the purpose of this Criterion?

The development of any quality assessment requires not only a comprehensive grasp of the content to be assessed but also a strategic plan for fair and equitable assessment design and delivery. The second criterion focuses on the assessment's alignment to the academic expectations of the targeted college- and career-ready standards, within items and among forms[1].

## Research Connection

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects
- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards
- Supplemental Information for Appendix A of the Common Core State Standards for English Language Arts and Literacy: New Research on Text Complexity
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (Eds.). (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.
- Shinn, E., & Ofiesh, N. S. (2012). Cognitive diversity and the design of classroom tests for all learners. *Journal of Postsecondary Education and Disability*, 25(3), 232-255.

## Scoring:

| | | |
|---|---|---|
| Meets Expectations | Partially Meets Expectations | Does Not Meet Expectations |

[1] Note: test forms and test events are synonymous in this document. EdReports will use these terms to describe a set of test items administered to a student in a single sitting either through a fixed form or computer adaptive test to measure an indicated construct.

| | | |
|---|---|---|
| • >79% of applicable points | • 50%-79% of applicable points | • <50% of applicable points |

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.a | The text passages are of high quality and aligned with the expectations of CCR standards and aligned to assessment design specifications.<br>• All texts or other stimuli included in the assessment are of publishable quality and are reflective of the requirements of CCR standards.<br>• Text complexity is determined and documented using at least one research-based instrument; overall text complexity is determined through a combination of qualitative and quantitative analyses.<br>• Texts are placed at the grade level indicated by the results of the text complexity process, with exceptions supported by explanation.<br>• The distribution of texts and/or other stimuli are representative of the balance and range of text types required by CCR standards.<br>• Texts provide adequate context for the design of multiple, meaningful items and are sufficiently developed to support logical inferences.<br>• Selected texts align to the assessment design specifications. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• All texts and/or other stimuli in the assessment are of publishable quality and reflective of the requirements of CCR standards.<br>• Text complexity is determined and documented using more than one research-based instrument; overall text complexity is measured through a combination of qualitative and quantitative analyses.<br>• Nearly all texts are placed at grade level/band correlated to the text complexity process; exceptions are supported by explanation.<br>• The distribution of texts and/or other stimuli are representative | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• Most texts and/or other stimuli in the assessment are of publishable quality and reflective of the requirements of CCR standards.<br>• Text complexity is determined and documented using at least one research-based instrument; overall text complexity is measured through a combination of qualitative and quantitative analyses.<br>• Most texts are placed at a grade level/band correlated to the text complexity process; exceptions may or may not be supported by explanation.<br>• The distribution of texts and/or other stimuli are most often | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• Few to no texts and/or other stimuli in the assessment are of publishable quality and reflective of the requirements of CCR standards.<br>• Text complexity may or may not be determined and documented using at least one research-based instrument; overall text complexity is not measured through a combination of qualitative and quantitative analyses.<br>• Some texts are placed at a grade level/band correlated to the text complexity process; exceptions are not supported by explanation.<br>• The distribution of texts and/or other stimuli are not |

| | | |
|---|---|---|
| of the balance and range of text types required by the CCSS.<br>● Texts provide adequate context for the design of multiple, meaningful items and are sufficiently developed to support logical inferences.<br>● The majority of assessment texts align to the assessment design specifications. | representative of the balance and range of text types required by the CCSS.<br>● Most texts provide adequate context for the design of multiple, meaningful items and are sufficiently developed to support logical inferences.<br>● Some of the assessment texts align to the assessment design specifications. | representative of the balance and range of text types required by the CCSS.<br>● Most texts do not provide adequate context for the design of multiple, meaningful items and are not sufficiently developed to support logical inferences.<br>● The assessment texts are not aligned to the assessment design specifications. |

## About this indicator:

### What is the purpose of this Indicator?

The Revised Publishers' Criteria (2012) repeatedly advises educators to ensure "the quality of the suggested texts is high — they are worth reading closely and exhibit exceptional craft and thought or provide useful information." The purpose of this indicator is to evaluate whether the stimulus texts are of a caliber worthy of students' time: rich enough in message to engage the test-takers, sufficiently crafted to act as exemplars of their respective test events, and complex enough to be reflective of CCR standards and targets. Noting text complexity is not only related to qualitative measures, but also driven by factors related to the text structure, conventionality, layers of meaning, and knowledge demands, this indicator evaluates whether the assessment places texts at appropriate grade levels in relation to text complexity. Additionally, this indicator evaluates whether the stimulus texts are sufficiently rich enough to provide meaningful context for item writers. Finally, this indicator evaluates the degree to which the selected texts align not only to CCR standards, but also to the assessment design specifications.

### Resources

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards
  - The Standards Approach to Text Complexity, pp. 4-10
- Supplemental Information for Appendix A of the Common Core State Standards for English Language Arts and Literacy: New Research on Text Complexity (pp. 4-10)
  - II. New Findings Regarding the Quantitative Dimension of Text Complexity
  - III. New Tools for Evaluating the Qualitative Dimension of Text Complexity
  - V. The Issue of Text Quality and Coherence in Text Selection
  - VI. Key Considerations in Implementing Text Complexity
  - VII. The Model in Action: Sample Annotated Reading Text
- CCSSO Criteria for High-Quality Assessments
  - B.1. Assessing student reading and writing achievement in both ELA and literacy
  - B.2. Focusing on complexity of texts
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - I. Key Criteria for Text Selection (p.1-5)
  - II. Key Criteria for Questions and Tasks (p. 6-8)
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2

    - Key Criteria for Text Selections, p. 5-7
    - Key Criteria for Questions and Tasks, p. 7-9
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
    - I.2.A. Curricula provide texts that are valuable sources of information.
    - I.2.B. Curricula include opportunities to combine quantitative information derived from charts and other visual formats and media with information derived from text.

## Indicator 1.2.a Guiding Questions:

- Are all texts or other stimuli included in the assessment of publishable quality and reflective of the requirements of CCR standards?
- Is text complexity analysis measured through at least one research-based instrument?
- In addition to quantitative analysis, is text complexity evaluated through a qualitative analysis of the four factors (i.e., levels of meaning, structure, language conventionality and clarity, and knowledge demands)?
- Is the placement of texts within a grade level or grade band indicated by the results of the text complexity process, quantitative and qualitative, with exceptions supported by explanation?
- Is the distribution of texts and/or other stimuli representative of the balance and range of text types required by CCR standards?
- Do texts provide adequate context for the design of multiple, meaningful items?
- Are texts sufficiently developed for the design of multiple, meaningful items?
- Are texts sufficiently developed to support logical inferences?
- Do the selected texts align to assessment design specifications?

## Evidence Collection

Text Quality

- Review listing of text titles presented for assessment.
- Note text details for publicly released item sets, public domain and/or permissioned passages.
- Note and review text details for commissioned passages.
- Review assessment texts and passages.
- Look for copyright permission related to texts and text excerpts.

Complexity Analysis

- Review complexity reports/documentation for quantitative and qualitative measures associated with specific text selections.
- Review media presented in the assessment (e.g., video, charts, graphs, and/or artwork).
    - Note evaluation documents and judgments regarding multimedia texts.

Off-Grade Texts

- Identify off-grade level/band texts in the assessment.
    - Note documented caveats for inclusion of off-grade level/band level/band passages, including annotations to justify the text's inclusion in the assessment.
    - Examine reading purpose associated with off-grade level/band texts.
- Cross check selected passages against passage selection documentation.

Text Distribution

- Review assessment texts, titles, and passages.
    - Identify text types and distribution among test events.

Text Meaningfulness

- Read the texts for depth and richness.
- Consider whether the texts, informational or literary, are age-appropriately provocative to intellectual thought.
- Review the texts for the level of inferences a reader may make, the knowledge a reader may acquire, and the connections the text weaves into the construction of the work as a whole.

Alignment to Development Documentation
- Note alignment between text selection guidelines in assessment design specifications and actual text passages (e.g., distributions of literary and informational texts; cultural diversity, content diversity, media diversity; complexity metrics).

## Cluster Meeting Discussion

Text Quality
- Which of the texts and/or other stimuli are currently published or on public display?
- If the texts have been previously published or available on public display, how does the testing instrument provide the background to the reader?
- Which assessment texts not currently published or on public display (e.g., commissioned texts) are of a caliber worthy of the time spent to read and consider?
- Do the texts presented create intellectual opportunities for deeper thought and/or careful consideration?
- Do the texts presented offer new ways of thinking? Add to existing knowledge? Provoke alternative views?
- How well does actual text quality reflect the assessment design specifications?

Complexity Analysis
- How does the assessment report the complexity levels, quantitative and qualitative, for all texts and/or passages?
- Which measure(s) of text complexity are research-based instruments?
- Which means for measuring text complexity of multimedia stimuli are reported?
- How do the means for measuring and reporting actual text complexity reflect the assessment design specifications?

Off-Grade Texts
- Are off-grade level/band texts included in the assessment?
- How well does the explanation for inclusion of off-grade level/band texts align to the explanation provided in Supplemental Information for Appendix A?
- How do the tasks and purposes associated with off-grade level/band texts affect the complexity of the text?
- How well are the off-grade level/band texts matched with the item task/s?
- How well do off-grade level/band texts accurately reflect the assessment design specifications regarding these selections?

Text Distribution
- Is the distribution between literary and informational texts in alignment with CCSS guidelines?
  - Grade 4: 50% literary; 50% informational
  - Grade 8: 45% literary; 55% informational
  - Grade 12: 30% literary; 70% informational?
- Do the assessment texts and/or other stimuli correspond to targeted CCR standards requirements?
- Do the assessment texts and/or other stimuli correspond to CCR standards' requirements related to a range of text types (e.g., poetry, drama, narrative fiction, speeches, multimedia, literary nonfiction, historical, scientific, and/or technical texts)?

- Among all test events reviewed, what is the distribution and/or range of texts from cultures and/or periods and what kinds of texts of diverse media and formats are included?
- How well does the distribution of texts in all aforementioned areas accurately reflect the assessment design specifications?

Text Meaningfulness

- Which of the assessment texts are substantial enough in content, design, and/or information to support a series of coherent reading and/or writing items aligned to the associated grade/band level?
- Which of the assessment texts are rich enough to support the construction of coherent items that engage readers more deeply in the text?
- Which of the assessment texts are rich enough to support the construction of coherent items that enable readers to generate and/or support inferential conclusions?
- Which of the assessment texts are rich enough to support the construction of a series of text-related selected response items that progress and may culminate in a constructed response and/or a performance task?
- How well does the meaningfulness of the assessment texts accurately represent the assessment design specifications?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.b | Test items are written to elicit evidence of learning relative to one or more CCR standard/s and aligned to assessment design specifications.<br>● Test items can be clearly identified as measuring one or more CCR standard/s without formally measuring knowledge and skills that are not included within CCR standards.<br>● Test items align to assessment design specifications.<br>● Items are content-accurate and reflect no technical or editorial flaws. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● At least 80% of the test items and test item sets can be aligned to exclusively measure CCR standards.<br>● Most assessment items align to the assessment design specifications.<br>● At least 97% of the test items contain no content flaws and are technically and editorially accurate. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● Between 50-79% of the test items and test item sets can be aligned to exclusively measure CCR standards.<br>● There is partial alignment between the actual assessment items and the assessment design specifications.<br>● At least 95% of the test items contain no content flaws, are technically and editorially accurate. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● Fewer than 50% of the test items and test item sets can be aligned to exclusively measure CCR standards.<br>● There is little to no alignment between the actual assessment items and the assessment design specifications.<br>● Numerous inaccuracies impede the demonstration of knowledge and skills. |

## About this indicator:

What is the purpose of this Indicator?

College- and career-ready standards are premised on reading, writing, speaking and listening as expanding cognitive processes built on increasingly complex texts, expanding vocabularies, analytic thinking, and evidence-based arguments and supports. The purpose of this indicator is to evaluate the degree of alignment between test items, CCR standards, and the assessment design specifications. Also evaluated are content and editorial accuracies and technical quality.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - B.1. Assessing student reading and writing achievement in both ELA and literacy

- B.3. Requiring students to read closely and use evidence from the texts
- B.5. Assessing writing
- B.6. Emphasizing vocabulary and language skills
- B.7. Assessing research and inquiry
- B.8. Assessing speaking and listening

## Indicator 1.2.b Guiding Questions:

- Can test items be clearly identified as measuring one or more CCR standards?
- Do test items avoid measuring knowledge and skills that are not included within CCR standards beyond their identified standard alignment?
- Do test items align to the assessment design specifications?
- Are items content-accurate and without technical or editorial flaws?

### Evidence Collection

Alignment to CCR Standards

- Examine items and descriptive metadata (if provided) for alignment to specific CCR standards.
- Identify items contained within an item set designed to measure the complexity of a specific CCR standard (e.g., two-part evidence-based responses).
- Look for items intended to elicit responses not reflective of CCR standards (e.g., items targeting personal connections, items not reflective of grade level or grade band standards, items measuring content knowledge outside of the scope--implied or stated--of CCR standards).

Alignment to Assessment Design Specifications

- Note alignment between assessment design specifications and the actual assessment items (e.g., content distribution [proposed and actual]; type and range of items [proposed and actual]).

Accuracy

- Look for inaccuracies in item presentation, including precision in the application of literary terms, reading comprehension relationships, and grammatical and/or usage conventions.
- Look for technical flaws (e.g., formatting, applications of technology enhanced items, integration of multimedia items, etc).
- Look for editorial inaccuracies (e.g., spelling, grammar, and punctuation).

### Cluster Meeting Discussion

Alignment to CCR Standards

- Does the pool of items reviewed provide enough information (e.g., metadata) to ensure alignment to one or more CCR standards?
- Considering the item pool, what percentage of the items can be clearly identified as measuring one or more of the tested domain's standards?
- How well does the descriptive metadata (if provided) associated with items and/or item sets (e.g., EBSR), correlate to the expectations of associated CCR standards?
- What proportion of item stems limit their questioning to CCR standards-related skills and concepts?
- What proportion of assessment keys and rubrics indicate item responses are limited to CCR standards related skills and concepts?
- What items or item sets (e.g., EBSR) can be identified as measuring the totality of a CCR standard?

Alignment to Assessment Design Specifications

- How well does the pool of items reviewed align to assessment design specifications?

- In distribution of content?
- In range and type of items?

Accuracy

- In relation to content, what percentage of items are accurate with no flaws?
- In relation to technical accuracy, what percentage of items are accurate with no flaws?
- In relation to editorial accuracy, what percentage of items are accurate with no flaws?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.c*<br>**This indicator may be considered "N/C" if it is not claimed by the publisher to be present in the assessment.* | The range of item types and cognitive demand among test events is sufficient to strategically assess the depth and complexity of CCR standards being addressed and is aligned to blueprints or assessment design specifications.<br>• The item stimuli are constructed to reach the depth and complexity of CCR standards expressing multiple cognitive goals (e.g., determine and analyze; describe and explain; make connections among and distinctions between).<br>• There is an appropriate distribution and/or range of cognitive demand exercised among test events submitted for review.<br>• The range of item types and cognitive demand among test events align to blueprints or assessment design specifications.<br>• If the assessment claims to measure writing, there is at least one extended constructed-response item to fulfill the expectations of CCR writing standards. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• Considering all test events reviewed, the range of item types fully matches targeted CCR standard/s and strategically assesses at least 80% of the targeted standard/s to their full depth and complexity.<br>• There is evidence of consistent processes to verify an appropriate range of cognitive demand among test events and/or standards being assessed.<br>• For the most part, the range of item types and cognitive demand among test events reflects the design proposed in blueprints or assessment design specifications. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• Considering all test events reviewed, the range of item types fully matches targeted CCR standard/s and strategically assesses between 79-50% of the targeted standard/s to their full depth and complexity.<br>• There is partial evidence of consistent processes to verify an appropriate range of cognitive demand among test events and/or standards being assessed.<br>• The range of item types and cognitive demand among test events partially reflects the design proposed in blueprints or assessment design specifications. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• Considering all test events reviewed, the range of item types does not fully match targeted CCR standard/s and strategically assesses fewer than 50% of the targeted standard/s to their full depth and complexity.<br>• There may or may not be evidence of consistent processes to verify an appropriate range of cognitive demand among test events and/or standards being assessed.<br>• The range of item types and cognitive demand among test events rarely reflects the design proposed in blueprints or assessment design |

| | | |
|---|---|---|
| • *If the assessment purpose states writing as a target of the assessment measure*, there is at least one extended constructed response represented in each test event. | • *If the assessment purpose states writing as a target of the assessment measure*, there is at least one extended constructed response represented in each test event. | specifications.<br>• *If the assessment purpose states writing as a target of the assessment measure*, there *may or may not be* at least one extended constructed response represented in each test event. |

## About this indicator:

What is the purpose of this Indicator?

College- and career-ready standards encompass a range of depth and complexity; aligned assessments must cover this depth and range. The purpose of this indicator is to evaluate whether test events fully align to targeted CCR standards. Full alignment requires that items among test events *fully* measure the targeted standards to ensure an accurate measure of proficiency level at the appropriate level of cognitive demand. For example, assessment items targeting CCR standards requiring explanation and/or evidence must provide a means for test-takers to offer explanations and/or evidence. Assessment items targeting CCR standards requiring comparison and contrast as aspects of the standard must provide a means for test-takers to do both. Items need to assess both literal and inferential reading skills. Items targeting Common Core Writing Standards 1-3 must provide a means for test-takers to actually write to appropriate prompts. Items targeting Common Core Writing Standard 8 must provide multiple sources from which test takers can draw information. Additionally, this indicator measures to what degree the test events represent the assessment design specifications or blueprints.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - B.3. Requiring students to read closely and use evidence from the texts
  - B.4. Requiring a range of cognitive demand
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - Introduction
  - Part 1. Section II. Key Criteria for Questions and Tasks
  - Part 2. Section II. Key Criteria for Questions and Tasks
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
  - Key Criteria for Questions and Tasks, p. 7-9

## Indicator 1.2.c Guiding Questions:

- Are item stimuli constructed to reach the depth and complexity of CCR standards expressing multiple cognitive goals (e.g., determine and analyze; describe and explain; make connections among and distinctions between)?
- Is there an appropriate distribution and/or range of cognitive demand exercised among test events?
- To what degree does the range of item types and cognitive demand among test events align to blueprints or assessment design specifications?
- If the assessment claims to measure writing, is there at least one extended constructed-response item to fulfill the expectations of CCR writing standards?

## Evidence Collection

Depth and Complexity of the Standards

- Record test experience related to item types and item sets (e.g., EBSR), item demands, cognitive demand, and technology interface.
- Review metadata and record data regarding the degree to which a standard is fully measured in depth and complexity.
- Review processes and/or assessment design specifications for ensuring a range of standards across test events is assessed.

Distribution of Cognitive Demand

- Look for various item types and item sets (e.g., EBSR) on test events. These items may include but are not limited to the following:
  - Multiple choice, multiple select, two-part evidence-based selected response, technology enhanced, technology enhanced constructed response, constructed extended response, short answer, performance tasks, other innovative item types.
- Note references to cognitive complexity associated with items and/or item sets (e.g., EBSR).
- Identify items and/or item sets (e.g., EBSR) that require test-takers to explain responses and/or provide evidence.
- Ensure that when technology enhanced items are used, they are relevant with value-added to the item.

Range of Items Matches Documentation

- Compare development rationale or purpose to actual test events.
- Note if ranges of cognitive demand described in the assessment design specifications are actualized among test events.
- Note alignment between assessment design specifications or blueprints and the actual test events (e.g., relationship of content distribution [proposed and actual]; distribution of score points [proposed and actual]; type and range of items [proposed and actual]).

Writing

- Identify items/prompts asking students to construct arguments, inform/explain, or write narrative responses.

## Cluster Meeting Discussion

Depth and Complexity of the Standards

- Considering all test events reviewed, what percentage of the  standards are measured to their full depth and complexity?
- Are the ranges of cognitive demand sufficient to measure the depth of targeted CCR standards?

Distribution of Cognitive Demand

- What variety of item types are represented in each test event?
- How well does the variety of item types meet the demands of the targeted standards?
- Do items targeting complex CCR standards provide sufficient context and options for response to meet the standards' expectations?
- If the assessment measures sub-scores, are the items tagged for measuring subskills or growth representative of content described in the assessment purpose?

Range of Items Matches Documentation

- What processes are documented to provide verification of the levels of cognitive demand assigned to items?

- To what degree are the ranges of cognitive demand represented on the test events aligned to the suggested ranges of cognitive demand provided on the test blueprint or assessment design specifications?
- How do the assessment design specifications or blueprints align with the actual test events? (e.g., relationship of content distribution [proposed and actual]; distribution of score points [proposed and actual]; type and range of items [proposed and actual]).

Writing

- If the assessment claims to measure writing standards, is there at least one extended constructed response on the assessment?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.d*<br>**This indicator may be considered "N/C" if it is not claimed by the publisher to be present in the assessment.* | The assessment is aligned to the reading expectations of CCR standards.<br>• Reading items require students to use or provide textual evidence in support of responses requiring close reading and analysis (e.g., constructed-response and/or two-part evidence-based selected-response item formats).<br>• Vocabulary items reflect the range of requirements for college and career readiness, including a focus on academic or Tier 2 words, the use of context to determine meaning, and an emphasis on words and phrases important to the central ideas of the text. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• The assessment requires close reading and analysis, consistently providing reading items that elicit the use of relevant textual evidence in response to the stimulus.<br>• Item responses consistently require the reader to use and explain both implicit and explicit text evidence and/or citations.<br>• The assessment comprehensively addresses CCR vocabulary and/or vocabulary standards (e.g., CCSS RI.4, RL.4, L.4, L.5, L.6) focusing on academic vocabulary and/or Tier 2 words.<br>• The assessment requires test-takers use context to determine meaning.<br>• The assessment focuses on words and phrases important to the central idea of the text. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• The assessment requires occasional close reading and analysis, sometimes providing items that elicit the use of relevant textual evidence in response to the stimulus.<br>• Item responses sometimes require the reader to use and or explain both implicit and explicit text evidence and/or citations.<br>• The assessment targets CCR vocabulary and/or vocabulary standards (e.g., CCSS RI.4, RL.4, L.4, L.5, L.6) and may or may not include items targeting academic vocabulary and/or Tier 2 words.<br>• The assessment tends to focus on vocabulary items with little or no context provided.<br>• The assessment partially focuses on words and phrases important to the central idea of the text. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• The assessment rarely or never requires close reading and analysis.<br>• Rarely or never do items elicit the use of relevant textual evidence in response to the stimulus.<br>• Rarely or never do item responses require the reader to use or explain implicit and/or explicit text evidence and/or citations.<br>• The assessment does not consistently or purposefully target vocabulary and/or CCR vocabulary standards.<br>• The assessment does not focus on words and phrases important to the central idea of the text. |

## About this indicator:

What is the purpose of this Indicator?

The intent of this indicator is to evaluate whether the assessment meaningfully connects to one of the key shifts of the CCSS: students should be able to answer a range of questions with answers grounded in evidence from texts, both literary and informational, in reading, writing, and speaking. Additionally, this indicator evaluates whether the assessment meaningfully connects to another key shift of the CCSS: regular practice with complex texts and their academic language. The indicator evaluates the assessment's adherence to the capacities of college and career-ready students to "set and adjust…language use as warranted…appreciate nuances…how the connotations of words affect meaning" and use "general and specialized reference materials when appropriate." Furthermore, the indicator evaluates whether the assessment meaningfully supports vocabulary skills as a measurement of literacy by emphasizing vocabulary both in the number of items assessing vocabulary standards and the scoring value attributed to those items.

Resources:

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects: Appendix A Research Supporting Key Elements of the Standards
  - Vocabulary, p. 32
  - Three Tiers of Words, p. 33-35
- CCSSO Criteria for High-Quality Assessments
  - B.3. Requiring students to read closely and use evidence from the texts
  - B.6. Emphasizing vocabulary and language skills
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - Introduction
  - Part 1. Section II. Key Criteria for Questions and Tasks
  - Part 1. Section III. Key Criteria for Academic Vocabulary
  - Part 2. Section II. Key Criteria for Questions and Tasks
  - Part 2. Section III. Academic (and Domain Specific) Vocabulary
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
  - Key Criteria for Reading Foundations, p. 3-5
  - Key Criteria for Text Selections, p. 5-7
  - Key Criteria for Questions and Tasks, p. 7-9

## Indicator 1.2.d Guiding Questions:

- Do reading items require students to use or provide textual evidence in support of responses requiring close reading and analysis, e.g., constructed-response and/or two-part evidence-based selected-response item formats?
- Does the assessment emphasize the importance of vocabulary by clearly targeting a range of CCR vocabulary standards (e.g., CCSS RI.4, RL.4, L.4, L.5, L.6)?
- Do vocabulary items reflect the range of requirements for college and career readiness, including a focus on academic or Tier 2 words, the use of context to determine meaning, and an emphasis on words and phrases important to the central ideas of the text?

### Evidence Collection

Close Reading/Text Evidence

- Look for reading items that indicate test-takers will be using and/or providing text evidence in support of their response. These items may include but are not limited to the following:
  - Multiple choice, multiple select, two-part evidence-based selected response answers, technology-enhanced, technology-enhanced constructed response, constructed extended response, and other innovative item types
- Review items to determine the consistency of alignment between CCR standards' expectation for textual dependency and items' actual demands for textual evidence in analysis and/or support of item response.
- Determine the distribution of textual demand: implicit or inferential use and explicit or citational use.
- Consider the grade appropriateness in relation to the distribution of items requiring implicit use of text and explicit reference to text.
- Reference the stimulus texts and review the question-text relationship.

Vocabulary Emphasis

- Look for reading items that indicate test-takers will be sufficiently assessed in the area of vocabulary. These items may include but are not limited to the following:
  - Multiple choice, multiple select, two-part evidence-based selected response answers, technology enhanced, technology enhanced constructed response, and other innovative item types.
- Review scoring tables for vocabulary items.

Academic or Tier 2 Vocabulary

- Look for vocabulary items clearly tied to contextual analysis. These items may include but are not limited to the following:
  - Multiple choice, multiple select, two-part evidence-based selected response answers, technology enhanced, technology enhanced constructed response, and other innovative item types.

## Cluster Meeting Discussion

Close Reading/Text Evidence

- Overall, how well does the assessment align to the Common Core's shift to close reading that requires text evidence as part of a response?
- What items require close reading (i.e., require readers to provide and/or use text evidence to form the response)?
- Which types of items are being used to demonstrate close reading and textual analysis (e.g., multiple selected response, technology enhanced, and/or constructed response)?
- Are the item types appropriate to the targeted CCR standard(s)?
- How well do items requiring close reading and analysis among test events develop a CCR pattern of CCR expectations?

Vocabulary Emphasis

- Does the assessment emphasize the importance of vocabulary by clearly targeting a range of vocabulary standards (e.g., CCSS RI.4, RL.4, L.4, L.5, L.6)?
- How do reviewed items target a range of vocabulary skills sufficient to emphasize the importance of vocabulary to the measurement of literacy skills?
- How does the assessment indicate the score points targeting vocabulary and vocabulary standards (e.g., CCSS RI.4, RL.4, L.4, L.5, L.6)?
- In general, does the assessment attribute a significant portion of vocabulary relative to the overall score?

Academic or Tier 2 Vocabulary

- Does the assessment focus vocabulary items on grade-level academic vocabulary and/or Tier 2 words?

- How often do vocabulary items use contextualized stems to engage students in using context in meaning making of academic and/or Tier 2 words?
- Does the assessment require test-takers to use passage context to determine meaning of unknown grade-level words and/or multiple-meaning words?
- Does the assessment focus vocabulary items on grade-level words and phrases important to the central idea of the text?
- How consistently are the vocabulary items aligned to standards across all assessments reviewed (e.g., CCSS RL.4, RI.4, L.4, L.5, L.6)?
- How well do vocabulary items engage students in making grade appropriate meaning of academic, technical, connotative and figurative language?
- How often do vocabulary items assess the test-takers' understanding of word parts and the use of general and specialized reference materials related to vocabulary acquisition?
- Are there uncontextualized vocabulary items? If so, how well do they assess the test-takers' understanding of word parts and the use of general and specialized reference materials related to vocabulary acquisition?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.e*<br>**This indicator may be considered "N/C" if it is not claimed by the publisher to be present in the assessment.* | The assessment is aligned to the writing, research, and language expectations of CCR standards.<br>• Writing prompts and tasks represent the distribution of content reflected in CCR standards.<br>• Writing tasks and/or prompts are text-based and require students to analyze, synthesize, organize, and write using evidence from a source or sources.<br>• The majority of score points for items meeting the language standards are found in written responses or editing items that reflect authentic writing applications.<br>• Items assessing conventions focus on common student errors that address conventions most critical for college and career readiness in real-world settings. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• The assessment targets CCR standards' distribution of text types and purposes (i.e., opinion/argument, informative/explanatory, and narrative).<br>• The assessment writing task/s consistently require students read a source or multiple source materials and use information from that reading to build and present knowledge in their writing.<br>• The assessment writing task(s) consistently encourage students to analyze and synthesize text evidence as they organize thoughts for composing a written response. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• The assessment includes writing tasks but writing items do not target CCR standards' distribution of text types and purposes (i.e., opinion/argument, informative/explanatory, and narrative).<br>• The assessment writing task(s) may or may not require test-takers to read a source or multiple source materials and use information from that reading to build and present knowledge in their writing.<br>• The assessment writing task(s) may or may not encourage students to analyze and synthesize text evidence as they organize thoughts for composing a written response. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• The assessment does not present writing prompts and/or tasks that ask students to respond to reading through writing.<br>• The assessment writing task(s) does not require test-takers to read a source or multiple source materials and use information from that reading to build and present knowledge in their writing.<br>• The assessment writing task(s) does not encourage students to analyze and synthesize text evidence as they organize thoughts for composing a written response.<br>• The assessment writing task/s does not remind students to cite evidence in their response. |

| | | |
|---|---|---|
| • The assessment writing task/s consistently remind students to cite evidence in their response.<br>• The majority of score points for items meeting the language standards are found in written responses or editing items that reflect authentic writing applications.<br>• Assessment items focus on common grade level conventions of English grammar and usage aligned to CCSS L.1, L.2, and/or L.3.<br>• When language standards are measured within authentic writing applications, assessment prompts make test-takers aware that conventions will be evaluated in writing samples and/or performance tasks. | • The assessment writing task/s may or may not remind students to cite evidence in their response.<br>• The majority of score points for items meeting the language standards are found in items other than those that reflect authentic writing applications.<br>• Assessment items may or may not focus on common grade level conventions of English grammar and usage aligned to CCSS L.1, L.2, and/or L.3.<br>• When language standards are measured within authentic writing applications, assessment prompts make or may not make test-takers aware that conventions will be evaluated in writing samples and/or performance tasks. | • The majority of score points for items meeting the language standards are found in written responses or editing items that do not reflect authentic writing applications.<br>• Assessment items do not focus on common grade level conventions of English grammar and usage aligned to CCSS L.1, L.2, and/or L.3.<br>• When language standards are measured within authentic writing applications, assessment prompts do not make test-takers aware that conventions will be evaluated in writing samples and/or performance tasks. |

## About this indicator:

What is the purpose of this Indicator?

Historically, much student writing was drawn from individual experience and opinion. Among the shifts of the Common Core State Standards (CCSS) is the expectation of evidence-based writing grounded in analysis of literary and informational texts as a means to inform and support arguments. The purpose of this indicator is to evaluate whether the assessment meaningfully supports the importance of writing as a measurement of literacy through an emphasis on writing tasks. Furthermore, this indicator evaluates whether the assessment targets the range of writing types and expectations described in the CCSS writing standards and further, adheres to the standards' assertion that students need "to use writing as a way of offering and supporting opinions, demonstrating understanding of the subjects they are studying, and conveying real and imagined experiences and events." Additionally, this indicator evaluates whether the assessment meaningfully targets a key shift of the CCSS, that reading, writing, and speaking are grounded in evidence from texts, both literary and informational.

Relatedly, this indicator evaluates whether the assessment meaningfully supports the CCSS' admonishment that "students must gain control over many conventions of standard English grammar, usage, and mechanics." Linked to assessment of language skills is an evaluation of the extent to which the assessment adheres to the standards' reminder that the rules of English are "inseparable" from their application in the context of literacy, in this case, specifically the context of writing.

A final purpose of this indicator is to evaluate whether the assessment meaningfully connects to the CCSS' emphasis on the use of appropriate conventions of standard English rather than measure obscurely known and infrequently-followed conventions of formal, standard English and/or mere knowledge of such conventions.

Resources:

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects: L1, L2, L3
  - Note on range and content of student language use, p. 25 and p. 51.
- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards
  - Writing, p. 23-25
- CCSSO Criteria for High-Quality Assessments
  - B.1. Assessing student writing achievement in both ELA and literacy
  - B.5. Assessing writing
  - B.6. Emphasizing language skills
  - B.7. Assessing research and inquiry
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - Introduction
  - Part 1. Section IV. Key Criteria for Writing to Sources and Research
  - Part 1. Section V. Additional Key Criteria for Student Reading, Writing, Listening, and Speaking
    - Materials embrace the most significant grammar and language conventions (p. 12)
  - Part 2. Section IV. Key Criteria for Writing to Sources and Research

## Indicator 1.2.e Guiding Question:

- Do writing prompts and tasks represent the distribution reflected in CCSS standards?
- Are writing tasks and/or prompts text based?
- Do writing tasks and/or prompts require students to analyze, synthesize, organize, and write using evidence from a source or sources?
- Are the majority of score points for items meeting CCR language standards found in written responses or editing items that reflect authentic writing applications?
- Are items assessing conventions focused on common student errors that address conventions most critical for college and career readiness in real-world settings?

### Evidence Collection

- Examine test events and/or related documentation for evidence of writing task distribution (i.e., argument, inform/explain, and narrative).
- Examine test items (prompt and corresponding text/s) to evaluate the demands of organization, analysis and synthesis required in the written response.
- Review assessment scoring guides, rubrics, and annotations; evaluate the assessment's expectations for organization, analysis, and synthesis in the written response.
- Review documentation for delineations among writing tasks (e.g., short answer, extended response, research tasks).
- Review writing items/prompts for references to CCSS language standards.
- Review documentation emphasizing the evaluation of more commonly found errors in English usage.

### Cluster Meeting Discussion

- Do the writing tasks prompted in the assessment target the standards' distribution of text types and purposes as described in CCSS, (i.e., opinion/argument, informative/explanatory, and narratives)?

- Do the assessment writing items and/or prompts target the CCSS' expectations that students respond to reading through writing, using evidence gained in their reading to support their writing?
  - encourage students to analyze and synthesize text evidence as they organize thoughts for composing a written response?
  - encourage students to cite or reference text evidence in their written response?
- Do the assessment writing items and/or prompts target the CCSS' expectations that students read a source or multiple source materials and write to build and present knowledge?
- How well do the items demonstrate consistency in the structure and design of writing items among various test events?
  - In earlier grades, do the actual writing items/prompts provide clear guidance or cues regarding the prompt purpose and/or the expected type of writing as a response?
  - In later grades, do the actual writing items/prompts allow students "to combine elements of different kinds of writing—for example, to use narrative strategies within argument and explanation within narrative—to produce complex and nuanced writing" (CCSS, p. 41)?
- How often are the writing items and/or prompts appropriately drafted to stimulate rich thought and not limit the test taker's response?
- How often are multiple texts and research techniques incorporated into assessment items (e.g., does the testing environment allow or provide for note taking while reading)?
- Where does the documentation include guidelines to ensure consistent alignment to CCSS grade-level standards related to standard English conventions (e.g., Language Standards 1-3)?
- How often are the score points for items meeting the language standards found in written responses or editing items that reflect authentic writing applications?
- How often do multiple choice items reflecting Common Core grade-level standards L1, L2, and L3 consistently offer plausible distractors among the selection items?
- Among all test events reviewed, how often does the assessment
  - measure a CCR progression of English language conventions for grammar and usage rather than obscure and/or advanced conventions?
  - measure a CCR progression of English language conventions for capitalization and punctuation rather than focus on obscure and/or advanced conventions?
  - measure a CCR progression of English language conventions for spelling by rewarding rather than penalizing errors in the use of mature but incorrectly spelled words?
  - make test-takers generally aware that conventions will be evaluated in writing samples and/or performance tasks?
  - ensure writers are not penalized for errors that do not distract from the overall meaning and/or integrity of their compositions?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.f*<br>**This indicator may be considered "N/C" if it is not claimed by the publisher to be present in the assessment.* | The assessment is aligned to the speaking and listening expectations of CCR standards.<br>• Speaking items assess students' ability to draw on diverse content to prepare for, participate in, or orally present findings in a performance task.<br>• Listening comprehension items assess students' ability to evaluate text and marshal information presented in diverse media forms. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• The assessment comprehensively targets CCR speaking and listening standards.<br>• The assessment requires the evaluation of text and the marshaling of information from text in the measurement of speaking and listening skills.<br>• The assessment requires the use of text evidence in the measurement of speaking and listening skills.<br>• The assessment provides a variety of media for measuring listening comprehension (e.g., radio, original recording, contemporary audio recording). | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• The assessment targets CCR speaking and listening standards.<br>• The assessment emphasizes the evaluation of text and the marshaling of information in the measurement of speaking and listening skills.<br>• The assessment requires little in the way of the use of text evidence in the measurement of speaking and listening skills.<br>• The assessment evaluates listening comprehension through a single medium in one form of oral text. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• The assessment, while claiming alignment to CCR standards, does not target speaking or listening expectations of CCR standards.<br>• The assessment does not require the evaluation of text and the marshaling of information from text in the measurement of speaking and listening skills.<br>• The assessment does not require the use of text evidence in the measurement of speaking and listening skills.<br>• The assessment tends to measure hearing ability rather than listening skills. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is twofold. In the evaluation of both speaking and listening, the indicator evaluates whether the assessment meaningfully connects with CCR speaking and listening standards by measuring those skills in academic practice (i.e., the contributions of pertinent disciplinary findings through conversations, presentation, and/or debates) appropriately supported by media as needed to support, clarify, inform and/or

persuade. Secondly, in the realm of listening, the indicator evaluates whether the assessment meaningfully connects with the CCR speaking and listening standards by measuring listening comprehension using college and career aligned materials.

Resources:

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards (p. 27)
- CCSSO Criteria for High-Quality Assessments
  - B.8. Assessing speaking and listening
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - II.1.c Questions and tasks require the use of textual evidence, including supporting valid inferences from the text.

## Indicator 1.2.f Guiding Questions:

- Do speaking items assess students' ability to draw on diverse content to prepare for, participate in, or orally present findings in a performance task?
- Do listening comprehension items assess students' ability to evaluate text and marshal information presented in diverse media forms?

### Evidence Collection

- Review the assessment to ascertain the range of speaking tasks presented.
- Review the speaking items to evaluate whether the items appropriately target grade-level standards.
- Review the assessment to ascertain the range of listening tasks presented.
- Review the listening comprehension items to evaluate whether the items appropriately target grade-level standards.

### Cluster Meeting Discussion

Speaking

- How well does the assessment target grade level speaking standards?
- As standard appropriate, how does the assessment
  - emphasize the use or marshaling of information in the measurement of speaking skills?
  - emphasize the use/evaluation of text in the measurement of speaking skills?
  - require the use of text evidence in the measurement of speaking skills?

Listening

- Which listening comprehension standards does the assessment target (e.g., CCSS SL.2 and SL.3)?
- Which forms for measuring listening comprehension does the assessment provide (e.g., original orations, single readers, dramatic dialogues)?
- How often does the assessment emphasize the marshaling of information in the measurement of listening skills?
- How often does the assessment emphasize the evaluation of text in the measurement of listening skills?
- How often does the assessment require the use of text evidence in the measurement of listening skills?
- How often does the assessment evaluate the use of active listening skills?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.2 | Text passages, assessment items, and resulting test forms align to expectations of ELA domains as outlined by college- and career-ready (CCR) standards. |
| Indicator 1.2.g*<br>**This indicator may be considered "N/C" if it is not claimed by the publisher to be present in the assessment.* | The assessment is aligned to the foundational skills expectations of reading CCR standards.<br>● Prompts and tasks represent the distribution of content reflected in CCR standards.<br>● The assessment provides a variety of item elicitation formats appropriate for measuring CCR foundational skills. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● The assessment comprehensively targets the foundational skills expectations of reading CCR standards.<br>● The assessment provides a variety of item elicitation formats appropriate for measuring CCR foundational skills. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● The assessment measures some foundational skills expectations of reading CCR standards.<br>● The assessment provides item elicitation formats appropriate for measuring foundational skills standards. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● The assessment, while claiming alignment to CCR standards, does not target grade level foundational skills expectations of reading CCR standards.<br>● The assessment does not provide item elicitation formats appropriate for measuring foundational skills standards. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate whether the assessment is aligned to the foundational skills expectations of CCR standards and meaningfully supports the standards' expectations targeting grade-level foundational skills in the areas of print concepts, phonological awareness, phonics, word recognition, and/or fluency. Furthermore, the indicator evaluates whether the item formats are appropriate for eliciting foundational skills knowledge from students.

Resources:

- Common Core State Standards for English Language Arts, Kindergarten-Grade 12: Reading: Foundational Skills
- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards (p. 17-22)
- CCSSO Criteria for High-Quality Assessments
  - B.6. Emphasizing Vocabulary and Language Skills

- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
  - Introduction: p. 1-3
  - Key Criteria for Reading Foundations, p. 3-5
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
  - I.1.b. All students (including those who are behind) have extensive opportunities to encounter grade-level complex text.

## Indicator 1.2.g Guiding Questions:

- Do prompts and tasks represent the distribution of content reflected in foundational skills reading CCR standards?
- Does the assessment provide a variety of item elicitation formats appropriate for measuring CCR foundational skills?

### Evidence Collection

- Review the assessment to ascertain the range of foundational skills assessed: print concepts, phonological awareness, phonics, word recognition, and/or fluency.
- Review the items to evaluate whether the items appropriately assess targeted CCR foundational skills standards.

### Cluster Meeting Discussion

- Which foundational skills standards does the assessment target?
- Through which item elicitation formats does the assessment evaluate foundational skills?

# Gateway 1: Alignment, Fairness, & Accessibility

## Criterion 1.3

### Fairness and Accessibility

The assessment is fair and accessible for all students in the intended test-taking population.

### What is the purpose of this Criterion?

The development of any quality assessment requires not only a comprehensive grasp of the content to be assessed but also a strategic plan for fair and equitable assessment design and delivery. The third criterion focuses on the interim assessment's adherence to universal design principles and the incorporation of design elements that allow for the widest range of test takers.

### Research Connection

- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects
- Common Core State Standards for English Language Arts & Literacy in History/Social Studies, Science, and Technical Subjects Appendix A: Research Supporting Key Elements of the Standards
- Supplemental Information for Appendix A of the Common Core State Standards for English Language Arts and Literacy: New Research on Text Complexity
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades K–2
- Revised Publishers' Criteria for the Common Core State Standards in English Language Arts and Literacy, Grades 3–12
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (Eds.). (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.
- Shinn, E., & Ofiesh, N. S. (2012). Cognitive diversity and the design of classroom tests for all learners. *Journal of Postsecondary Education and Disability*, 25(3), 232-255.

### Scoring:

| Meets Expectations | Partially Meets Expectations | Does Not Meet Expectations |
|---|---|---|
| • 10-12 points | • 6-9 points | • <6 points |

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.3 | The assessment is fair and accessible for all students in the intended test taking population. |
| Indicator 1.3.a | Items and test events are developed and reviewed using procedures that ensure fairness.<br>• Item development documentation/procedures clearly demonstrate adherence to the principles of universal design.<br>• Item rendering specifications clearly reflect the principles of universal design.<br>• Item review processes are designed to minimize construct-irrelevant variance.<br>• Items and passages go through a content bias/sensitivity review to make sure they are appropriate and fair for all relevant student groups.<br>• Procedures are in place to evaluate the technical quality and appropriateness of items and test events for student subgroups and students utilizing different accommodations. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• Item development documentation/procedures clearly demonstrate adherence to principles of universal design.<br>• Test and item rendering specifications clearly reflect the principles of universal design.<br>• Item review processes are effectively designed to mitigate construct-irrelevant variance.<br>• Items go through a consistent content bias/sensitivity review to make sure they are appropriate and fair for all relevant student groups<br>• Passages go through a consistent content bias/sensitivity review to make sure they are appropriate for all relevant student groups. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• Item development documentation/procedures demonstrates adherence to the principles of universal design.<br>• Test and item rendering specifications reflect the principles of universal design.<br>• Item review processes attempt to mitigate construct-irrelevant variance.<br>• Items go through a content bias/sensitivity review to make sure they are appropriate and fair for all relevant student groups<br>• Passages go through a content bias/sensitivity review to make sure they are appropriate for all relevant student groups.<br>• Procedures are in place to evaluate the technical quality | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• Item development guidelines make no reference to universal design or other means to ensure fairness in the test design.<br>• Test and item rendering specifications reflect little or no concern for principles of fairness.<br>• Item review processes give little to no attention to concerns for construct-irrelevant variance.<br>• There may or may not be a process for content bias/sensitivity review; on its face, the assessment shows little or no concern for appropriateness and/or fairness.<br>• Passages do not go through a content bias/sensitivity review to make sure they are |

| | | |
|---|---|---|
| • Consistent procedures are in place to evaluate the technical quality and appropriateness of items and test events for student subgroups and students using varied accommodations. | and appropriateness of items and test events for student subgroups and students using varied accommodations. | appropriate for all relevant student groups.<br>• Procedures are not in place to evaluate the technical quality or the appropriateness of items and test events for student subgroups and students using varied accommodations. |

## About this indicator:

What is the purpose of this Indicator?

At all levels, assessments must be held accountable for fairness. To ensure students' scores are not influenced by disability, ethnicity, culture, geographic location, socioeconomic condition, or gender, assessments in both development and final design should conform to the principles of universal design. By following these principles of fairness and incorporating elements of universal design, the widest possible range of students will be offered opportunities for full participation in the test experience. The purpose of this indicator is to determine to what degree the assessment adheres to such principles.

Resources:

- Cognitive diversity and the design of classroom tests for all learners.
- *Universal Design for Learning: Theory and Practice*.
- *Standards for Educational and Psychological Testing*.
  - Chapter 3: "Fairness in Testing" (p. 49-62)
    - Cluster 1: Test Design, Development, Administration, and Scoring Procedures (p. 63-65)
  - Chapter 12: "Educational Testing and Assessment: Accommodations and Modifications" (p. 183-194)
- CCSSO Criteria for High-Quality Assessments
  - A.2 Ensuring that assessments are valid for required and intended purposes.
    - Evidence that the assessments lead to the intended consequences
    - The set of content standards against which the assessments are designed is provided.
    - Evidence is provided to ensure the content validity of test forms and the usefulness of score reports.
  - A.5 Providing accessibility to *all* students, including English learners and students with disabilities
    - Follow the principles of universal design
    - Offer appropriate accommodations and modifications
    - Assessments provide valid and reliable scores for English learners
    - Assessments provide valid and reliable scores for students with disabilities

## Indicator 1.3.a Guiding Questions:

- Does test development documentation clearly demonstrate adherence to principles of universal design?
- Do test development procedures clearly demonstrate adherence to principles of universal design?
- Do item rendering specifications clearly reflect the principles of universal design?
- Does the item review process mitigate threats of construct-irrelevant variance?
- Do items go through a content bias/sensitivity review to make sure they are appropriate for all relevant student groups?

- Do items go through a content bias/sensitivity review to make sure they are fair for all relevant student groups?
- Do passages go through a content bias/sensitivity review to make sure they are fair for all relevant student groups?
- Are procedures in place to evaluate the technical quality of test items and test events for student subgroups and students using varied accommodations?
- Are procedures in place to evaluate the appropriateness of test items and test events for student subgroups and students using varied accommodations?
- Are there procedures in place or validity evidence to support that reported information (e.g., achievement scores, predictive information, etc.) does not differ in meaning for relevant subgroups in the examinee population?
- Are there procedures in place or validity evidence to support that reported information (e.g., achievement scores, predictive information, etc.) does not differ in meaning for students using supported accommodations?

## Evidence Collection

Adherence to Principles of Universal Design

- Review item development and item writing guidance, and training materials specifically for adherence to the core principles of universal design.
- Review item development and training materials, specifically for efforts to minimize instances of construct-irrelevant variance.

Content Bias/Sensitivity Review

- Review processes and guidelines for mitigating bias and sensitivity conflicts within assessment items and passages and among test events.

Technical Quality

- Review documentation summarizing scaling and equating procedures and how they are monitored and evaluated over time (e.g., scale drift).
- Review item and test development specifications, to better understand the level of detail provided to support consistency in the development of items and test events.
- Review procedures and policies used to evaluate newly developed items for potential bias or Differential Item Functioning (DIF) prior to operational use.
- Review procedures used to evaluate the reliability of assessment results across disaggregated student groups and for students utilizing different accommodations.

## Cluster Meeting Discussion

Adherence to Principles of Universal Design

- Do item development guides emphasize adherence to principles of universal design?
- How are item development and review processes designed to mitigate instances of construct-irrelevant variance due to factors such as disability, ethnicity, culture, geographic location, socioeconomic condition, or gender?
- How are item development and review processes designed to mitigate instances of construct-irrelevant variance due to factors such as clutter on the page, graphics, reading load, flawed items, etc.?
- How do test development specifications ensure that assessments are clear and comprehensible for all students?

- How do test rendering specifications and/or explanations indicate fair and accessible assessment design and delivery practices?

Content Bias/Sensitivity Review

- Is the makeup of the bias/sensitivity committee(s) representative of the intended student population? Are all major student groups represented?
- Are guidelines for avoiding bias and sensitivity conflicts adequate to ensure items, passages, and test events are free of questionable or offensive language or attitudes?

Technical Quality

- Are flagging rules and/or evaluation criteria for items demonstrating differential performance described or provided in conjunction with a defensible rationale?
- If differential item functioning or differential test functioning is detected, are prescriptive actions detailed to review, revise, and/or drop items from the item pool (or an operational test event)? Are those actions justified and justifiable?
- When credible evidence indicates that test scores may differ in meaning for relevant subgroups, is validity evidence supporting the score interpretations for individuals from those subgroups provided?
- If predictive information is reported, does predictive validity evidence indicate that the prediction does not over- or under-predict future performance for particular sub-groups of students?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.3 | The assessment is fair and accessible for all students in the intended test taking population. |
| Indicator 1.3.b | Appropriate accommodations and supports are in place to ensure the assessment is accessible to all students in the intended test taking population, including special populations of students and English Learners.<br>• The test-taking population for which the assessment was/was not designed to support is clearly documented.<br>• The list of accommodations is aligned to the vendor's definition of the assessment's intended uses.<br>• The list of accommodations is sufficient to serve the needs of the full population of intended test takers.<br>• Evidence is available to support the validity and fairness of the intended interpretations and uses for those students who access the exam using the supported accommodations.<br>• Evidence is available that supports the quality and appropriateness of provided accommodations.<br>• The administration manual is clearly worded and supports teachers and other educational personnel in providing an appropriate testing experience for all students.<br>• Sample forms or released test items are available to stakeholders at each grade level. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• The test-taking population for which the assessment was/was not designed to support is clearly documented.<br>• The provided accommodations are sufficient to support the intended use of results.<br>• The list of accommodations provided is sufficient for the intended population of test-taking students, including special populations of students and English Learners.<br>• Evidence is available to support the validity and fairness of the intended interpretations and | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• The test-taking population for which the assessment was/was not designed to support is documented.<br>• The provided accommodations support the intended use of results.<br>• The list of assessment accommodations meets the needs of some students targeted by the assessment.<br>• Evidence partially supports the validity and fairness of the intended interpretations and uses for those who access the exam using the supported | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• The test-taking population for which the assessment was designed is not clearly provided.<br>• The list of assessment accommodations meets few or none of the targeted students' needs.<br>• Few or none of the provided assessment accommodations align to the purposes or outcomes of the assessment.<br>• No evidence is provided to support the quality and appropriateness of provided accommodations. |

| | | |
|---|---|---|
| uses for those students who access the exam using the supported accommodations.<br>● Evidence is available to support the quality and appropriateness of provided accommodations.<br>● The administration manual is clearly worded and supports teachers and other educational personnel in providing an appropriate testing experience for all students.<br>● Sample forms or released test items are available to stakeholders at each grade level. | accommodations.<br>● Some of the provided assessment accommodations align to the assessment's purposes and uses.<br>● Some evidence is provided to support the quality and appropriateness of provided accommodations.<br>● The administration manual supports teachers and other educational personnel in providing a testing experience for all students.<br>● Sample forms or released test items may or may not be available to stakeholders. | ● The administration manual is poorly written and fails to support teachers and other educational personnel in providing a testing experience for all students.<br>● Sample forms or released test items are not available to stakeholders. |

## About this indicator:

What is the purpose of this Indicator?

Assessment systems should accommodate all test-takers in the intended testing population, including English Learners and special populations of students. The purpose of this indicator is to evaluate accessibility (i.e., supports provided to educators to establish an appropriate testing environment for all students). In addition to evaluating accessibility, the indicator also evaluates the appropriateness and the quality of the provided accommodations as well as the means by which the accommodation is provided.

Resources:

- Cognitive diversity and the design of classroom tests for all learners.
- *Standards for Educational and Psychological Testing*.
  - Chapter 3: "Fairness in Testing" (pp. 49-62)
  - Chapter 7: "Supporting Documentation for Tests" (pp.123-129)
  - Chapter 12: "Educational Testing and Assessment: Accommodations and Modifications" (pp. 183-194)
- CCSSO Criteria for High-Quality Assessments
  - A.5 Providing accessibility to *all* students, including English learners and students with disabilities
    - Follow the principles of universal design
    - Offer appropriate accommodations and modifications
    - Provide valid and reliable scores for English learners

## Indicator 1.3.b Guiding Questions:

- Is the test-taking population for which the assessment was/was not designed to support clearly documented?
- Is the list of accommodations aligned to the vendor's definition of the assessment's intended uses?
- Is the list of accommodations provided sufficient given what the vendor has defined as the population of students for whom the assessment was designed, including special populations of students and English Learners?

- Is evidence available in the accessibility and accommodations manual or guidelines to support the integrity of the intended score interpretations for all test-takers?
- Is evidence available to support the quality of provided accommodations for specific users or groups of users?
- Is evidence available to support the appropriateness of provided accommodations for specific users or groups of users?
- Is the administration manual clearly worded to support teachers and other educational personnel in providing an appropriate testing experience for all students?
- Are sample forms or released test items available to stakeholders at each grade level?

## Evidence Collection

Testing Population
- Review documentation overview of intended test taking population.

Sufficiency of Accommodations for Demonstration of Knowledge/Skill
- Review list of assessment accommodations offered to ensure the assessment accessibility to all students in that population.
- Review accessibility and accommodations manuals, instructions, guidance.
- Review test development and item writing guidelines to ensure test and/or accessibility features do not hinder access to item content.

Quality and Appropriateness of Accommodations
- Review assessment data, research reports, or other documentation addressing validity and reliability questions associated with supported accommodations.
- Ensure testing guidelines address assessment presentation, response, setting, timing and scheduling.
- Determine test accessibility of online glossaries and/or translation processes.
- Read the test administration manual for adherence to best practices.

Sample Forms
- Review sample forms or released items.

## Cluster Meeting Discussion

Testing Population
- Is the test-taking population for which the assessment was/was not designed clearly documented?

Sufficiency of Accommodations for Demonstration of Knowledge/Skill
- Is there evidence that test items and accessibility features permit English Learners to demonstrate their knowledge and abilities?
- Is there evidence that test items do not contain features that unnecessarily prevent test-takers from accessing the content of the item?
- Are allowed accommodations appropriate for removing construct-irrelevant barriers and enabling test takers to demonstrate their knowledge and skills?

Alignment of Accommodations for Comparative Purposes
- Does documentation show the accommodations provided on the interim assessment are comparable with accommodations provided for other relevant predicted criterion measures?
- Is the list of accommodations aligned to the vendor's definition of the assessment's intended uses?
- Is the list of accommodations aligned to the vendor's definition of the assessment's intended test-takers?

Quality and Appropriateness of Accommodations

- Are research-based studies or empirical evidence provided to show the effectiveness of suggested accommodations?
- Are the accommodations likely to remove construct-irrelevant barriers without interfering with the measurement of the intended construct? For example, read-aloud accommodation may not be appropriate when assessing reading comprehension.
- Are literature reviews provided to document appropriate use of the accommodations included?
- Are studies included to show differential boost for special populations of students when the appropriate accommodation is provided?
- Is evidence provided to show test accommodations address presentation, response requirements, timing/scheduling and setting?

Administration Manual

- How easily is the test administration manual accessed?
- How easy is the test administration manual to read and navigate?
- Does the test administration manual provide guidance in use and monitoring of accessibility features and allowed accommodations?
- Does the test administration manual provide guidance for administering the assessment in a proper testing environment for all students?

Sample Forms

- Do sample forms or released items provide a fair representation of the test experience?

# Gateway 1: Alignment, Fairness, & Accessibility

| | |
|---|---|
| Criterion 1.3 | The assessment is fair and accessible for all students in the intended test taking population. |
| Indicator 1.3.c | The range and types of technology provided within the assessment support the validity of assessment outcomes.<br>• Guidance is provided to support accessibility to the assessment system on a variety of platforms.<br>• Auditory supports present stimuli and items in a natural voice and at a cadence that can be adjusted to accommodate the learner.<br>• Overall visual design, including digital tools (e.g., dictionaries, thesauri, sticky notes, and highlighters) enhances the test-taking experience, does not distract or clutter the digital workspace, and can be easily navigated by students. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• Guidance is provided to support accessibility to the assessment system on a variety of platforms.<br>• Auditory supports present stimuli and items in a natural voice and at a cadence that can be adjusted to accommodate the learner.<br>• Overall visual design, including digital tools (e.g., dictionaries, thesauri, sticky notes, and highlighters) enhances the test-taking experience, does not distract or clutter the digital workspace, and can be easily navigated by students. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• Guidance is provided to support accessibility to the assessment system on more than one platform.<br>• Auditory supports are present.<br>• Digital tools (e.g., dictionaries, thesauri, sticky notes, and highlighters) are included, however, they may distract or clutter the digital workspace. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• Vague or overly technical guidance is provided to support accessibility to the assessment system.<br>• Auditory supports are not present to accommodate the learner.<br>• Overall visual design, including digital tools, distracts from the test-taking experience or clutters the digital workspace; digital tools challenge the student test-takers. |

## About this indicator:

What is the purpose of this Indicator?

Technology should enhance rather than constrain student performance, thereby influencing the validity of the test-taking experience and resulting outcomes. This indicator examines whether technology supports students in thoughtful engagement with the assessment content and avoids inadvertent construct-irrelevant variance.

Resources:

- Cognitive diversity and the design of classroom tests for all learners.
- CCSSO Criteria for High-Quality Assessments
  - A.5 Providing accessibility to all students, including English learners and students with disabilities
    - Following the principles of universal design
    - Offering appropriate accommodations and modifications
    - Assessments produce valid and reliable scores for students with disabilities

## Indicator 1.3.c Guiding Questions:

- Is guidance provided to support accessibility to the assessment system on a variety of platforms?
- Do auditory supports present stimuli and items in a natural voice and at a cadence that can be adjusted to accommodate the learner?
- Does the overall visual design, including digital tools (e.g., dictionaries, thesauri, sticky notes, and highlighters), enhance the test-taking experience?
- Does the overall visual design, including digital tools (e.g., dictionaries, thesauri, sticky notes, and highlighters), distract or clutter the digital workspace?
- Is the overall assessment design, including digital tools (e.g., dictionaries, thesaurus, sticky notes, and highlighters), easily navigated by students?

### Evidence Collection

- Review the technical assistance documents and/or services support documentation.
- Test the accessibility of digital materials (including all test events, teacher/administrator tools, and other potential school interfaces) to determine web-based compatibility with multiple Internet browsers (e.g., Firefox, Google Chrome).
- Test the assessment environment through actual application of tools.
- Review auditory controls: volume controls, voice modulation, speed, etc.
- Note the organization of space on the page, digital or print, including the use of graphics, borders, and text layout.
- In digital layouts, note means of transitions within and between passages, pages, etc.

### Cluster Meeting Discussion

Accessibility Guidance

- Are the assessments compatible with multiple Internet browsers (e.g., Firefox, Google Chrome)?
- Are assessments platform neutral (i.e., compatible with multiple operating systems such as Windows and Apple)?
- Do the assessments follow universal programming style?
- Do the assessments allow the use of tablets and mobile devices?

Auditory Supports

- Does the assessment provide auditory options that can be turned on and off based on the needs/requirements of a population?
- Are auditory supports delivered in a natural voice that can be adjusted to meet the needs of the student?

Digital Design and Supports

- Do digital tools provided for test-takers support the test environment?

- Does the overall visual design of the assessment enhance the test-taking experience?
- How easily can students navigate the testing environment?
- Are there distractions in the assessment layout?

# Gateway 2: Technical Quality

## Criterion 2.1

### Overall Achievement

The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain.

## What is the purpose of this Criterion?

The first criterion defines evidence necessary to support the interpretation of interim assessment test scores as measures of student achievement in the assessed content domains.

**A Note on Gateway 2 Reviews:** The Technical Quality criteria are evaluated by statisticians and psychometricians trained by the Center of Assessment. These criteria evaluate the validity, reliability, and the quality of scores created by the interim assessments to ensure the data is high quality. This review requires a deep understanding of the information and scores generated by the assessment and how the information addresses the purpose of the assessment.

## Potential Sources of Evidence for Criterion 2.1

- Technical Reports or Summaries
- Item development specifications and processes, and qualitative and quantitative item review and piloting procedures
- Test development and review procedures, including test blueprints and or adaptive specifications
- Standard setting procedures (if applicable)
- Procedures for establishing performance level descriptors (if applicable)
- Norming studies (if applicable), or summaries of any samples used to support reporting of national norms
- Summaries of validity analyses supporting the intended interpretations and uses of all the achievement scores.
- Procedures and results of any conducted reliability and precision analyses
- Equating and scaling procedures and scale score characteristics
- Test security and administration procedures

## Scoring:

| Meets Expectations | Partially Meets Expectations | Does Not Meet Expectations |
|---|---|---|
| • 7-8 points | • 5-6 points | • <5 points |

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.1 | The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain. |
| Indicator 2.1.a | Item and form development procedures result in high-quality test events.<br>• Item development, review, and piloting procedures and materials are designed to ensure all newly developed items meet technical quality standards.<br>• Assessment design specifications and test development and review procedures ensure test events meet content and statistical quality criteria. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |

## About this indicator:

What is the purpose of this Indicator?

A test score is only useful to the extent that it provides valid information about the degree to which a student understands the content standards targeted for assessment. While ensuring alignment between test blueprints, test items and the content standards is necessary, it is not sufficient. It is also necessary to show that test items and events were designed, developed and evaluated using high quality, technically sound procedures.

## Indicator 2.1.a

### What is reviewed?

- The assessment's technical report or related documentation that provides information about the design of the assessment, including the domain it is intended to assess and when/how frequently it is intended to be administered within an instructional sequence.
- Assessment design specifications – focusing on details reflecting the intended representation of test content, text complexity, item types, etc.

- Item development procedures and specifications, including content review and piloting procedures and outcomes
- Test development and review procedures (and/or specification underlying the development of adaptive test events)
- Information summarizing the required technical properties of test items and test events, including criteria underlying the selection/rejection of test items and properties of test events
- Item and test-level statistics associated with the test events provided for review
- Validity evidence demonstrating the relationship between test scores and other indicators of performance in the content domain
- Validity evidence demonstrating that assessment items elicit the knowledge and skills intended by the content standards

## Evidence Necessary to Meet Expectations for Indicator 2.1.a

Item development, review, and piloting procedures and materials were designed to ensure all newly developed items meet technical quality standards.

- ✔ All newly developed items are piloted with a sample of students representing the intended test taking population prior to operational use.
- ✔ There are clear, reasonable criteria in place for evaluating the quality of newly developed test items based on pilot test performance (e.g., fit and discrimination, constructed response performance distributions, procedures for flagging items that require additional content review or removal from the item bank).
- ✔ Documentation suggests that items are modified based on information provided by content reviewers or resulting from cognitive labs.
- ✔ Item-level and summary statistics adhere to the vendor's defined specifications for statistical quality.

Assessment design specifications and test development and review procedures ensure test events meet content specifications and statistical quality criteria.

- ✔ Procedures used to establish assessment design specifications are provided (e.g., specifically how decisions were made to ensure operational results would support desired claims about achievement in the content domain).
- ✔ Test specifications indicate the minimum expectations that must be met in order for any form (fixed or adaptive) to be considered compliant from a content and statistical standpoint.
- ✔ When applicable, adaptive test development specifications exist and clearly describe item selection procedures, and criteria for exiting/finishing a testing event to ensure the result is a valid measure of student knowledge in the content domain (e.g., requirements related to content coverage have been met).
- ✔ Evaluation procedures are in place to ensure that test forms (fixed or adaptive) meet the technical requirements defined within test development specifications (e.g., psychometric review of fixed forms; evaluation of simulated adaptive test events at different points along the ability distribution).
- ✔ Test events and associated test maps are provided for review and demonstrate the statistical characteristics defined within test specification documents.
- ✔ For adaptive assessments, summary data collected by the vendor demonstrates that assessments consistently meet the requirements of the test blueprints and specifications for students at all ability levels.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.1 | The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain. |
| Indicator 2.1.b | Achievement scores are reliable.<br>• Item/test development and review procedures facilitate the reliability of test scores.<br>• Procedures for calculating and evaluating reliability are well-documented and appropriate.<br>• Obtained reliability indices and estimates of precision are at an appropriate level to support the use of results as intended. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |

## About this indicator:

What is the purpose of this Indicator?

The focus of reliability analysis is to quantify the precision of test scores. While every score has some amount of error, it should not be so much that a test user lacks confidence that the score reflects what the student actually knows. The evidence listed below reflects that necessary to evaluate the adequacy of procedures used to control, evaluate and report the impact of measurement error on assessment results.

### Indicator 2.1.b

#### What is reviewed?

- Item and test development specifications to better understand the criteria and level of detail provided to support consistency in the development of items and test events (e.g., discrimination, fit, etc.).
- The assessment's technical report for information on the procedures used to calculate and evaluate test score reliability and, when appropriate, classification accuracy/decision consistency.

- Test administration specifications - for details designed to control the impact of extraneous factors on assessment results
- Reliability coefficients associated with provided test events and/or summary information describing the range of observed test score reliabilities across test events by grade and content area (for fixed or adaptive tests).
- If there are constructed response items, information about the training and procedures used to ensure reliability in the scoring of these items.
- Any studies evaluating test score reliability work).

## Evidence Necessary to Meet Expectations for Indicator 2.1.b

Item/test development and review procedures facilitate the reliability of test scores.

- ✔ Consistency in the scoring of constructed response items, when necessary, is evaluated prior to operational use (i.e., the extent to which scoring rubrics provide for consistent responses).
- ✔ For fixed forms - test score reliability is estimated and evaluated (against a defined criterion) prior to test administration.
- ✔ For adaptive, test cases are conducted to ensure the item bank supports the development of reliable assessments for students along the full range of the ability continuum.
- ✔ For adaptive tests that incorporate variable length stopping rules, CAT criteria specify the minimum standard error that must be achieved before a student is exited from a testing event.

Procedures for calculating and evaluating reliability are well documented and appropriate given the psychometric model.

- ✔ The type of reliability index that is reported makes sense given the psychometric model that is being used (e.g., classical or IRT or another model).
- ✔ When human judgment enters into scoring, procedures and methods for gathering and evaluating inter-rater, and within-examinee score reliability are provided.

Obtained reliability indices and estimates of precision are at an appropriate level to support the use of results as intended.

- ✔ While acceptable values for reliability are context dependent, a general rule of thumb is that the minimum score reliability for low-stakes use is generally around .70.
- ✔ There is reasonable measurement precision along the full range of the score continuum and/or near the cut-scores used to support decision making.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.1 | The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain. |
| Indicator 2.1.c | Achievement scores support intended interpretations of student performance.<br>• Evidence is provided to support the intended interpretations of student achievement.<br>• Equating/linking procedures supporting the comparability of achievement scores and score-based inferences across test events/administrations are described and reasonable.<br>• Item development specifications, task models, and scoring rubrics include enough detail to support consistency in the presentation, format, and degree of scaffolding observed in items and associated stimuli across test events.<br>• There is empirical evidence and an active research agenda supporting the validity of achievement scores as measures of the intended knowledge and skills. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |

## About this indicator:

What is the purpose of this Indicator?

Overall achievement scores can be reported in a variety of different ways. For example, scores may be reported using scaled scores, in performance categories (e.g., Below Basic, Basic, etc.), or using percentile ranks. Different representations answer different questions about a student's performance, such as: How did my student perform relative to other students in the state at his/her grade? Did my student perform at the level expected for students in his/her grade? In all cases, the procedures and data used to support intended interpretations of student achievement must be well documented and meet best practice.

## Indicator 2.1.c

### What is reviewed?

- *If assessment results are reported as scaled scores the evaluator should review:*
  - Procedures used to establish the scaled score metric and the characteristics of the scale (e.g., LOSS/HOSS).
- *If assessment results are reported in terms of performance categories or levels evaluators should review, as appropriate*
  - Descriptions of performance levels (e.g., performance level descriptors or achievement level descriptors) and the procedures by which they were established.
  - Procedures for establishing the cut-scores that define the different performance levels.
  - Evidence supporting the claims underlying performance in a particular category or beyond an established threshold (e.g., on-track, college and career ready, on-grade level).
- *If assessment results are interpreted in consideration of the performance of a norm group the evaluator should review:*
  - Procedures used to establish any national norms (e.g., norming study) and a detailed summary of the characteristics of the norm group.
  - Procedures used to calculate and define reported norms (e.g., stanine.) including any business rules detailing criteria for inclusion in the norm group (e.g., how many items must a student respond to be eligible for inclusion in local norms).
- Documentation summarizing scaling and equating procedures and how they are monitored and evaluated over time (e.g., scale drift).

### Evidence Necessary to Meet Expectations for Indicator 2.1.c

Evidence is provided to support the intended interpretations of student achievement.

- ✔ The manner in which overall achievement results (e.g., scaled scores, performance levels, etc.) are to be interpreted are clearly articulated.
- ✔ If assessment results are reported as scaled scores:
  - The procedures used to translate student performance to the scaled score metric are transparent and documented in enough detail to support consistent application across test events and administrations.
  - The properties of the reportable scale (e.g., range and spread) provide for an appropriate floor and ceiling given the range of achievement expected (over the first few years) and intended uses of assessment results
- ✔ If assessment results are reported in terms of performance categories or levels (e.g., Basic/Proficient/Advanced, pass/fail, or on-track/not on track):
  - The expectations associated with performance in a given level (including above/below a defined threshold) are clearly defined.
  - The process used to establish cut scores is well documented and utilizes data and procedures that support the intended interpretations (e.g., college ready; on grade level, etc.). *For example, the performance of students at Grade 5 may be used to establish the cut score defining on-track performance in Grade 4. Performance of a national sample of students could be used to define expectations for "on grade level." Review of the items associated with a given scaled score or range might be used to establish what it means to be "proficient" on a given assessment.*

- ✔ If *assessment results are interpreted in consideration of the performance of a norm group* (e.g., percentile rank, grade equivalent):
  - A clear description of the norm group is provided (e.g., if a norming study was conducted the procedures and samples used should be provided).
  - The norm group is relevant (given the manner in which results are intended to be used), representative of the examinee population of interest and large enough to be reliable.
  - The test design and scale specifications support valid and appropriate normative inferences (i.e., the assessment and scale will result in an appropriate range of student performance; the score scale is broad enough to spread students along the ability continuum).

Equating/linking procedures supporting the comparability of achievement scores and score-based inferences across test events/administrations are described and *reasonable*.

- ✔ Characteristics of linking sets are described, when appropriate.
- ✔ Procedures utilize appropriate data and statistics (e.g., appropriate sample sizes, stable item parameters).
- ✔ Procedures are in place to calculate and evaluate the standard error of equating.
- ✔ Procedures are in place to monitor equating stability over time and detect scale drift.

Item development specifications, task models, and scoring rubrics include enough detail to support consistency in the presentation, format, and degree of scaffolding observed in items and associated stimuli across test events.

- ✔ If an adaptive engine is used, item development specifications include details related to how content and skill characteristics required by the items should be coded to support the requirements of the adaptive algorithm and provide for the selection/administration of appropriate sets of items.

There is empirical evidence and an active research agenda supporting the validity of achievement scores as measures of the intended knowledge and skills.

- ✔ Correlations between assessment results and other reliable measures of the construct are positive and strong (e.g., assessment results, grades, etc.).
- ✔ Cognitive labs provide evidence supporting items developed to assess difficult to measure standards.
- ✔ The vendor has a research agenda in place which outlines validation activities that have been or will be conducted and the process by which research will be used to inform future test/item development activities.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.1 | The interim assessment provides for valid inferences about a student's current overall achievement in the target content domain. |
| Indicator 2.1.d | Achievement scores are appropriate for supporting their intended uses.<br>● The intended uses for the achievement scores are clearly and consistently articulated.<br>● There is sufficient theoretical and empirical evidence supporting the intended uses of achievement scores. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |

## About this indicator:

What is the purpose of this Indicator?

The validity of an assessment not only depends on the accuracy of score interpretations, but also on the degree to which theory or evidence supports the intended uses of the scores.

## Indicator 2.1.d

### What is reviewed?

- Documentation of test uses as provided in technical manuals, score reports, and interpretive guides
- Marketing materials used to advertise the utility of the assessment
- Validity evidence provided to support each intended use of overall achievement scores

### Evidence Necessary to Meet Expectations for Indicator 2.1.d

The intended uses for the achievement scores are clearly and consistently articulated.

✔ The recommended uses of the achievement scores are clearly provided.

- ✔ There is an alignment between the uses supported in technical documentation and score reports, and those uses advertised in assessment marketing materials.

There is sufficient theoretical and empirical evidence supporting the intended uses of achievement scores.

- ✔ The vendor supplies evidence to support the reasonableness of the intended uses, including empirical research. For example, if the assessment is used to recommend a particular instructional intervention for a set of similarly scoring students, evidence supporting the effectiveness of that intervention for the relevant subset of students is provided.
- ✔ If a study cited by the test publisher is not published, summaries are made available.

# Gateway 2: Technical Quality

## Criterion 2.2

### Predicted Student Performance

The interim assessment provides valid information regarding predicted student performance on a state's summative assessment or other intended criterion measure(s).

## What is the purpose of this Criterion?

Criteria 2-4 reflect the evidence necessary to evaluate the quality of additional information provided by interim assessments to inform decision making, including: predicted performance on the state summative assessment or a different criterion measure, performance on specific sub-skills, and growth measures or representations of progress over time.

**A Note on Gateway 2 Reviews:** The Technical Quality criteria are evaluated by statisticians and psychometricians trained by the Center of Assessment. These criteria evaluate the validity, reliability, and the quality of scores created by the interim assessments to ensure the data is high quality. This review requires a deep understanding of the information and scores generated by the assessment and how the information addresses the purpose of the assessment.

## Potential Sources of Evidence for Criterion 2.2

- Summaries of predictive validity studies that describe the relationships between the interim assessment and state summative assessments (for those states in which a prediction claim is made) or other intended criterion measures.
- Procedures and data used to support criterion or norm-referenced interpretations of predicted future performance.
- Summaries of studies evaluating the validity of the predicted classifications (e.g., decision consistency).
- Summaries of studies evaluating the predictive validity of interim test scores for predicting summative test scores or other criterion measures.
- Summaries of studies supporting the intended uses of the predicted information.

## Scoring

Points may vary based on indicators that are claimed by the publisher to be assessed.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.2 | The interim assessment provides valid information regarding predicted student performance on a state's summative assessment or other intended criterion measure(s). |
| Indicator 2.2.a*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | The design of the interim assessment supports its use in predicting performance on one or more external measures.<br>● Sufficient information is provided to evaluate the degree to which the construct or content domain targeted by the interim assessment is similar to that assessed by the criterion measure(s).<br>● The intended use of the interim assessment does not invalidate or contradict its appropriateness for predicting performance on the intended criterion measure(s).<br>● If an interim assessment was designed to predict performance on specific assessments (e.g., ACT, SAT), evidence supporting that claim is provided. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | | |

## About this indicator:

What is the purpose of this Indicator?

Due to a variety of factors, a student's performance on any two academic assessments will be correlated at some level. Consequently, any assessment can be used to predict performance on another if performance data for both assessments are available. In order to add value and ensure predicted results reflect future performance on a particular assessment, evidence must be provided that the design of the interim assessment facilitates the use of results for this purpose.

## Indicator 2.2.a

### What is reviewed?

- The assessment's technical report or related documentation that provides information about the design of the assessment, including the domain it is intended to assess and when/how frequently it is intended to be administered within an instructional sequence.
- Assessment design specifications – focusing on details reflecting the intended representation of test content, item complexity, item types, etc.
- When appropriate, documentation outlining the procedures used to ensure the interim assessment would predict performance on a specific criterion measure.
- Any documentation describing the conditions that should hold in order to use the interim assessment to predict performance on an external measure.

### Evidence Necessary to Meet Expectations for Indicator 2.2.a

Sufficient information is provided to evaluate the degree to which the construct or content domain targeted by the interim assessment is similar to that assessed by the criterion measure(s).

- ✔ The knowledge and skills addressed by the interim assessment are clearly related to those measured on the state summative assessment or criterion measure(s) for which predicted scores will be generated.
- ✔ If an interim assessment measures a small subset of the domain reflected by the summative assessment or criterion measure, evidence and a clear rationale are provided to support the use of the assessment for this purpose.
- ✔ If an interim assessment measures related content but at a grade level far removed from that being predicted, evidence and a clear rationale are provided to support the use of the assessment for this purpose

The intended use of the interim assessment does not invalidate or contradict it's appropriateness for predicting performance on the intended criterion measure(s).

- ✔ For *example, if the assessment was intended to only measure a specific sub-domain, it should not be used to predict performance on a summative assessment that tests the full content domain.*

If an interim assessment was designed to predict performance on a specific assessment (e.g., ACT, SAT) evidence supporting that claim is provided.

- ✔ The process used by the test vendor to design for this purpose is articulated in technical documentation

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.2 | The interim assessment provides valid information regarding predicted student performance on a state's summative assessment or other intended criterion measure(s). |
| Indicator 2.2.b*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | Predicted results are reliable.<br>• Procedures used for calculating and evaluating the reliability of predicted scores/classifications are well documented and appropriate.<br>• The reliability of the predicted result is calculated in a manner that is consistent with the inferences they were designed to support (e.g., CCR).<br>• The predictions demonstrate sufficient reliability to support their intended uses. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | | |

## About this indicator:

**What is the purpose of this Indicator?**

With any prediction there is going to some degree of error. This indicator is intended to rate the degree to which a predicted score/measure provides dependable information about how a student will perform on a future assessment.

## Indicator 2.2.b

### What is reviewed?

- Procedures used to estimate the reliability of predicted scores
- Standard errors of the prediction

- Classification accuracies

## Evidence Necessary to Meet Expectations for Indicator 2.2.b

Procedures used for calculating and evaluating the reliability of predicted scores/classifications are well documented and appropriate.

The reliability of the predicted result is calculated in a manner that is consistent with the inferences it was designed to support (e.g., CCR).

The predictions demonstrate sufficient reliability to support their intended uses.

- ✔ The standard errors around the prediction are reasonable and not so large as to potentially interfere with the intended interpretations and uses of the information.
- ✔ If predicted classifications are provided, classification accuracies are statistically higher than chance.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.2 | The interim assessment provides valid information regarding predicted student performance on a state's summative assessment or other intended criterion measure(s). |
| Indicator 2.2.c*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | Predicted results (e.g., expected scaled scores, performance levels, passing status, etc.) reflect a student's likely future performance on the state summative assessment or other intended criterion measure(s).<br>● The data and procedures used to establish and evaluate the predictive relationship for a given test-taking sample are documented and reasonable.<br>● The procedures used to support intended interpretations are clearly articulated.<br>● Studies support the appropriateness of the predicted result as a measure of future performance. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | | |

## About this indicator:

What is the purpose of this Indicator?

When the result of the interim assessment is used to predict student performance on a criterion measure such as the statewide summative assessment or a college entrance exam, evidence should be provided demonstrating that the data and procedures used make sense and that predicted scores can be interpreted and used in the manner intended. Predictions are typically conducted to support specific interpretations regarding a student's future performance. In many cases the interpretation is clear (e.g., Maya is predicted to obtain a score in the "proficient range" on the end of year assessment.) However in other cases the interpretation relies on additional information or a business rule that have been defined by the state (e.g., Based on performance on the interim assessment, Maya should be "on-grade level" by the end of the year or the "85$^{th}$ percentile" among her

academic peers.) As with overall achievement scores, the procedures and data used to support intended interpretations of predicted results should be well documented and meet best practice.

## Indicator 2.2.c

### What is reviewed?

- Procedures for establishing the predictive relationship (e.g., correlations, standard setting)
- Business rules for reporting predicted scores
- Interpretive guides for supporting interpretation and use of predictive information
- Validity studies conducted to support intended uses of the predictive scores

### Evidence Necessary to Meet Expectations for Indicator 2.2.c

The data and procedures used to establish and evaluate the predictive relationship for a given test taking sample are documented and reasonable.

- ✔ The sample of students used to establish the predictive relationship and evaluate the relationship between variables is representative of the full range of achievement on the interim assessment.
- ✔ The procedures used to link the assessments are appropriate given the type of prediction being made (e.g., score to score, score to performance level, performance level to performance level, etc.).
  - *For example, for purposes of linking specific test scores to specific levels of criterion performance regressions, equations are more useful than correlation coefficients; for dichotomous categorical variables logistic regression should be used.*[2]

The procedures used to support intended interpretations are clearly articulated.

- ✔ If the predicted result is used to support criterion-referenced interpretations of future performance (e.g., *on-track/not on track; college ready/not college ready; below, meeting or exceeding expectations*)
  - The expectations associated with performance in each level are clearly defined (e.g., What does it mean to be on track)
  - The process and data used to establish the cut scores defining performance in each level is well documented

If the score is used to support norm-referenced interpretations of future performance (percentile rank, grade equivalent).

- a clear description of the norm group is provided (e.g., If a norming study was conducted, the procedures and samples used.)
- the norm group is relevant (given the manner in which results are intended to be used), representative of the examinee population of interest and large enough to be reliable
- The test design and scale specifications support valid and appropriate normative inferences (i.e., the assessment and scale will result in an appropriate range of student performance; the score scale is broad enough to spread students along the ability continuum).

Studies support the appropriateness of the predicted result as a measure of future performance.

- ✔ Predictive validity studies clearly demonstrate that predictions of future performance are realized.

[2] Taken from Standard 1.18 of *Standards for Educational and Psychological Testing*

- ✔ Results should be provided for every assessment to which a prediction is made. In the case where an interim assessment program provides predictions to the state assessment in multiple states, there should be a predictive validity study for each state.
- ✔ Evidence is provided demonstrating the quality and utility of interim assessment scores for predicting future performance above and beyond information that is already freely and readily available to the end users (e.g., by comparing the quality of the prediction against that which would have been established using student performance on the previous year's summative assessment).

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.2 | The interim assessment provides valid information regarding predicted student performance on a state's summative assessment or other intended criterion measure(s). |
| Indicator 2.2.d*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | Predicted results are appropriate for supporting their intended uses.<br>• The intended uses for the predicted results are clearly and consistently articulated.<br>• There is sufficient theoretical and empirical evidence to support the appropriateness of the intended uses of predicted results. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | | |

## About this indicator:

What is the purpose of this Indicator?

The validity of an assessment not only depends on the accuracy of score interpretations, but also on the degree to which theory or evidence supports the intended uses of the scores.

## Indicator 2.2.d

### What is reviewed?

- Documentation of test uses as provided in technical manuals, score reports, and interpretive guides
- Marketing materials used to advertise the utility of the assessment
- Validity evidence supporting test use

### Evidence Necessary to Meet Expectations for Indicator 2.2.d

The intended uses for the predicted results are clearly and consistently articulated

- ✔ The recommended uses of the predictive information are clearly provided.
- ✔ There is an alignment between the uses supported in technical documentation and score reports, and those uses advertised in assessment marketing materials.

There is sufficient theoretical and empirical evidence to support the intended uses of predicted results.

- ✔ The vendor supplies evidence to support the reasonableness of the intended uses, including empirical research. For example, if the assessment is used to recommend an instructional intervention to a particular subset of students on the basis of the predictive information, evidence supporting the effectiveness of that intervention for the students in question is provided.
- ✔ If a study cited by the test publisher is not published, summaries are made available.

# Gateway 2: Technical Quality

## Criterion 2.3

### Sub-scores

The interim assessment provides for valid inferences about a student's specific areas of strength and need (e.g., at the reportable category, content strand or objective level).

## What is the purpose of this Criterion?

Criteria 2-4 reflect the evidence necessary to evaluate the quality of additional information provided by interim assessments to inform decision making, including: predicted performance on the state summative assessment or a different criterion measure, performance on specific sub-skills, and growth measures or representations of progress over time.

**A Note on Gateway 2 Reviews:** The Technical Quality criteria are evaluated by statisticians and psychometricians trained by the Center of Assessment. These criteria evaluate the validity, reliability, and the quality of scores created by the interim assessments to ensure the data is high quality. This review requires a deep understanding of the information and scores generated by the assessment and how the information addresses the purpose of the assessment.

## Potential Sources of Evidence for Criterion 2.3

- Test blueprints, test specifications
- Any validity studies conducted to support the use and interpretation of sub-score results as intended
- Scaling and norming procedures for sub-scores.
- Procedures for setting performance standards, including the writing of performance level descriptors (if applicable)
- Procedures and results of any conducted reliability and precision analyses for the sub-score results
- Score reports
- Use and interpretive guides

## Scoring

Points may vary based on indicators that are claimed by the publisher to be assessed.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.3 | The interim assessment provides for valid inferences about a student's specific areas of strength and need (e.g., at the reportable category, content strand or objective level). |
| Indicator 2.3.a*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | Test events are designed to provide specific information about a student's areas of strength and need in the content domain.<br>• The assessment design supports the reporting of sub-scores at each level of granularity for which they are provided<br>• The assessment design supports interpretations of students' areas of strength and need in the content domain. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | | |

## About this indicator:

What is the purpose of this Indicator?

Information about student strengths and weaknesses within the content domain are often reported on interim assessment score reports. Such information may be provided in the form of sub-scores, leveled performance categories (e.g., checkmarks and stop signs), or general statements that account for student performance across sets of items addressing similar or related skills. The purpose of this indicator is to evaluate the degree to which assessment design and development procedures support these types of interpretations overall and/or for each level of granularity at which sub-scores are reported (e.g., by standard, strand, objective, reportable category, etc.).

## Indicator 2.3.a

### What is reviewed?

- Construct definitions, score reports, and interpretive guides.
- Samples of test events that indicate which items are used to calculate/inform the reported sub-scores.
- Business rules and scaling procedures for calculating or aggregating sub-scores, when provided.
- Assessment design specifications that indicate:
    - the minimum number of items/points necessary to report sub-scores at each level of granularity for which they are provided, or inform statements regarding general areas of strength and need in the content domain;
      OR
    - the statistical criteria (e.g., minimum standard error threshold) necessary to support inferences about general areas of strength and need in the content domain given the implemented measurement model.

## Evidence Necessary to Meet Expectations for Indicator 2.3.a

The assessment design supports the reporting of sub-scores  at each level of granularity for which they are provided.

- ✔ Reported sub-scores reflect the content emphases depicted in assessment design documentation such as test blueprints and specifications.
- ✔ Test blueprints and specifications highlight content and statistical requirements underlying the reporting of sub-scores (e.g., minimum number of items/points, content representation).

The assessment design supports interpretations of students' areas of strength and need in the content domain.

- ✔ Test development documentation describes how items are tagged and aggregated to support inferences regarding areas of strength and need.
- ✔ Assessment design specifications highlight content and statistical requirements underlying the reporting of areas of strength and need based on student performance.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.3 | The interim assessment provides for valid inferences about a student's specific areas of strength and need (e.g., at the reportable category, content strand or objective level). |
| Indicator 2.3.b*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | Reported sub-scores are reliable.<br>• Estimates of reliability/precision are provided for all reported sub-scores.<br>• Procedures for calculating reliability indices and precision for the sub-score results are defensible and well documented.<br>• The calculated reliability and precision indices indicate adequate support for the intended interpretations and uses. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | | |

## About this indicator:

What is the purpose of this Indicator?

Sub-scores are particularly susceptible to low reliability due to the potential for only small numbers of items measuring each sub-domain. The purpose of this indicator is to evaluate the degree to which the reliabilities of reported sub-scores are sufficient to allow for intended inferences about achievement in the sub-domain and inform instructional decision making.

## Indicator 2.3.b

### What is reviewed?

- Procedures for calculating sub-score reliabilities and reported reliabilities
- Samples of sub-score reports and interpretive guides

## Evidence Necessary to Meet Expectations for Indicator 2.3.b

Estimates of reliability/precision are provided for all reported sub-scores.

Procedures for calculating reliability indices and precision for the sub-score results are defensible and well documented.

- ✔ Sub-scores reliabilities are calculated and reported in a manner that is consistent with the inferences the sub-scores were designed to support (e.g., on grade level).

The calculated reliability and precision indices indicate adequate support for the intended interpretations and uses.

- ✔ The sub-scores have adequate numbers of items to support reliable use.
- ✔ While acceptable values for reliability are context dependent, a general rule of thumb is that the minimum score reliability for low-stakes use is generally around .70.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.3 | The interim assessment provides for valid inferences about a student's specific areas of strength and need (e.g., at the reportable category, content strand or objective level). |
| Indicator 2.3.c*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | Reported sub-scores support intended interpretations of student performance in defined sub-skill areas.<br>● Evidence is provided to support intended interpretations of all reported sub-scores.<br>● Empirical data suggest sub-scores represent distinct sub-domains and should be reported separately. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | | |

## About this indicator:

What is the purpose of this Indicator?

Like overall achievement scores, sub-scores are often reported in ways that are intended to support specific interpretations of student performance, including criterion-referenced (e.g., on-grade level, on-track, mastered, etc.) and norm-referenced (e.g., percentile ranks, grade equivalents, etc.) Since these interpretations directly influence instructional decision making, the procedures and data used to support intended interpretations of sub-scores must be well documented and reported separately for each type of intended interpretation.

## Indicator 2.3.c

What is reviewed?

- *If sub-scores are reported as a transformation of raw scores or ability estimates to scale scores, the evaluator should review:*
  - Procedures used to establish the scaled scores, the characteristics of the scale (LOSS/HOSS) and any interpretations the scale was developed to support.
- *If sub-scores results are reported in terms of performance categories or levels, evaluators should review, as appropriate:*
  - Descriptions of performance levels and the procedures by which they were established (which may be content-based or empirically derived)
  - Procedures for establishing cut-scores that define the different performance levels
  - Evidence supporting the claims underlying performance in a particular category or beyond an established threshold (e.g., on-track, college and career ready, on-grade level).
- *If sub-scores are interpreted in consideration of the performance of a norm group the evaluator should review:*
  - A detailed summary of the characteristics of the norm group.
  - Procedures used to calculate and define reported norms (e.g., stanine, grade equivalents, etc.) including any business rules detailing criteria for inclusion in the norm group (e.g., how many items must a student respond to in order to be included in local norms).
- Dimensionality studies and/or studies evaluating convergent/discriminant validity for the different sub-scores.
- Validity studies that provide evidence of the appropriateness of the sub-scores as reflecting achievement in the intended sub-domain area.

## Evidence Necessary to Meet Expectations for Indicator 2.3.c

Evidence is provided to support intended interpretations of all reported sub-scores.

- ✔ If sub-scores are reported as a transformation of raw scores or ability estimates to scale scores:
  - The procedures used to translate student performance to scaled sub-scores are transparent and documented in enough detail to support consistent application across test events and administration.
  - The properties of the reportable scale facilitate the intended use and interpretation of results.
- ✔ If sub-scores results are reported in terms of performance categories or levels (e.g., Basic/Proficient/Advanced, pass/fail, on-track/not on track).
  - The expectations associated with performance in a given level (including above/below a defined threshold) are clearly defined.
  - The process used to establish cut scores used to inform sub-score reporting are well documented and reasonable given the associated interpretation (e.g., mastery, on grade level, etc.). For example, utilizing the cut score associated with “proficiency” on the total test may not appropriately reflect “proficiency” in a given sub-score area.
- ✔ If sub-scores are interpreted in consideration of the performance of a norm group:
  - A clear description of the norm group is provided. Note: in some cases sub-score interpretations may be purely relative within the own student’s performance—e.g., highlighting students areas of relative strength and weakness. In other cases, the norm group might be the student’s own class or school.
  - The norm group is relevant given the manner in which results are intended to be used.
  - The test design and scale specifications support valid and appropriate normative inferences (i.e., the assessment and scale will result in an appropriate range of student performance; the score scale is broad enough to spread students along the ability continuum).

- The reports do not include sub-scores that are purely descriptive (i.e., number of items correct) or without properties that allow for meaningful interpretations of student performance or support instructional use.

Empirical data suggest sub-scores represent distinct sub-domains and should be reported separately.

- ✔ Dimensionality analyses and/or correlations among sub-scores support claims that the sub scores represent distinct sub-domains rather than duplicative information.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.3 | The interim assessment provides for valid inferences about a student's specific areas of strength and need (e.g., at the reportable category, content strand or objective level). |
| Indicator 2.3.d*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | Reported sub-scores are appropriate for supporting their intended uses.<br>• The intended uses for the sub-scores are clearly and consistently articulated.<br>• There is sufficient theoretical and empirical evidence supporting the intended uses for the sub-scores. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | | |

## About this indicator:

What is the purpose of this Indicator?

The validity of an assessment not only depends on the accuracy of score interpretations, but also on the degree to which theory or evidence supports the intended uses of the scores.

### Indicator 2.3.d

#### What is reviewed?

- Documentation of test uses as provided in technical manuals, score reports, and interpretive guides
- Marketing materials used to advertise the utility of the assessment
- Validity evidence supporting test use

#### Evidence Necessary to Meet Expectations for Indicator 2.3.d

The intended uses for the sub-scores are clearly and consistently articulated.

- ✔ The recommended uses of the sub-scores are clearly provided.
  - There is an alignment between the uses supported in technical documentation and score reports, and those uses advertised in assessment marketing materials.

There is sufficient theoretical and empirical evidence supporting the intended uses for the sub-scores.

- ✔ The vendor supplies evidence to support the reasonableness of the intended uses, including empirical research. For example, if the assessment is used to recommend a particular instructional intervention for a student with a particular sub-score profile, evidence supporting the effectiveness of that intervention for students with the same or similar profiles of achievement is provided.

If a study cited by the test publisher is not published, summaries are made available.

# Gateway 2: Technical Quality

## Criterion 2.4

### Student Progress

The interim assessment provides valid information regarding student progress in the content domain.

## What is the purpose of this Criterion?

Criteria 2-4 reflect the evidence necessary to evaluate the quality of additional information provided by interim assessments to inform decision making, including: predicted performance on the state summative assessment or a different criterion measure, performance on specific sub-skills, and growth measures or representations of progress over time.

**A Note on Gateway 2 Reviews:** The Technical Quality criteria are evaluated by statisticians and psychometricians trained by the Center of Assessment. These criteria evaluate the validity, reliability, and the quality of scores created by the interim assessments to ensure the data is high quality. This review requires a deep understanding of the information and scores generated by the assessment and how the information addresses the purpose of the assessment.

## Potential Sources of Evidence for Criterion 2.4

- Test blueprints, test specifications
- Scaling and norming procedures
- Studies investigating the effect of ceiling or floor effects on the estimated growth scores
- Any validity studies collected to support the use and interpretation of growth information
- Procedures and results of any conducted precision analyses on the estimated growth scores

## Scoring

Points may vary based on indicators that are claimed by the publisher to be assessed.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.4 | The interim assessment provides valid information regarding student progress in the content domain. |
| Indicator 2.4.a*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | The interim assessment is designed to support measures of growth.<br>● Test design and content specifications (within and across grades) support the use of assessment results as a means of evaluating growth in the manner specified by the vendor.<br>● The technical characteristics of the test and reportable scale support the reported growth measure. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | | |

## About this indicator:

What is the purpose of this Indicator?

Interim assessments are often used to track student learning progress in the content domain over time. The purpose of this indicator is to evaluate the effectiveness of the interim assessment at providing valid information regarding student growth.

## Indicator 2.4.a

### What is reviewed?

- Test design and specification documents
- Procedures for calculating reported measures of student progress
- Validity studies related to the interpretation and use of growth scores

## Evidence Necessary to Meet Expectations for Indicator 2.4.a

Test design and content specifications (within and across grades) support the use of assessment results as a means of evaluating progress in the manner specified by the vendor.

- ✔ For example, the nature of how the assessed construct changes across years should align with the intended growth interpretation.

The technical characteristics of the test and reportable scale support the reported growth measure.

- ✔ There is sufficient variability in the scale score continuum to support inferences about student growth
- ✔ If vertical scale scores are used to make growth interpretations:
  - There is sufficient distinctness in the scale score ranges within and across grade levels to help prevent misinterpretations associated with the vertical scale
  - Documentation describing the construction of the vertical scale (e.g., design) and procedures for ongoing monitoring are provided.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.4 | The interim assessment provides valid information regarding student progress in the content domain. |
| Indicator 2.4.b*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress..* | Student growth scores are reliable.<br>● Procedures for estimating standard errors around the growth estimates are appropriate and well documented.<br>● The reliability of the growth scores have been evaluated for students at different places along the ability scale.<br>● The calculated reliability and precision indices indicate adequate support for the intended uses of the reported growth scores. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | | |

## About this indicator:

What is the purpose of this Indicator?

Due to measurement error in both achievement scores necessary to calculate growth, estimates of growth can often suffer from substantially increased measurement error. The purpose of this indicator is to evaluate the strength of the evidence for supporting the reliability of the reported growth scores for their intended uses.

## Indicator 2.4.b

### What is reviewed?

- Procedures for estimating reliability of growth scores and the resulting reliability estimates
- Samples of score reports and user interpretive guides regarding the reporting growth scores

## Evidence Necessary to Meet Expectations for Indicator 2.4.b

Procedures for estimating standard errors around the growth estimates are appropriate and well documented.

The reliability of the growth scores is evaluated for students at different places along the ability scale.

- ✔ Evidence is provided to support the reliability of the growth estimates for students across the full achievement continuum. When errors in the growth estimate change significantly as a result of the student achievement score, student-level errors, rather than mean errors, are provided in technical documentation.

The calculated reliability and precision indices indicate adequate support for the intended uses of the reported growth scores.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.4 | The interim assessment provides valid information regarding student progress in the content domain. |
| Indicator 2.4.c*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress..* | Student growth scores support the intended interpretations.<br>• The procedures and measures for calculating student growth are clearly documented and appropriate.<br>• If significant modifications are made to the interim assessment that might break the trend line (i.e., test design changes, rescaling, and shifts in performance standards), empirical evidence is provided to support the intended interpretations and uses of growth scores.<br>• Empirical evidence confirms that growth scores provide for valid intended inferences about student learning in the content domain. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | | |

## About this indicator:

What is the purpose of this Indicator?

Due to the great variability in how growth scores are calculated and reported across different assessment programs, it is critical that all assessment vendors clearly document the meaning of growth information. Additionally, the intended growth interpretation(s) must be consistent with the manner in which growth scores were calculated. For example, value-added models and student growth percentiles support norm-referenced interpretations of growth while gain score models are designed to support criterion-referenced interpretations.

## Indicator 2.4.c

### What is reviewed?

- Technical reports
- Sample score reports and interpretive guides for student growth scores
- Methods for calculating individual and aggregate student growth scores

## Evidence Necessary to Meet Expectations for Indicator 2.4.c

The procedures and measures for calculating student growth are clearly documented and appropriate.

- ✔ Calculated measures of student growth align with the intended interpretation (e.g., criterion-referenced vs. norm-referenced interpretations).
- ✔ When appropriate, procedures and business rules for calculating aggregate scores are provided (e.g., mean student growth percentiles at the teacher or school level).
- ✔ When student growth is interpreted with respect to attainment of a specified target, the process used to establish the target is clear.

If significant modifications are made to the interim assessment that might break the trend line (i.e., test design changes, rescaling, and shifts in performance standards), empirical evidence is provided to support the intended interpretations and uses of growth scores.

- ✔ Technical reports should clearly document changes made to the assessment design and provide evidence that they do not impact the validity of intended interpretations and uses.

Empirical evidence confirms that growth scores provide for valid inferences about student learning in the content domain.

- ✔ Evidence is provided which indicates that students who show growth on the assessment demonstrate improved performance in the content domain.

# Gateway 2: Technical Quality

| | |
|---|---|
| Criterion 2.4 | The interim assessment provides valid information regarding student progress in the content domain. |
| Indicator 2.4.d*<br>**This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress..* | Student growth scores support the intended interpretations.<br>● The intended uses for the growth scores are clearly and consistently articulated.<br>● There is sufficient theoretical and empirical evidence supporting the intended uses for the growth scores. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Meets expectations<br><br>There is sufficient, high quality evidence provided to support the range of expectations associated with the indicator. Expectations that are not supported by evidence are either reasonably explained by the vendor or not applicable given the assessment design. | 1 point<br><br>Partially meets expectations<br><br>There is some evidence provided to support the range of expectations associated with the indicator, but the evidence varies in quality and/or additional evidence is required to fully meet expectations. | 0 points<br><br>Does not meet expectations<br><br>No evidence or minimal evidence has been provided to support the indicator, OR the evidence provided is low quality and does not appropriately address the expectations outlined for this indicator. |
| **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | | |

## About this indicator:

What is the purpose of this Indicator?

The validity of an assessment not only depends on the accuracy of score interpretations, but also on the degree to which theory of evidence supports the intended uses of the scores.

## Indicator 2.4.d

### What is reviewed?

- Documentation of test uses as provided in technical manuals, score reports, and interpretive guides
- Marketing materials used to advertise the utility of the assessment
- Validity evidence supporting test use

### Evidence Necessary to Meet Expectations for Indicator 2.4.d

The intended uses for the growth scores are clearly and consistently articulated.

- ✔ The recommended uses of the growth information are clearly provided.
  - There is an alignment between the uses supported in technical documentation and score reports, and those uses advertised in assessment marketing materials.

There is sufficient theoretical and empirical evidence supporting the intended uses for the growth scores.

- ✔ The vendor supplies evidence to support the reasonableness of the intended uses, including empirical research. For example, if the assessment is used to recommend a particular instructional intervention for students in a particular growth range, evidence supporting the effectiveness of that intervention for students in that range of growth is provided.

If a study cited by the test publisher is not published, summaries are made available.

# Gateway 3: Score Reports and Interpretive Guides

## Criterion 3.1

### Overall Achievement

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of overall achievement results.

## What is the purpose of this Criterion?

Score reports and the resources[3] developed to guide each type of score-report user are vital to ensuring test results are interpreted and used in the manner intended. The criteria and indicators in Gateway 3 focus on the degree to which adequate information is provided to help intended users (e.g., educators, parents, students, administrators, or other specified users) interpret and use test results to appropriately inform decision making. When educator and psychometric reviewers conduct their evaluation of Gateway 3, they will only be evaluating the criteria in Gateway 3 that connect back to the criteria evaluated in Gateway 2.

## Research Connection

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system* [Policy brief]. p.1-8.

## Scoring:

| Meets Expectations | Partially Meets Expectations | Does Not Meet Expectations |
|---|---|---|
| ● 8-10 points | ● 5-7 points | ● <5 points |

[3] Such as interpretive guides, user manuals, and other informational documents and/or videos

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.1 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of overall achievement results. |
| Indicator 3.1.a | The design of the score reports and supporting materials (e.g., user manuals and interpretive guides) and the types of information provided are consistent with the intended interpretations and uses for specific users (e.g., educators, parents, students, or administrators).<br>● Score reports effectively represent the intended interpretations and uses of overall achievement results.<br>● The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses.<br>● Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when needed.<br>● Evidence (e.g., studies, focus groups) is provided that users are able to effectively interpret and use reports in the manner intended.<br>● The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences.<br>● Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly).<br>○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with information about score reports and the supporting materials, the design of those reports and the attention paid to users within the design process. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, OR the evidence is low quality and does not appropriately address the expectations associated with information about score reports and the supporting materials, the design of those |

| | | |
|---|---|---|
| | | reports, and the attention paid to users within the design process. |

## About this indicator:

What is the purpose of this Indicator?

Score reports are the vehicle of communication between assessment results and stakeholders. Because stakeholders may have different interests and, therefore, different purposes in mind for assessment outcomes, reports must be designed to effectively support test users in making the appropriate score interpretations and carrying out the intended score uses. The uses evaluated in this Gateway and through the EdReports review are only those that have been specified by the vendor as uses the assessment has been designed to support.

The purpose of this indicator is to evaluate the extent to which score reports and supporting materials effectively represent the information needed by each group of stakeholders. In addition, it is important that score reports and/or supporting materials warn of misuse, identify results that may compromise the integrity of the test, and clearly communicate the conditions that cause compromised results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
    - Cluster 3: Reporting and Interpretation (p. 119-120)
  - Chapter 7: “Supporting Documentation for Tests”
    - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
  - Chapter 12: “Educational Testing and Assessment”
    - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
  - D.1 Focusing on student achievement and progress toward readiness
  - D.2 Providing timely data that inform instruction

## Indicator 3.1.a Guiding Questions:

- Are score reports and supporting materials and the information provided designed to be consistent with the interpretations and uses for different types of users?
- Do the score reports and supporting materials effectively represent the intended interpretations and uses of the overall achievement results?
- Is the grain size of the information provided appropriate for effectively serving the intended interpretations and uses?
- Is evidence provided that shows that attention was paid to different audiences and users during the design process?
- Were there focus groups and/or studies in place to collect feedback from stakeholders about the ability to effectively interpret and use the reports in the manner intended?
- Does the documentation effectively warn against potential or common misuses of results that could result in negative unintended consequences for students?
- Do the reports identify or flag students for whom the integrity of the test interpretations may be compromised?

- Are the conditions which bring about a flag stated in reports or interpretive guides?

## Evidence Collection

- Find the list of types of score reports for various audiences (e.g., educators/professional learning community teams, parents, students, or administrators).
- Review the different types of reports.
- Review the organization style of score reports, e.g., headings, explanations, graphs & charts.
- Review the user manuals and interpretive guides to determine the degree of educational detail provided, including interpretations, score explanations, and how to use the reports.
- Identify how the score is reported and the grain size of the information.
- Consider the ease of accessibility for score reports.
- Evaluate the readability of score reports and variations of readability among various reports.
- Find any translated versions of the score reports and/or supporting materials.
- Review summaries of focus groups, studies, and any report development documentation provided that were used to inform the design of the reports for the various audiences.
- Review any focus group feedback or studies provided that show the ability of users to effectively interpret and use the reports.
- Note warnings associated with misuse of results.
- Review examples of score reports that have students flagged for whom test integrity may be compromised.
- Review interpretive guides or score reports for information on conditions which flag compromised tests.

## Cluster Meeting Discussion

- What types of reports are provided for users?
- Who are the different users listed for each type of score report?
- Do the score reports represent the intended interpretations and uses in a manner that is appropriate for the specific user?
- Do supporting materials provide enough information that supports the intended interpretations and uses of the results?
- Do supporting materials provide appropriate information to support the intended interpretations and uses of the results?
- How are the scores reported out?
- What is the grain size of the information?
    - Is the grain size appropriate for the intended uses?
- Are the score reports designed in such a way that any stakeholder group would understand?
    - Do headings and content organization make the intended interpretations clear and easy to understand in each version?
- Does the score report design confuse or conflate the intended uses with uses that come from other types of data?
- Is the cognitive load for the reports and the supporting materials appropriate for the interpretations and uses of that specific audience?
    - Is the readability for the parent/family report at an appropriate level?
- Are the score reports and supporting materials accessible to all stakeholder groups, through translations or other features?
- Is the information provided on each type of report appropriate to effectively serve the intended uses?
- Is the information provided the right grain size to represent the intended interpretations and uses?
- Is information provided on how the findings of focus groups, studies, etc. show that users are able to interpret and use the score reports as intended?

- Is information provided on how the feedback of focus groups, studies, etc. was used to make changes or improvements to specific score reports?
- Do sample reports clearly indicate when the integrity of a test has been compromised?
- Are warnings safeguarding the misinterpretation or misuse of scores clear and apparent?
- Are flags or other markers provided to identify students for whom the integrity of the test interpretations may be compromised?
    - Students not attempting a large number of items
    - Students with interrupted test administration
    - Students with an unreasonable response time
- If flags or markers for the score are provided, are the flags or markers clearly defined in the score report or interpretive guides?
- Are conditions mentioned in reports or interpretive guides that bring about a flag for the integrity of a test being compromised?
    - If conditions are mentioned, what are they and are they clearly articulated?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.1 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of overall achievement results. |
| Indicator 3.1.b | Score reports include information about the degree of error associated with the achievement score.<br>● For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability.<br>● Supports (e.g., illustrative examples, informational text) are provided to facilitate accurate interpretations of error estimates and clarify the practical implications of error on score use. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with information about the degree of error related to the achievement score. | 1 point<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with information about the degree of error related to the achievement score, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with information about the degree of error related to the achievement score, OR the evidence is low quality and does not appropriately address the expectations associated with information about the degree of error related to the achievement score. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the inclusion of information about the degree to which overall achievement results may be impacted by measurement error, and whether that information is appropriate and supported by clear guidance for interpretation. The guidance should clarify how measurement error should influence the interpretation and use of the results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”

        - Cluster 3: Reporting and Interpretation (p. 119-120)
    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.1.b Guiding Questions:

- Do score reports include information about the degree of error associated with the achievement score/classification and how it should be interpreted?
- Is the degree of error provided in a format that is clear and easy to understand?
- Is information necessary to support accurate interpretations of error estimates and clarify the implications of error on score uses provided in user guides, interpretive materials, and/or on score reports?

### Evidence Collection

- Review score reports for degree of error (e.g., confidence intervals, error bands, probability statements).
- Review any support materials provided (e.g., parent portals, data analysis guides for educators) for explanations about degree of error.
- Read the explanations provided for different audiences related to error in data interpretation.

### Cluster Meeting Discussion

- Is information about the degree of error provided?
- What is the format for the degree of error?
- Do the reports provide audience-appropriate reliability information in a manner that supports accurate interpretations regarding the degree of error associated with achievement scores?
    - Who are the audiences for the information?
- Do the support materials provide audience-appropriate explanations of “degree of error” and how it can be interpreted?
    - Who are the audiences represented in the explanations?
    - What is the quality of the explanations?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.1 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of overall achievement results. |
| Indicator 3.1.c | Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.<br>● Guidance is aligned to the use.<br>● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data.<br>● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with the guidance provided to support the intended interpretations and uses, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with the guidance provided to support the intended interpretations and uses, OR the evidence is low quality and does not appropriately address the expectations associated with the guidance provided to support the intended interpretations and uses. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the guidance that is provided to support the understanding and intended use of the score reports. These include any instructional or curricular supports that are provided in the score reports, manuals, or guides. The guidance should be aligned to the intended uses and should have a foundation in research or input from educators versed in using educational data.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.

    - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
        - Cluster 3: Reporting and Interpretation (p. 119-120)
    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.1.c Guiding Questions:

- Is the guidance provided sufficient and appropriate to support intended score interpretations and uses?
- Is there clear alignment between the guidance provided and the intended use?
- Is any guidance that is provided based on research and/or feedback from educators experienced in using educational data?
- Does the guidance provided support appropriate use for students at the full range of performance outcomes?

### Evidence Collection

- Identify any guidance (e.g., instructional or curricular supports) that is provided to support the interpretations and uses of the results.
- Review any research related to the creation of the guidance provided.
- Review any feedback from educators relative to the creation of the guidance provided.
- Identify guidance provided to support appropriate use for students at all ranges of performance outcomes.

### Cluster Meeting Discussion

- How much guidance is provided to support the intended interpretations and uses of the results?
    - Is it an appropriate amount?
- Is the guidance provided appropriate information to support the intended interpretations and uses of the results?
- How well does the guidance provided align with the intended uses?
- Is guidance provided that does not align with the intended uses?
- Is evidence provided that shows guidance was created using research?
- Is evidence provided that shows guidance was created in consultation with educators experienced in using educational data?
- Is there guidance provided to support appropriate use for students in the full range of performance outcomes (i.e., not just the lowest groups)?

# Gateway 3: Score Reports and Interpretive Guides

## Criterion 3.2

### Predicted Student Performance

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of predicted student performance.

## What is the purpose of this Criterion?

Score reports and the resources developed to guide each type of score-report user are vital to ensuring test results are interpreted and used in the manner intended. The criteria and indicators in Gateway 3 focus on the degree to which adequate information is provided to help intended users (e.g., educators, parents, students, administrators, or other specified users) interpret and use test results to appropriately inform decision making. When educator and psychometric reviewers conduct their evaluation of Gateway 3, they will only be evaluating the criteria in Gateway 3 that connect back to the criteria evaluated in Gateway 2.

## Research Connection

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system* [Policy brief]. p.1-8.

## Scoring

Points may vary based on indicators that are claimed by the publisher to be assessed.

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.2 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of predicted student performance. |
| Indicator 3.2.a*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | The design of the score reports and supporting materials (e.g., user manuals and interpretive guides) and the types of information provided are consistent with the intended interpretations and uses for specific users (e.g., educators, parents, students, or administrators).<br>• Score reports effectively represent the intended interpretations and uses of predicted student performance results.<br>• The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses.<br>• Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when applicable.<br>• Evidence is provided that users are able to effectively interpret and use reports in the manner intended.<br>• The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences.<br>• Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly).<br>  ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• Sufficient, high quality evidence supports the range of expectations associated with information about score reports and the supporting materials, the design of those reports and the attention paid to users within the design process. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• There is some evidence to support the range of expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• No evidence or minimal evidence supports the expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, OR the evidence is low quality and does not appropriately address the expectations associated with information about score reports and the supporting materials, the design of those |

| | | |
|---|---|---|
| | | reports, and the attention paid to users within the design process. |

## About this indicator:

What is the purpose of this Indicator?

Score reports are the vehicle of communication between assessment results and stakeholders. Because stakeholders may have different interests and, therefore, different purposes in mind for assessment outcomes, reports must be designed to effectively support test users in making the appropriate score interpretations and carrying out the intended score uses. The uses evaluated in this Gateway and through the EdReports review are only those that have been specified by the vendor as uses the assessment has been designed to support.

The purpose of this indicator is to evaluate the extent to which score reports and supporting materials effectively represent the information needed by each group of stakeholders. In addition, it is important that score reports and/or supporting materials warn of misuse, identify results that may compromise the integrity of the test, and clearly communicate the conditions that cause compromised results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation
    - Cluster 3: Reporting and Interpretation (p. 119-120)
- CCSSO Criteria for High-Quality Assessments
  - D.1 Focusing on student achievement and progress toward readiness
  - D.2 Providing timely data that inform instruction

## Indicator 3.2.a Guiding Questions:

- Are score reports and supporting materials and the information provided designed to be consistent with the interpretations and uses for different types of users?
- Do the score reports and supporting materials effectively represent the intended interpretations and uses of the predicted performance results?
- Is the grain size of the information provided appropriate for effectively serving the intended interpretations and uses?
- Is evidence provided that shows that attention was paid to different audiences and users during the design process?
- Were there focus groups and/or studies in place to collect feedback from stakeholders about the ability to effectively interpret and use the reports in the manner intended?
- Does the documentation effectively warn against potential or common misuses of results that could result in negative unintended consequences for students?
- Do the reports identify or flag students for whom the integrity of the test interpretations may be compromised?
- Are the conditions which bring about a flag stated in reports or interpretive guides?

## Evidence Collection

- Find the list of types of score reports for various audiences (e.g., educators/professional learning community teams, parents, students, or administrators).
- Review the different types of reports.
- Review the organization style of score reports, e.g., headings, explanations, graphs & charts.
- Review the user manuals and interpretive guides to determine the degree of educational detail provided, including interpretations, score explanations, and how to use the reports.
- Identify how the score is reported and the grain size of the information.
- Consider the ease of accessibility for score reports.
- Evaluate the readability of score reports and variations of readability among various reports.
- Find any translated versions of the score reports and/or supporting materials.
- Review summaries of focus groups, studies, and any report development documentation provided that were used to inform the design of the reports for the various audiences.
- Review any focus group feedback or studies provided that show the ability of users to effectively interpret and use the reports.
- Note warnings associated with misuse of results.
- Review examples of score reports that have students flagged for whom test integrity may be compromised.
- Review interpretive guides or score reports for information on conditions which flag compromised tests.

Identify Audience and Reports
- Find the list of types of score reports for various audiences (e.g., educators/professional learning community teams, parents, students, or administrators).
- Review the different types of reports.

Evaluate Design
- Review the organization style of score reports, e.g., headings, explanations, graphs & charts.
- Consider the ease of accessibility for score reports.
- Evaluate the readability of score reports and variations of readability among various reports.

Scores & Detailed Information
- Review the user manuals and interpretive guides to determine the degree of educational detail provided, including interpretations, score explanations, and how to use the reports.
- Identify how the score is reported and the grain size of the information.
- Find any translated versions of the score reports and/or supporting materials.
- Note warnings associated with misuse of results.

Studies of Score Report Quality
- Review summaries of focus groups, studies, and any report development documentation provided that were used to inform the design of the reports for the various audiences.
- Review any focus group feedback or studies provided that show the ability of users to effectively interpret and use the reports.

Flagging Scores with Possible Issues
- Review examples of score reports that have students flagged for whom test integrity may be compromised.
- Review interpretive guides or score reports for information on conditions which flag compromised tests.

## Cluster Meeting Discussion

- What types of reports are provided for users?

- Who are the different users listed for each type of score report?
- Do the score reports represent the intended interpretations and uses in a manner that is appropriate for the specific user?
- Do supporting materials provide enough information that supports the intended interpretations and uses of the results?
- Do supporting materials provide appropriate information to support the intended interpretations and uses of the results?
- How are the scores reported out?
- What is the grain size of the information?
    - Is the grain size appropriate for the intended interpretations and uses?
- Are the score reports designed in such a way that any stakeholder group would understand?
    - Do headings and content organization make the intended interpretations clear and easy to understand in each version?
- Does the score report design confuse or conflate the intended uses with uses that come from other types of data?
- Is the cognitive load for the reports and the supporting materials appropriate for the interpretations and uses of that specific audience?
    - Is the readability for the parent/family report at an appropriate level?
- Are the score reports accessible to all stakeholder groups, through translations or other features?
- Is the information provided on each type of report appropriate to effectively serve the intended uses?
- Is the information provided the right grain size to represent the intended interpretations and uses?
- Is information provided on how the findings of focus groups, studies, etc. show that users are able to interpret and use the score reports as intended?
- Is information provided on how the feedback of focus groups, studies, etc. was used to make changes or improvements to specific score reports?
- Do sample reports clearly indicate when the integrity of a test has been compromised?
- Are warnings safeguarding the misinterpretation or misuse of scores clear and apparent?
- Are flags or other markers provided to identify students for whom the integrity of the test interpretations may be compromised?
    - Students not attempting a large number of items
    - Students with interrupted test administration
    - Students with an unreasonable response time
- If flags or markers for the score are provided, are the flags or markers clearly defined in the score report or interpretive guides?
- Are conditions mentioned in reports or interpretive guides that bring about a flag for the integrity of a test being compromised?
    - If conditions are mentioned, what are they and are they clearly articulated?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.2 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of predicted student performance. |
| Indicator 3.2.b*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | Score reports include information about the degree of error associated with the predicted performance score.<br>● For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability.<br>● Supports (e.g., illustrative examples, informational text) are provided to facilitate accurate interpretations of error estimates and clarify the practical implications of error on score use. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with information about the degree of error related to the predicted performance score. | 1 point<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with information about the degree of error related to the predicted performance score, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with information about the degree of error related to the predicted performance score, OR the evidence is low quality and does not appropriately address the expectations associated with information about the degree of error related to the predicted performance score. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the inclusion of information about the degree to which predicted performance results may be impacted by measurement error, and whether that information is appropriate and supported by clear guidance for interpretation. The guidance should clarify how measurement error should influence the interpretation and use of the results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”

        - Cluster 3: Reporting and Interpretation (p. 119-120)
    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.2.b Guiding Questions:

- Do score reports include information about the degree of error associated with the predicted student performance score and how it should be interpreted?
- Is the degree of error provided in a format that is clear and easy to understand?
- Is information necessary to support accurate interpretations of error estimates and clarify the implications of error on score uses provided in user guides, interpretive materials, and/or on score reports?

### Evidence Collection

- Review score reports for degree of error (e.g., confidence intervals, error bands, probability statements).
- Review any support materials provided (e.g., parent portals, data analysis guides for educators) for explanations about degree of error.
- Read the explanations provided for different audiences related to error in data interpretation.

### Cluster Meeting Discussion

- Is information about the degree of error provided?
- What is the format for the degree of error?
- Do the reports provide audience-appropriate reliability information in a manner that supports accurate interpretations regarding the degree of error associated with predicted performance scores?
    - Who are the audiences for the information?
- Do the support materials provide audience-appropriate explanations of “degree of error” and how it can be interpreted?
    - Who are the audiences represented in the explanations?
    - What is the quality of explanations?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.2 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of predicted student performance. |
| Indicator 3.2.c*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.* | Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.<br>● Guidance is aligned to the use.<br>● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data.<br>● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with the guidance provided to support the intended interpretations and uses, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with the guidance provided to support the intended interpretations and uses, OR the evidence is low quality and does not appropriately address the expectations associated with the guidance provided to support the intended interpretations and uses. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the guidance that is provided to support the understanding and intended use of the score reports. These include any instructional or curricular supports that are provided in the score reports, manuals, or guides. The guidance should be aligned to the intended uses and should have a foundation in research or input from educators versed in using educational data.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.

    - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
        - Cluster 3: Reporting and Interpretation (p. 119-120)
    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.2.c Guiding Questions:

- Is the guidance provided sufficient and appropriate to support intended score interpretations and uses?
- Is there clear alignment between the guidance provided and the intended use?
- Is any guidance that is provided based on research and/or feedback from educators experienced in using educational data?
- Does the guidance provided support appropriate use for students at the full range of performance outcomes?

### Evidence Collection

- Identify any guidance (e.g., instructional or curricular supports) that is provided to support the interpretations and uses of the results.
- Review any research related to the creation of the guidance provided.
- Review any feedback from educators relative to the creation of the guidance provided.
- Identify guidance provided to support appropriate use for students at all ranges of performance outcomes.

### Cluster Meeting Discussion

- How much guidance is provided to support the intended interpretations and uses of the results?
    - Is it an appropriate amount?
- Is the guidance provided appropriate information to support the intended interpretations and uses of the results?
- How well does the guidance provided align with the intended uses?
- Is guidance provided that does not align with the intended uses?
- Is evidence provided that shows guidance was created using research?
- Is evidence provided that shows guidance was created in consultation with educators experienced in using educational data?
- Is there guidance provided to support appropriate use for students in the full range of performance outcomes (i.e., not just the lowest groups)?

# Gateway 3: Score Reports and Interpretive Guides

## Criterion 3.3

### Sub-scores

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of sub-scores.

## What is the purpose of this Criterion?

Score reports and the resources developed to guide each type of score-report user are vital to ensuring test results are interpreted and used in the manner intended. The criteria and indicators in Gateway 3 focus on the degree to which adequate information is provided to help intended users (e.g., educators, parents, students, administrators, or other specified users) interpret and use test results to appropriately inform decision making. When educator and psychometric reviewers conduct their evaluation of Gateway 3, they will only be evaluating the criteria in Gateway 3 that connect back to the criteria evaluated in Gateway 2.

## Research Connection

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing.* American Educational Research Association.
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system* [Policy brief]. p.1-8.

## Scoring

Points may vary based on indicators that are claimed by the publisher to be assessed.

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.3 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of sub-scores. |
| Indicator 3.3.a*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | The design of the score reports and supporting materials (e.g., user manuals and interpretive guides) and the types of information provided are consistent with the intended interpretations and uses for specific users (e.g., educators, parents, students, or administrators).<br>• Score reports effectively represent the intended interpretations and uses of sub-scores.<br>• The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses.<br>• Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when applicable.<br>• Evidence is provided that users are able to effectively interpret and use reports in the manner intended.<br>• The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences.<br>• Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly).<br>  ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• Sufficient, high quality evidence supports the range of expectations associated with information about score reports and the supporting materials, the design of those reports and the attention paid to users within the design process. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• There is some evidence to support the range of expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• No evidence or minimal evidence supports the expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, OR the evidence is low quality and does not appropriately address the expectations associated with information about score reports and the supporting materials, the design of those |

| | | |
|---|---|---|
| | | reports, and the attention paid to users within the design process. |

## About this indicator:

What is the purpose of this Indicator?

Score reports are the vehicle of communication between assessment results and stakeholders. Because stakeholders may have different interests and, therefore, different purposes in mind for assessment outcomes, reports must be designed to effectively support test users in making the appropriate score interpretations and carrying out the intended score uses. The uses evaluated in this Gateway and through the EdReports review are only those that have been specified by the vendor as uses the assessment has been designed to support.

The purpose of this indicator is to evaluate the extent to which score reports and supporting materials effectively represent the information needed by each group of stakeholders. In addition, it is important that score reports and/or supporting materials warn of misuse, identify results that may compromise the integrity of the test, and clearly communicate the conditions that cause compromised results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
    - Cluster 3: Reporting and Interpretation (p. 119-120)
- CCSSO Criteria for High-Quality Assessments
  - D.1 Focusing on student achievement and progress toward readiness
  - D.2 Providing timely data that inform instruction

## Indicator 3.3.a Guiding Questions:

- Are score reports and supporting materials and the information provided designed to be consistent with the interpretations and uses for different types of users?
- Do the score reports and supporting materials effectively represent the intended interpretations and uses of the sub-scores?
- Is the grain size of the information provided appropriate for effectively serving the intended interpretations and uses?
- Is evidence provided that shows that attention was paid to different audiences and users during the design process?
- Were there focus groups and/or studies in place to collect feedback from stakeholders about the ability to effectively interpret and use the reports in the manner intended?
- Does the documentation effectively warn against potential or common misuses of results that could result in negative unintended consequences for students?
- Do the reports identify or flag students for whom the integrity of the test interpretations may be compromised?
- Are the conditions which bring about a flag stated in reports or interpretive guides?

## Evidence Collection

- Find the list of types of score reports for various audiences (e.g., educators/professional learning community teams, parents, students, or administrators).
- Review the different types of reports.
- Review the organization style of score reports, e.g., headings, explanations, graphs & charts.
- Review the user manuals and interpretive guides to determine the degree of educational detail provided, including interpretations, score explanations, and how to use the reports.
- Identify how the score is reported and the grain size of the information.
- Consider the ease of accessibility for score reports.
- Evaluate the readability of score reports and variations of readability among various reports.
- Find any translated versions of the score reports and/or supporting materials.
- Review summaries of focus groups, studies, and any report development documentation provided that were used to inform the design of the reports for the various audiences.
- Review any focus group feedback or studies provided that show the ability of users to effectively interpret and use the reports.
- Note warnings associated with misuse of results.
- Review examples of score reports that have students flagged for whom test integrity may be compromised.
- Review interpretive guides or score reports for information on conditions which flag compromised tests.

## Cluster Meeting Discussion

- What types of reports are provided for users?
- Who are the different users listed for each type of score report?
- Do the score reports represent the intended interpretations and uses in a manner that is appropriate for the specific user?
- Do supporting materials provide enough information that supports the intended interpretations and uses of the results?
- Do supporting materials provide appropriate information to support the intended interpretations and uses of the results?
- How are the scores reported out?
- What is the grain size of the information?
    - Is the grain size appropriate for the intended uses?
- Are the score reports designed in such a way that any stakeholder group would understand?
    - Do headings and content organization make the intended interpretations clear and easy to understand in each version?
- Does the score report design confuse or conflate the intended uses with uses that come from other types of data?
- Is the cognitive load for the reports and the supporting materials appropriate for the interpretations and uses of that specific audience?
    - Is the readability for the parent/family report at an appropriate level?
- Are the score reports and supporting materials accessible to all stakeholder groups, through translations or other features?
- Is the information provided on each type of report appropriate to effectively serve the intended uses?
- Is the information provided the right grain size to represent the intended interpretations and uses?
- Is information provided on how the findings of focus groups, studies, etc. show that users are able to interpret and use the score reports as intended?
- Is information provided on how the feedback of focus groups, studies, etc. was used to make changes or improvements to specific score reports?
- Do sample reports clearly indicate when the integrity of a test has been compromised?
- Are warnings safeguarding the misinterpretation or misuse of scores clear and apparent?

- Are flags or other markers provided to identify students for whom the integrity of the test interpretations may be compromised?
  - Students not attempting a large number of items
  - Students with interrupted test administration
  - Students with an unreasonable response time
- If flags or markers for the score are provided, are the flags or markers clearly defined in the score report or interpretive guides?
- Are conditions mentioned in reports or interpretive guides that bring about a flag for the integrity of a test being compromised?
  - If conditions are mentioned, what are they and are they clearly articulated?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.3 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of sub-scores. |
| Indicator 3.3.b*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | Score reports include information about the degree of error associated with sub-scores.<br>● For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability.<br>● Supports (e.g., illustrative examples, informational text) are provided to facilitate accurate interpretations of error estimates and clarify the practical implications of error on score use. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with information about the degree of error related to sub-scores. | 1 point<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with information about the degree of error related to sub-scores, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with information about the degree of error related to sub-scores, OR the evidence is low quality and does not appropriately address the expectations associated with information about the degree of error related to sub-scores. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the inclusion of information about the degree to which sub-scores may be impacted by measurement error, and whether that information is appropriate and supported by clear guidance for interpretation. The guidance should clarify how measurement error should influence the interpretation and use of the results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
    - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
        - Cluster 3: Reporting and Interpretation (p. 119-120)

    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.3.b Guiding Questions:

- Do score reports include information about the degree of error associated with reported sub-score results and how it is to be interpreted?
- Is the degree of error provided in a format that is clear and easy to understand?
- Is information necessary to support accurate interpretations of error estimates and clarify the implications of error on score uses provided in user guides, interpretive materials, and/or on score reports?

### Evidence Collection

- Review score reports for degree of error (e.g., confidence intervals, error bands, probability statements).
- Review any support materials provided (e.g., parent portals, data analysis guides for educators) for explanations about degree of error.
- Read the explanations provided for different audiences related to error in data interpretation.

### Cluster Meeting Discussion

- Is information about the degree of error provided?
- What is the format for the degree of error?
- Do the reports provide audience-appropriate reliability information in a manner that supports accurate interpretations regarding the degree of error associated with sub-scores?
    - Who are the audiences for the information?
- Do the support materials provide audience-appropriate explanations of “degree of error” and how it can be interpreted?
    - Who are the audiences represented in the explanations?
    - What is the quality of explanations?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.3 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of sub-scores. |
| Indicator 3.3.c*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed with the use of sub-scores.* | Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.<br>● Guidance is aligned to the use.<br>● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data.<br>● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with the guidance provided to support the intended interpretations and uses, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with the guidance provided to support the intended interpretations and uses, OR the evidence is low quality and does not appropriately address the expectations associated with the guidance provided to support the intended interpretations and uses. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the guidance that is provided to support the understanding and intended use of the score reports. These include any instructional or curricular supports that are provided in the score reports, manuals, or guides. The guidance should be aligned to the intended uses and should have a foundation in research or input from educators versed in using educational data.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”

        - Cluster 3: Reporting and Interpretation (p. 119-120)
    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.3.c Guiding Questions:

- Is the guidance provided sufficient and appropriate to support intended score interpretations and uses?
- Is there clear alignment between the guidance provided and the intended use?
- Is any guidance that is provided based on research and/or feedback from educators experienced in using educational data?
- Does the guidance provided support appropriate use for students at the full range of performance outcomes?

### Evidence Collection

- Identify any guidance (e.g., instructional or curricular supports) that is provided to support the interpretations and uses of the results.
- Review any research related to the creation of the guidance provided.
- Review any feedback from educators relative to the creation of the guidance provided.
- Identify guidance provided to support appropriate use for students at all ranges of performance outcomes.

### Cluster Meeting Discussion

- How much guidance is provided to support the intended interpretations and uses of the results?
    - Is it an appropriate amount?
- Is the guidance provided appropriate information to support the intended interpretations and uses of the results?
- How well does the guidance provided align with the intended uses?
- Is guidance provided that does not align with the intended uses?
- Is evidence provided that shows guidance was created using research?
- Is evidence provided that shows guidance was created in consultation with educators experienced in using educational data?
- Is there guidance provided to support appropriate use for students in the full range of performance outcomes (i.e., not just the lowest groups)?

# Gateway 3: Score Reports and Interpretive Guides

## Criterion 3.4

### Student Progress

Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of student growth or progress results.

## What is the purpose of this Criterion?

Score reports and the resources developed to guide each type of score-report user are vital to ensuring test results are interpreted and used in the manner intended. The criteria and indicators in Gateway 3 focus on the degree to which adequate information is provided to help intended users (e.g., educators, parents, students, administrators, or other specified users) interpret and use test results to appropriately inform decision making. When educator and psychometric reviewers conduct their evaluation of Gateway 3, they will only be evaluating the criteria in Gateway 3 that connect back to the criteria evaluated in Gateway 2.

## Research Connection

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). *The role of interim assessments in a comprehensive assessment system* [Policy brief]. p.1-8.

## Scoring

Points may vary based on indicators that are claimed by the publisher to be assessed.

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.4 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of student growth or progress results. |
| Indicator 3.4.a*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | The design of the score reports and supporting materials (e.g., user manuals and interpretive guides) and the types of information provided are consistent with the intended interpretations and uses for specific users (e.g., educators, parents, students, or administrators).<br>• Score reports effectively represent the intended interpretations and uses of student growth or progress results.<br>• The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses.<br>• Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when applicable.<br>• Evidence is provided that users are able to effectively interpret and use reports in the manner intended.<br>• The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences.<br>• Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly).<br>  ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>• Sufficient, high quality evidence supports the range of expectations associated with information about score reports and the supporting materials, the design of those reports and the attention paid to users within the design process. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>• There is some evidence to support the range of expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• No evidence or minimal evidence supports the expectations associated with information about score reports and the supporting materials, the design of those reports, and the attention paid to users within the design process, OR the evidence is low quality and does not appropriately address the expectations associated with information about score reports and the supporting materials, the design of those |

| | | |
|---|---|---|
| | | reports, and the attention paid to users within the design process. |

## About this indicator:

What is the purpose of this Indicator?

Score reports are the vehicle of communication between assessment results and stakeholders. Because stakeholders may have different interests and, therefore, different purposes in mind for assessment outcomes, reports must be designed to effectively support test users in making the appropriate score interpretations and carrying out the intended score uses. The uses evaluated in this Gateway and through the EdReports review are only those that have been specified by the vendor as uses the assessment has been designed to support.

The purpose of this indicator is to evaluate the extent to which score reports and supporting materials effectively represent the information needed by each group of stakeholders. In addition, it is important that score reports and/or supporting materials warn of misuse, identify results that may compromise the integrity of the test, and clearly communicate the conditions that cause compromised results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
    - Cluster 3: Reporting and Interpretation (p. 119-120)
- CCSSO Criteria for High-Quality Assessments
  - D.1 Focusing on student achievement and progress toward readiness
  - D.2 Providing timely data that inform instruction

## Indicator 3.4.a Guiding Questions:

- Are score reports and supporting materials and the information provided designed to be consistent with the interpretations and uses for different types of users?
- Do the score reports and supporting materials effectively represent the intended interpretations and uses of the student growth or progress results?
- Is the grain size of the information provided appropriate for effectively serving the intended interpretations and uses?
- Is evidence provided that shows that attention was paid to different audiences and users during the design process?
- Were there focus groups and/or studies in place to collect feedback from stakeholders about the ability to effectively interpret and use the reports in the manner intended?
- Does the documentation effectively warn against potential or common misuses of results that could result in negative unintended consequences for students?
- Do the reports identify or flag students for whom the integrity of the test interpretations may be compromised?
- Are the conditions which bring about a flag stated in reports or interpretive guides?

## Evidence Collection

- Find the list of types of score reports for various audiences (e.g., educators/professional learning community teams, parents, students, or administrators).
- Review the different types of reports.
- Review the organization style of score reports, e.g., headings, explanations, graphs & charts.
- Review the user manuals and interpretive guides to determine the degree of educational detail provided, including interpretations, score explanations, and how to use the reports.
- Identify how the score is reported and the grain size of the information.
- Consider the ease of accessibility for score reports.
- Evaluate the readability of score reports and variations of readability among various reports.
- Find any translated versions of the score reports and/or supporting materials.
- Review summaries of focus groups, studies, and any report development documentation provided that were used to inform the design of the reports for the various audiences.
- Review any focus group feedback or studies provided that show the ability of users to effectively interpret and use the reports.
- Note warnings associated with misuse of results.
- Review examples of score reports that have students flagged for whom test integrity may be compromised.
- Review interpretive guides or score reports for information on conditions which flag compromised tests.

## Cluster Meeting Discussion

- What types of reports are provided for users?
- Who are the different users listed for each type of score report?
- Do the score reports represent the intended interpretations and uses in a manner that is appropriate for the specific user?
- Do supporting materials provide enough information that supports the intended interpretations and uses of the results?
- Do supporting materials provide appropriate information to support the intended interpretations and uses of the results?
- How are the scores reported out?
- What is the grain size of the information?
    - Is the grain size appropriate for the intended uses?
- Are the score reports designed in such a way that any stakeholder group would understand?
    - Do headings and content organization make the intended interpretations clear and easy to understand in each version?
- Does the score report design confuse or conflate the intended uses with uses that come from other types of data?
- Is the cognitive load for the reports and the supporting materials appropriate for the interpretations and uses of that specific audience?
    - Is the readability for the parent/family report at an appropriate level?
- Are the score reports and supporting materials accessible to all stakeholder groups, through translations or other features?
- Is the information provided on each type of report appropriate to effectively serve the intended uses?
- Is the information provided the right grain size to represent the intended interpretations and uses?
- Is information provided on how the findings of focus groups, studies, etc. show that users are able to interpret and use the score reports as intended?
- Is information provided on how the feedback of focus groups, studies, etc. was used to make changes or improvements to specific score reports?
- Do sample reports clearly indicate when the integrity of a test has been compromised?
- Are warnings safeguarding the misinterpretation or misuse of scores clear and apparent?

- Are flags or other markers provided to identify students for whom the integrity of the test interpretations may be compromised?
  - Students not attempting a large number of items
  - Students with interrupted test administration
  - Students with an unreasonable response time
- If flags or markers for the score are provided, are the flags or markers clearly defined in the score report or interpretive guides?
- Are conditions mentioned in reports or interpretive guides that bring about a flag for the integrity of a test being compromised?
  - If conditions are mentioned, what are they and are they clearly articulated?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.4 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of student growth or progress results. |
| Indicator 3.4.b*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | Score reports include information about the degree of error associated with the student progress score.<br>• For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability.<br>• Supports (e.g., illustrative examples, informational text) are provided to facilitate accurate interpretations of error estimates and clarify the practical implications of error on score use. |

| Scoring | | |
|---|---|---|
| 2 points<br><br>Materials meet expectations of this indicator.<br><br>• Sufficient, high quality evidence supports the range of expectations associated with information about the degree of error related to the student growth or progress score. | 1 point<br><br>Materials partially meet expectations of this indicator.<br><br>• There is some evidence to support the range of expectations associated with information about the degree of error related to the student growth or progress score, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>• No evidence or minimal evidence supports the expectations associated with information about the degree of error related to the student growth or progress score, OR the evidence is low quality and does not appropriately address the expectations associated with information about the degree of error related to the student growth or progress score. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the inclusion of information about the degree to which student growth or progress results may be impacted by measurement error, and whether that information is appropriate and supported by clear guidance for interpretation. The guidance should clarify how measurement error should influence the interpretation and use of the results.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].

- *Standards for Educational and Psychological Testing*.
  - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”
    - Cluster 3: Reporting and Interpretation (p. 119-120)
  - Chapter 7: “Supporting Documentation for Tests”
    - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
  - Chapter 12: “Educational Testing and Assessment”
    - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
  - D.1 Focusing on student achievement and progress toward readiness
  - D.2 Providing timely data that inform instruction

## Indicator 3.4.b Guiding Questions:

- Do score reports include information about the degree of error associated with student growth or progress measures and how it should be interpreted?
- Is the degree of error provided in a format that is clear and easy to understand?
- Is information necessary to support accurate interpretations of error estimates and clarify the implications of error on score uses provided in user guides, interpretive materials, and/or on score reports?

### Evidence Collection

- Locate explanations for effects of measurement error on growth reports; consider explanations related to error in data interpretation.
- Review score reports for degree of error representation (e.g., confidence intervals, error bands, probability statements).
- Review any support materials provided (e.g., parent portals, data analysis guides for educators) for explanations about degree of error.
- Read the explanations provided for different audiences related to error in data interpretation.

### Cluster Meeting Discussion

- Is information about the degree of error provided?
- What is the format for the degree of error?
- Do the reports provide audience-appropriate reliability information in a manner that supports accurate interpretations regarding the degree of error associated with the student growth or progress scores?
  - Who are the audiences for the information?
- Do the support materials provide audience-appropriate explanations of “degree of error” and how it can be interpreted?
  - What audiences are represented in the explanations?
  - What is the quality of the explanations?

# Gateway 3: Score Reports and Interpretive Guides

| | |
|---|---|
| Criterion 3.4 | Score reports and other resources (e.g., user manuals, interpretive guides, instructional or curricular resources) are appropriate for facilitating the intended interpretations and uses of student growth or progress results. |
| Indicator 3.4.c*<br>**These claims should match claims made and evaluated in Gateway 2. This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.* | Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.<br>● Guidance is aligned to the use.<br>● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data.<br>● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes. |

| Scoring | | |
|---|---|---|
| 4 points<br><br>Materials meet expectations of this indicator.<br><br>● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses. | 2 points<br><br>Materials partially meet expectations of this indicator.<br><br>● There is some evidence to support the range of expectations associated with the guidance provided to support the intended interpretations and uses, but the evidence varies in quality and/or sufficiency. | 0 points<br><br>Materials DO NOT meet expectations of this indicator.<br><br>● No evidence or minimal evidence supports the expectations associated with the guidance provided to support the intended interpretations and uses, OR the evidence is low quality and does not appropriately address the expectations associated with the guidance provided to support the intended interpretations and uses. |

## About this indicator:

What is the purpose of this Indicator?

The purpose of this indicator is to evaluate the guidance that is provided to support the understanding and intended use of the score reports. These include any instructional or curricular supports that are provided in the score reports, manuals, or guides. The guidance should be aligned to the intended uses and should have a foundation in research or input from educators versed in using educational data.

Resources:

- *The role of interim assessments in a comprehensive assessment system* [Policy brief].
- *Standards for Educational and Psychological Testing*.
    - Chapter 6: “Test Administration, Scoring, Reporting, and Interpretation”

        - Cluster 3: Reporting and Interpretation (p. 119-120)
    - Chapter 7: “Supporting Documentation for Tests”
        - Cluster 3: Content of Test Documents: Test Administration and Scoring (p. 127-129)
    - Chapter 12: “Educational Testing and Assessment”
        - Cluster 2: Use and Interpretation of Educational Assessments 197-200)
- CCSSO Criteria for High-Quality Assessments
    - D.1 Focusing on student achievement and progress toward readiness
    - D.2 Providing timely data that inform instruction

## Indicator 3.4.c Guiding Questions:

- Is the guidance provided sufficient and appropriate to support intended score interpretations and uses?
- Is there clear alignment between the guidance provided and the intended use?
- Is any guidance that is provided based on research and/or feedback from educators experienced in using educational data?
- Does the guidance provided support appropriate use for students at the full range of performance outcomes?

### Evidence Collection

- Identify any guidance (e.g., instructional or curricular supports) that is provided to support the interpretations and uses of the results.
- Review any research related to the creation of the guidance provided.
- Review any feedback from educators relative to the creation of the guidance provided.
- Identify guidance provided to support appropriate use for students at all ranges of performance outcomes.

### Cluster Meeting Discussion

- How much guidance is provided to support the intended interpretations and uses of the results?
    - Is it an appropriate amount?
- Is the guidance provided appropriate information to support the intended interpretations and uses of the results?
- How well does the guidance provided align with the intended uses?
- Is guidance provided that does not align with the intended uses?
- Is evidence provided that shows guidance was created using research?
- Is evidence provided that shows guidance was created in consultation with educators experienced in using educational data?
- Is there guidance provided to support appropriate use for students in the full range of performance outcomes (i.e., not just the lowest groups)?