edreports.org

Center for Assessment

Evidence Guide

Interim Assessment

Mathematics

Grades 3-8

v1.0

Index

Gateway 1

Gateway 2

Gateway 3

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.

In mathematics, EdReports designed Criterion 1.1 and 1.2 indicators to reflect the importance of focus and rigor, but allow for test publishers to have flexibility within those constraints.

Focus:
Focus allows students to delve deeper and develop mastery over critical topics. Because students master topics they can move on to new topics that build from these foundations. Major and supporting clusters consist of these topics and standards, which are utilized to keep building on these learning progressions toward college and career readiness.

- The focus requirements in indicators 1.1 are around assessments demonstrating the following: in elementary grades, there should be more emphasis placed on numbers and operations, while in middle school, ratios and proportional relationships plus algebra are key topics. While CCSSO Criteria for High Quality Assessments influenced the need to have numerical thresholds in assessment for major work topics, EdReports worked to balance that with findings in curriculum reviews-thus determining a minimum similar to curriculum reviews.
- Major work from previous or future grade levels that are part of test events will be counted toward focus, so that the innovation of focusing strongly on the most important content can be met within CATs.

Rigor:
In order to demonstrate knowledge of standards and proficiency toward mastery, EdReports believes it is important that questions address all 3 aspects of rigor and have some degree of balance in approach (i.e. not just fluency questions).

- Having indicators 1.2.d-1.2.g that concentrate solely on conceptual understanding, application, procedural skills and fluency, and mathematical practices shows the emphasis EdReports deems is necessary in rigor for any assessment. Although EdReports requires these indicators to be reviewed, we will be looking more for representation and a significant proportion to be there rather than have a strict upper threshold.

Coherence:

- There are no indicators around coherence because, while critical in materials and instruction throughout the year and across grades, assessments do not typically build on concepts and learning in the same way as instructional materials. However, coherence between mathematical concepts within a grade or course may be reflected when students encounter assessment items that examine an aspect of rigor, especially conceptual understanding or application. It may also be demonstrated in score reports that suggest actionable steps which connect previous or future math concepts to those reflected in the assessment data. If the assessment is assessing the standards in a focused and rigorous way, it likely supports coherence.

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?

Quality mathematics interim assessments reflect mathematics as a coherent topic focused on major topics that assess conceptual understanding, procedural skills, fluencies, applications, and mathematical practices. The development of these assessment systems require a strategic plan for the design of the assessment, including test blueprints or assessment design specifications which focus on the most important topics for college- and career-readiness as described in Criterion 1.1.

Research Connection

- Common Core State Standards for Mathematics (CCSSM)
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments
- 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.
- 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.
- Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal Design for Learning: Theory and Practice*. Wakefield, MA: CAST. pp. 73-76.

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. ● Assessment rationale explains the design of the assessment, the benefits of the assessment, and a research foundation grounding the assessment process. ● Item development documentation is sufficiently robust to support the writing and review of items measuring CCR standards. ● Across all item types, assessment design specifications provide clear scoring information and/or rubrics to evaluate students' levels of understanding with respect to the standards. ● Item development documentation includes a description of processes used to ensure items are content-accurate and without technical or editorial flaws. ● The suggested ranges of cognitive demand reflected in assessment design specifications are sufficient to measure the depth of the standards.

Scoring		
4 points Materials meet expectations of this indicator. ● 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. ● Item development documentation and writing guidelines provide clear direction in the writing of high-quality CCR items. ● Item development documentation provides clear direction for a review process	2 points Materials partially meet expectations of this indicator. ● 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. ● Item development documentation and writing guidelines provide guidance in the writing of high-quality items, but are not specific or comprehensive enough to ensure all items are high quality. ● Item development documentation provides direction for a review process	0 points Materials DO NOT meet expectations of this indicator. ● The development documentation lacks any rationale regarding the design, benefits, or research grounding the assessment or process. ● Item development documentation and writing guidelines lack guidance in the writing of high-quality items. ● Item development documentation lacks guidance for a review process resulting in high-quality CCR items. ● Scoring specifications and criteria are either not present or not aligned. ● Assessment documentation lacks review processes and/or

resulting in high-quality CCR items. ● Scoring specifications and criteria are standards aligned, clearly communicated, and adequate to ensure inter-rater reliability. ● There is evidence of processes used to ensure the technical quality and editorial accuracy of the items. ● Assessment design specifications provide clear evidence that the cognitive demand is sufficient to measure the standards.	resulting in high-quality CCR items. ● Scoring specifications and criteria are present and aligned but not specific enough to ensure consistent score reporting for constructed response items. ● Processes are in place to support alignment to standards, technical quality, and editorial accuracy of items, but these processes lack the specificity necessary to ensure consistency. ● Assessment design specifications indicate some evidence the cognitive demand is sufficient to measure the standards.	guidance to ensure alignment to standards, technical quality, and editorial accuracy of items. ● Assessment design specifications lack clear evidence that the cognitive demand is sufficient to measure the standards.

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, including: item writer guidelines, rubric design, item review guidelines, assessment creation, and technical processes for review, revision/correction, and ultimate approval. Additionally, this indicator evaluates whether adequate processes are in place to ensure the technical quality and editorial accuracy of the assessment items and how this documentation ensures there is an appropriate range of cognitive demand.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - C.5. Ensuring high-quality items and a variety of item types
  - 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
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics part 1
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments
- *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 the 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 rubrics to evaluate students' levels of proficiency with respect to the standards?
- Does item development documentation include a description of processes used to ensure items are content-accurate and without technical or editorial flaws?
- Do assessment design specifications include a description of processes used to ensure items and test events assess a range of cognitive demands as specified in CCR standards?

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 Writing Guidance

- Review guidelines and processes for the production of high-quality items consistently aligned to CCR standards.
- Review the assessment's stance toward cognitive complexity in assessment and assessment design.
- Review assessment's means for measuring cognitive complexity relating to item development.
- Review processes for ensuring a range of CCR standards are assessed.
- Identify level of specificity in item writer training materials.
- Review technical processes used for item review and approval.
- Review documentation regarding the testing or piloting of items on assessments.

Scoring Information and Rubric Development

- Review scoring guides for level of detail in point values attributable to individual items.
- 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.

Content and Technical Quality

- 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).

Cognitive Demand

- Review assessment design specifications for cognitive demand information.

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 Documentation
- Does item development documentation provide enough guidance and support for the creation of an item bank that ensures alignment to the full range and intent of targeted CCR standards?
- Are the item writing materials robust enough to support the development of high-quality selected response items as prescribed by the targeted standards?
    - Are descriptions/samples of procedural skill, conceptual understanding, and application items included?
    - Are clear guidelines for the construction of items including mathematical practices included?
    - Are sample items incorporating CCR standards provided to illustrate the design of a range of cognitive complexity and item formats included?
    - 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?
- If necessary, are the item writing materials robust enough to support the development of technology-enhanced stems requiring relevant evidence of mathematical skills as prescribed by targeted CCR standards in the response?
    - Are guidelines for the construction of technology-enhanced stems clear and supported with examples that incorporate CCR standards?
    - Are item writer checklists included?
- Are the item writing materials robust enough to support the development of an item bank that can assess the full intent of the mathematical standards?
- Is the review process thorough enough to ensure high-quality test items on test events?
- What materials are weak or missing, if any?

Scoring Information and Rubric Design
- Do scoring matrices, guides, and/or exemplars note the point-value attributed to responses and aid in the scoring of student responses?
- 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?

Content and Technical Accuracy
- Does the documentation provide enough information to support content accuracy in regard to item development and review?
- Does the documentation provide enough information to support technical quality, alignment to standards, and editorial accuracy of the items?
- If sub-score analyses are reported, does the documentation provide enough information to support the technical quality of the reported sub-score(s)?

Cognitive Demand
- Do the assessment design specifications provide enough information to ensure the items and test events measure the cognitive demand required by the targeted CCR standards?

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 focus strongly on the content that is most important for students to master by reflecting an appropriate expected distribution of content and related score points. ● The expected distribution of score points across the K-8 content blueprints and/or assessment design specifications focuses strongly on the major work as established in CCR standards. ● The expected distribution of score points across the HS course content blueprints and/or assessment design specifications focuses on the content and skills students need to be successful in college and careers as established in CCR standards.

Scoring		
4 points Materials meet expectations of this indicator. ● The blueprints, item-selection criteria, and/or assessment design specifications assure that points on the assessments given will focus at least 65% of score points on the major work in elementary or middle grades or, 50% for high school, on the widely applicable prerequisites for college and career success.	2 points Materials partially meet expectations of this indicator. ● The blueprints or item-selection criteria establish that between 50-64% of score points on the assessment focus on the major work in elementary or middle grades or, between 35-49% for high school, on the widely applicable prerequisites for college and career success. OR ● Assessment design specifications and other evidence, such as item-pool distributions, analyses of the population of test events administered, technical reports including content distribution, etc., prioritize that at least 50% of the items on the assessment will assess the major work in elementary or middle grades or, at least 35% for high school, on the widely applicable	0 points Materials DO NOT meet expectations of this indicator. ● The blueprints, item-selection criteria, and/or assessment design specifications do not establish that points on the assessments given will focus fewer than 50% of score points on the major work in elementary or middle grades or, for fewer than 35% for high school, on the widely applicable prerequisites for college and career success.

	prerequisites for college and career success.	

About this indicator:

What is the purpose of this Indicator?

Assessment design specifications ensure assessments yield information on the most important topics and ensure the assessment matches math expectations. The purpose of this indicator is to evaluate whether the majority of expected score points across all administrations focus on the content that is most important for students to master in order to reach college and career readiness.

For each grade band, this content consists of:

- Elementary grades – number and operations;
- Middle school – ratio, proportional relationships, pre-algebra, and algebra (See Appendix 1.1 for specific clusters); and
- High school – prerequisites for careers and a wide range of postsecondary studies, particularly algebra, functions, statistics, and modeling applications. (See Appendix 1.2 for specific topics).

Resources:

- CCSSO Criteria for High Quality Assessments
  - C.1
- Appendix 1.1 Major Clusters of the Grade
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
  - I. Focus, Coherence, and Rigor in the Common Core State Standards for Mathematics
  - 1. Focus on Major Work
  - 2. Focus in Early Grades
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
  - I. Focus, Coherence, and Rigor in the High School Standards
  - 1. Focus on Widely Applicable Prerequisites

Indicator 1.1.b Guiding Questions:

- Do the majority of the total score points on the assessment blueprints, item-selection criteria, or assessment design specifications focus strongly on the content most needed for success in later mathematics?
- For an assessment without blueprints or item-selection criteria, does other evidence provided (item-pool distributions, analyses of the population of test events administered, technical reports including content distribution, etc.) prioritize that the assessment will focus strongly on the content most needed for success in later mathematics?

Evidence Collection

Assessments with Blueprints

- Find test blueprints for assessments. (If an assessment uses the same blueprint for each test event, then we will evaluate the blueprint used across these test events.)
- Look for assessment blueprint point totals or percentages of points.

- Use test blueprints to total the points by content category. Calculate the points that measure the major work and the total number of points, then find the percentage of major work. Compare these percentages to the percentages given below:
  - 65% or more of score points in elementary grades align to major work.
  - 65% or more of score points in middle-school grades align to major work.
  - 50% or more of score points in high-school grades align to prerequisites for careers and a wide range of postsecondary studies.
- Report out the score points aligned to major work for the given blueprint.

Assessments with Item-Selection Criteria

- Search for item-selection algorithms for a computer adaptive test (CAT).
- Examine CAT algorithm rules that relate to item selection and content coverage.
- Evaluate the item-selection algorithm for rules that require certain content alignments and how these rules are ranked against other rules that guide item selection.
  - Determine if these rules assure a test event in which 65% or more of the score points align to the major work.
  - Report out the score points aligned to major work for the test events.

Assessments without Blueprints or Item-Selection Criteria

- If the algorithm does not have requirements around content alignment, refer to the other evidence provided to determine the percentage of items aligned to major work.
  - Look at distributions of content across the entire item pool.
  - Look at a large number of test events administered and the populations they are administered to.
- Calculate the items in the pool that measure the major work and the total number of items in the pool, then find the percentage of major work. Compare these percentages to the percentages given below:
  - 50% or more of items in elementary grades align to major work.
  - 50% or more of items in middle-school grades align to major work.
  - 35% or more of items in high-school grades align to prerequisites for careers and a wide range of postsecondary studies.
- Report out the item pool content distribution for items as they align to major work.
- Use studies of historical test events administered to analyze the student populations receiving those events to determine what percentage of students receive test events focused on major work.

Cluster Meeting Discussion

- Does the documentation provide enough information to ensure the expected distribution of score points assures students will primarily be assessed on their skills and knowledge in the major work?
- Do test blueprints, item-selection criteria, or assessment design specifications ensure the focus on the major work?
- If there are no test blueprints or item-selection criteria, to what extent does item pool distribution prioritize a focus on major work?
  - Do historical test events show a focus on the major work?
  - What percentage of test events show a focus on the major work?
- How do these percentages fall into the guidelines mentioned above?
  - At least 65% of the expected score points in test events are focused on the major work for elementary and middle school grades.
  - At least 50% of the expected points in each course align exclusively to prerequisites for careers and a wide range of postsecondary studies.

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2

Item and Form Alignment

Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.

What is the purpose of this Criterion?

Quality mathematics interim assessments reflect mathematics as a coherent topic focused on major topics that assess conceptual understanding, procedural skills, fluencies, applications, and mathematical practices. The development of these assessment systems require a strategic plan for the design of the assessment, including test blueprints or assessment design specifications which focus on the most important topics for college- and career-readiness as described in Criterion 1.1. These plans are used to create items, items banks, and forms that reflect this focus while assessing the full range of mathematics described by the college- and career-ready standards as described in Criterion 1.2.

Research Connection

- Common Core State Standards for Mathematics (CCSSM)
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments
- 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.
- 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.
- Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal Design for Learning: Theory and Practice*. Wakefield, MA: CAST. pp. 73-76.

Scoring:

Meets Expectations	Partially Meets Expectations	Does Not Meet Expectations
• 13-16 points	• 8-12 points	• <8 points

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.
Indicator 1.2.a	Test forms/events delivered to students reflect an appropriate distribution of content and related score points and item types within forms/events. ● The test forms/events delivered to students focus strongly on the major and supporting clusters as established in CCSS standards. ● The test forms/events delivered to students focus on the content and skills students need to be successful in college and careers as established in CCR standards.

Scoring		
4 points Materials meet expectations of this indicator. ● For all test forms/events reviewed, at least 65%[1] of score points are focused on the major clusters for elementary and middle school grades, or 50% on widely applicable prerequisites in high school.	2 points Materials partially meet expectations of this indicator. ● For all test forms/events reviewed, between 50-64% of score points are focused on the major clusters for elementary and middle school grades, or between 35-49% of widely applicable prerequisites in high school.	0 points Materials DO NOT meet expectations of this indicator. ● For all test forms/events reviewed, less than 50% of score points are focused on the major clusters for elementary and middle school grades, or less than 35% of widely applicable prerequisites in high school.

About this indicator:

What is the purpose of this Indicator?

Each set of test forms or events must be focused on the most important content to generate college- and career-readiness-aligned assessment data. The purpose of this indicator is to evaluate whether the vast majority of score points among interim assessment test events focus on the content that is most important for students to master in order to reach college and career readiness.

This content consists of the following major and supporting clusters:

- Kindergarten - K.CC.A, K.CC.B, K.CC.C, K.OA.A, K.NBT.A, K.MD.B, K.G.B
- Grade 1 - 1.OA.A, 1.OA.B, 1.OA.C, 1.OA.D, 1.NBT.A, 1.NBT.B, 1.NBT.C, 1.MD.A, 1.MD.C
- Grade 2 - 2.OA.A, 2.OA.B, 2.OA.C, 2.NBT.A, 2.NBT.B, 2.MD.A, 2.MD.B, 2.MD.C, 2.MD.D
- Grade 3 - 3.OA.A, 3.OA.B, 3.OA.C, 3.OA.D, 3.NF.A, 3.MD.A, 3.MD.B, 2.MD.C, 3.G.A
- Grade 4 - 4.OA.A, 4.OA.B, 4.NBT.A, 4.NBT.B, 4.NF.A, 4.NF.B, 4.NF.C, 4.MD.A, 4.MD.B
- Grade 5 - 5.NBT.A, 5.NBT.B, 5.NF.A, 5.NF.B, 5.MD.A, 5.MD.B, 5.MD.C

[1] If the percent of score points is close to 65%, items in supporting clusters will be taken into account. See Evidence Collection below.

- Grade 6 - 6.RP.A, 6.NS.A, 6.NS.C, 6.EE.A, 6.EE.B, 6.EE.C, 6.G.A
- Grade 7 - 7.RP.A, 7.NS.A, 7.EE.A, 7.EE.B, 7.SP.A, 7.SP.C
- Grade 8 - 8.NS.A, 8.EE.A, 8.EE.B, 8.EE.C, 8.F.A, 8.F.B, 8.G.A, 8.G.B, 8.SP.A
- High school – content from CCSSM widely applicable as prerequisites for a range of college majors, postsecondary programs, and careers, particularly algebra, functions, statistics, and modeling applications. (See Appendix 1.2 for specific topics).

Resources:

- CCSSO Criteria for High Quality Assessments
  - C.1
- Appendix 1.1 Major Clusters of the Grade
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
  - I. Focus, Coherence, and Rigor in the Common Core State Standards for Mathematics
  - 1. Focus on Major Work
  - 2. Focus in Early Grades
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
  - I. Focus, Coherence, and Rigor in the High School Standards
  - 1. Focus on Widely Applicable Prerequisites

Indicator 1.2.a Guiding Question:

Do the majority of the score points for the assessment focus strongly on the content most needed for success in later mathematics?

Evidence Collection

- Record the alignments of the items from the test forms/events.
- Create a table to show the percentage of points, among all test forms/events, that focus on the major and supporting clusters, or widely-applicable prerequisites for college and careers.
- If provided, evaluate data tables that summarize historical test forms/events that show the alignments for items given in those test events.
- Compare these percentages to the percentages in the scoring bullets,
  - 65% or more of score points align to major clusters for elementary and middle school grades
    - If the percent of score points is close to 65%, add the score points for supporting work clusters to the major work total to determine if this increases the percent to above 65%.
  - 50% or more of score points in high-school grades align to widely applicable as prerequisites for careers and a wide range of postsecondary studies.

Cluster Meeting Discussion

- What percentage of the score points align to major clusters?
- Do the percentages fall into guidelines mentioned above?
  - At least 65% of score points in test events is focused on major clusters for elementary and middle school grades.
    - Is the percent close to 65%? Should the score points from supporting clusters be added to the total score points for the major clusters?
  - In high school, at least 50% or more of score points align to widely applicable as prerequisites for careers and a wide range of postsecondary studies.

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards 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. ● 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. ● Test items align to the assessment design specifications. ● Items are content-accurate, reflecting no technical or editorial flaws.

Scoring		
2 points Materials meet all expectations of this indicator. ● At least 80% of the test items can be aligned to exclusively measure CCR standards. ● Most assessment items align to the assessment design specifications. ● At least 97% of the test items contain no content flaws and are technically and editorially accurate.	1 point Materials partially meet expectations of this indicator. ● Between 50-79% of the test items can be aligned to exclusively measure CCR standards. ● There is partial alignment between the actual assessment items and the assessment design specifications. ● At least 95% of the test items contain no content flaws, are technically and editorially accurate.	0 points Materials DO NOT meet expectations of this indicator. ● Fewer than 50% of the test items can be aligned to exclusively measure CCR standards. ● There is little to no alignment between the actual assessment items and the assessment design specifications. ● Numerous inaccuracies impede the demonstration of knowledge and skills.

About this indicator:

What is the purpose of this Indicator?

Items must be well aligned and of high quality to ensure they align with the expectations of the college- and career-ready standards and generate data matching the assessment's purpose. This indicator evaluates the degree of alignment between the 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

    - C.5 Ensuring high-quality items and a variety of item types
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
    - 2. Focus in Early Grades
    - 3. Focus and Coherence
    - 5. Consistent Progressions
    - 6. Coherent Connections
    - Indicators of Quality in instructional materials and tools for mathematics
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
    - 2. Rigor and Balance
    - 3. Consistent Content
    - 4. Coherent Connections

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 with CCR standards?
- Do test items avoid measuring knowledge and skills expected in subsequent years of 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.
- Look for items intended to elicit responses not reflective of the CCR standards (e.g., items not reflective of CCR standards, items measuring content knowledge outside of the scope, implied or stated, of the standards).

Alignment to Assessment Design Specifications

- Note alignment between assessment design specifications and the items.

Accuracy

- Look for inaccuracies in how the items are presented, including precision in the application of mathematical terms, number concepts, 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

- Do the 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 CCR standards?
- Does the descriptive metadata (if provided) associated with items, correlate to the expectations of the associated CCR standards?
- Do most of the items that assess supporting standards/clusters follow the intent of the standard/cluster?
- What connections between supporting and major standards/clusters were used to tightly align to the standards within supporting work?

- Do item stems limit their questioning to CCR standards-related skills and concepts?
- Do assessment keys and rubrics evaluate item responses using criteria that are limited to standards-related skills and concepts?

Alignment to Assessment Design Specifications
- Do the items reflect the intentions of the assessment design specifications?

Accuracy
- In relation to content, are the items accurate with no flaws?
- In relation to technical accuracy, are the items accurate with no flaws?
- In relation to editorial accuracy, are the items accurate with no flaws?

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.
Indicator 1.2.c	The range of item types and cognitive demand among test events is sufficient to strategically assess the full intent and complexity of CCR standards being addressed and is aligned to blueprints or assessment design specifications. ● The items are constructed to reach the depth and complexity of CCR standards and cover the expected range of content measured by any assessment sub-scores. ● There is an appropriate distribution and/or range of cognitive demand exercised among test events submitted for review. ● The range of item types and cognitive demand among test events align to blueprints or assessment design specifications.

Scoring		
2 points Materials meet expectations of this indicator. ● Overall, 80% of CCR standards targeted by the assessment are fully measured among test events to the totality of the standards' expectations. ● For the most part, the range of item types and cognitive demand among test events reflects the design proposed in the blueprints or assessment design specifications. ● There is evidence of consistent processes to verify an appropriate range of cognitive demand among test events and/or standards being assessed. ● Considering all test events reviewed, item types are strategically used to measure the targeted standards.	1 point Materials partially meet expectations of this indicator. ● Overall, between 50-79% of CCR standards targeted by the assessment are fully measured among test events to the totality of the standards' expectations. ● The range of item types and cognitive demand among test events partially reflects the design proposed in the blueprints or assessment design specifications. ● There is evidence of processes to verify an appropriate range of cognitive demand among test events and/or standards being assessed. ● Considering all test events reviewed, item types are usually strategically used to measure the targeted standards.	0 points Materials DO NOT meet expectations of this indicator. ● Overall, fewer than 50% of CCR standards targeted by the assessment are fully measured among test events to the totality of the standards' expectations. ● The range of item types and cognitive demand among test events rarely reflects the design proposed in the blueprints or assessment design specifications. ● There may or may not be evidence of processes to verify an appropriate range of cognitive demand among test events and/or standards being assessed. ● Considering all test events reviewed, item types are usually strategically used less than 50% of the time to measure the targeted standards.

About this indicator:

What is the purpose of this Indicator?

College- and career-ready standards include a range of depth and complexity, and aligned assessments must cover this depth and range. Assessment items should fully align to targeted standards using item formats that effectively measure the content. Full alignment requires items included in test events to fully measure the targeted standards in content and context, in word and in complexity to ensure an accurate measure of proficiency level at the appropriate level of cognitive demand. For example, assessment items targeting standards requiring procedural skills must target the procedure. Items targeting standards stating applications of operations must provide culturally-responsive real-world contexts that lend themselves to solutions using the specified operations.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - C.3. Requiring a range of cognitive demand
  - C.5 Ensuring high-quality items and a variety of item types
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments

Indicator 1.2.c Guiding Questions:

- Are items constructed to reach the depth and complexity of standards expressing multiple cognitive goals (e.g., determine and analyze; describe and explain; represent and model)?
- Do students answer questions through the assessment(s) that include the important foundations in CCR standards as well as application of these skills and concepts?
- 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?

Evidence Collection

Depth and Complexity of the Standards

- Record test experience related to item types, 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 on test events. These items may include but are not limited to the following:
  - Multiple choice, multiple select, technology enhanced, technology enhanced constructed response, constructed extended response, short answer, performance tasks, and other innovative item types.
- Identify items that require test-takers to explain responses and/or or provide evidence.
- Identify items asking students to construct arguments, create representations, or perform non-routine procedures.
- 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 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]).

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 the standards assessed?

Distribution of Cognitive Demand

- Is given metadata about cognitive demand accurate?
- 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?
- If the assessment measures sub-scores, are the items tagged for measuring sub-scores representative of content described in the assessment purpose?

Range of Items Matches Documentation

- What variety of item types are represented in the test events?
- Do the item types used match the types of skills or understandings assessed?
- Are item types appropriate and sufficient to meet the demands of all targeted standards?
- What processes are documented to provide verification of the levels of cognitive demand assigned to items?

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.
Indicator 1.2.d	The assessment is aligned to the procedural skill and fluency expectations of CCR standards. • The item development documentation and distributions of points from the assessment design specifications that directly address standards requiring procedural skill and fluency are reflected in the assessment items and resulting forms/test events.

Scoring		
2 points Materials meet expectations of this indicator. • Over 80% of the set of items measuring standards attending to procedural skill and fluency represent a range of procedural skills and fluencies as expected by those standards. • Items measuring procedural skills balance friendly and unfriendly numbers along with unusually complex execution of the procedures. • Most items measuring procedural skill and fluency standards measure procedural skills and fluency exclusively, not applications or conceptual understandings.	1 point Materials partially meet expectations of this indicator. • Between 50-79% of the set of items measuring standards attending to procedural skill and fluency represent a range of procedural skills and fluencies as expected by those standards. • Items measuring procedural skills provide some balance of friendly and unfriendly numbers along with unusually complex execution of the procedures. • Fewer than 50% of items measuring procedural skill and fluency measure procedural skills and fluency exclusively, not applications or conceptual understandings.	0 points Materials DO NOT meet expectations of this indicator. • Fewer than 50% of the set of items measuring standards attending to procedural skill and fluency represent a range of procedural skills and fluencies as expected by those standards. • Items measuring procedural skills do not balance friendly and unfriendly numbers along with unusually complex execution of the procedures. • Fewer than 25% of items measuring procedural skill and fluency measure procedural skills and fluency exclusively, not applications or conceptual understandings.

About this indicator:

What is the purpose of this Indicator?

Procedural skills and fluencies are explicitly measured in college- and career-ready standards, and should be directly assessed. Aligned assessments measure whether students have the procedural skills and fluencies they will need to solve problems. Direct assessment of these procedural skills and fluencies is important to ensure that this content, as outlined in CCR standards, is measured. Students must gain these skills as reflected in the progression of CCR standards to ensure they have the tools necessary to solve problems and understand future procedural skills.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - C.2 Assessing a balance of concepts, procedures, and applications
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
  - 3. Focus and Coherence
  - 4. Rigor and Balance
  - 5. Consistent Progressions
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
  - 2. Rigor and Balance
  - 3. Consistent Content
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments

Indicator 1.2.d Guiding Questions:

- As specified in CCR standards, are procedural skills and fluencies directly measured?
- Do items among test events measure a range of common and unconventional procedures?

Evidence Collection

- Review the items measuring standards that attend to procedural skills and fluencies to ensure they only require expectations of CCR standards.
- Review the items measuring standards that attend to procedural skills and fluencies to determine whether they represent a range from common procedural skills to unconventional procedures that are unique or unfamiliar.
- Look at metadata for the items to ensure most items aligned to standards requiring procedural skills and fluencies actually measure procedural skills or fluencies.
- Consider whether assessment design specifications ensure sets of assessments directly measure an appropriate range of procedural skills and fluencies.

Cluster Meeting Discussion

- Are there items which directly measure standards that attend to procedural skills and fluencies?
- How well do these items measure a range of complexities from straightforward calculations to non-routine procedures?
- How well do some items include friendly numbers while others include unfriendly numbers or unusually complex execution of the procedures?
- What percentage of items aligned to standards measuring procedural skill and fluency only measure the indicated procedural skill or fluencies?
- How well do the items among all test events evaluated measure the full range of procedural skills and fluencies expected by CCR standards?

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.
Indicator 1.2.e	The assessment is aligned to the conceptual understanding expectations of CCR standards. • The item development documentation and distributions of points from the assessment design specifications that directly address standards requiring conceptual understanding are reflected in the assessment items and resulting forms/test events.

Scoring		
2 points Materials meet expectations of this indicator. • Over 80% of the set of items measuring standards attending to conceptual understanding represent a range of conceptual understandings as expected by those standards. • Items measuring conceptual understanding provide a range of complexity from creating representations to sophisticated chains of reasoning. • Items aligned to conceptual understanding standards do not measure simple recall.	1 point Materials partially meet expectations of this indicator. • Between 50-79% of the set of items measuring standards attending to conceptual understanding represent a range of conceptual understandings as expected by those standards. • Items measuring conceptual understanding provide some range of complexity from creating representations to sophisticated chains of reasoning. • Less than 20% of items evaluated directly measuring conceptual understandings are simple recall.	0 points Materials DO NOT meet expectations of this indicator. • Fewer than 50% of the set of items measuring standards attending to conceptual understanding standards represent a range of conceptual understandings as expected by those standards. • Items measuring conceptual understanding do not provide a range of complexity from creating representations to sophisticated chains of reasoning. • More than 20% of items evaluated directly measuring conceptual understandings are simple recall.

About this indicator:

What is the purpose of this Indicator?

College- and career-ready standards require conceptual understanding of key mathematical concepts that build across grades and courses. Aligned assessments measure whether students have conceptual understanding as specified in CCR standards. Direct assessment of these conceptual understandings using representations, reasoning, justification, and connections to procedures and applications are required to ensure students have a firm background in the content standards. Conceptual understanding does not mean simple recall of definitions; it requires students to integrate ideas of multiple grade-level concepts, create representations of the concept,

and engage in reasoning and/or justification with the concept. These concepts are aligned in progressions to ensure students have the foundational understandings necessary to be successful with the new conceptual understandings, procedures, and applications.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - C.2 Assessing a balance of concepts, procedures, and applications
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
  - 3. Focus and Coherence
  - 4. Rigor and Balance
  - 5. Consistent Progressions
  - 10. Emphasis on Mathematical Reasoning
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
  - 2. Rigor and Balance
  - 3. Consistent Content
  - 8. Emphasis on Mathematical Reasoning
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments

Indicator 1.2.e Guiding Questions:

- As specified in the CCR standards, are conceptual understandings measured in items?
- Do items measure understanding of concepts including evidence of reasoning, justification, planning, analysis, judgment?
- Do items among test events measure understanding using a range of items, some that directly measure understanding of specific concepts while others integrate multiple concepts or require sophisticated lines of reasoning?

Evidence Collection

- Review the items measuring standards attending to conceptual understanding to ensure they directly assess whether students understand the concept.
- Review the items measuring standards attending to conceptual understanding to determine whether they represent a range from relating concepts and creating representations to solving problems integrating multiple concepts or requiring sophisticated lines of reasoning.
- Review the items aligned to standards attending to conceptual understanding to ensure they are not simple recall items. Items measuring conceptual understanding should at least relate CCR-aligned concepts or connect concepts with procedures, applications, and reasoning.
- Look at metadata for the items to ensure items aligned to standards requiring conceptual understanding actually measure conceptual understanding.
- Consider whether assessment design specifications ensure sets of assessments directly measure an appropriate range of conceptual understandings.

Cluster Meeting Discussion

- Are there items which directly measure conceptual understandings?
- How often are items aligned to standards attending to conceptual understanding simple recall items that do not require students to use or reason with the concept?
- How well do these items measure a range of complexities from creating representations to sophisticated chains of reasoning?

- How well do items aligned to conceptual understandings measure the indicated conceptual understandings?
- How well do the items among all test events evaluated measure the full range of conceptual understandings expected by CCR standards?

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.
Indicator 1.2.f	The assessment is aligned to the application expectations of CCR standards. • The item development documentation and distributions of points from the assessment design specifications that directly address standards requiring application are reflected in the assessment items and resulting forms/test events. • For high school assessments, items that attend to the full intent of the modeling process are administered to all students.

Scoring		
2 points Materials meet expectations of this indicator. • Over 70% of the set of items measuring standards attending to application represent a range of application items as expected by those standards. • Items measuring application provide a range of complexity from interpreting and solving problems to formulating solution strategies and, in middle and high school, parts of the modeling process. • Items aligned to application standards do not make the required mathematics obvious.	1 point Materials partially meet expectations of this indicator. • Between 50-69% of the set of items measuring standards attending to application represent a range of application items as expected by those standards. • Items measuring application provide some range of complexity from interpreting and solving problems to formulating solution strategies and, in middle and high school, parts of the modeling process. • Less than 20% of application items evaluated make the mathematics obvious.	0 points Materials DO NOT meet expectations of this indicator. • Fewer than 50% of the set of items measuring standards attending to application represent a range of application items as expected by those standards. • Items measuring application do not provide a range of complexity from interpreting and solving problems to formulating solution strategies and, in middle and high school, parts of the modeling process. • More than 20% of application items evaluated make the mathematics obvious.

About this indicator:

What is the purpose of this Indicator?

Aligned CCR assessments provide a wide variety of measures for applying CCR-aligned mathematics. Application items require students to interpret a situation, determine the mathematics applicable given this information, and use one or more solution steps to determine a solution and finally to precisely communicate justification for that solution. Application items do not directly indicate or make plainly obvious what mathematics is applicable. In the middle and high school grades, students use steps within the modeling process to work with culturally-responsive real-world situations and data to determine relevant solutions.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - C.2 Assessing a balance of concepts, procedures, and applications
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
  - 3. Focus and Coherence
  - 4. Rigor and Balance
  - 5. Consistent Progressions
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
  - 2. Rigor and Balance
  - 3. Consistent Content
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments

Indicator 1.2.f Guiding Questions:

- As specified in the CCR standards, are applications of content directly measured?
- Do items measuring applications require students to interpret a context and determine the procedure(s)/concept(s) relevant to determining a solution?
- Do items among test events measure application using a range of items, some that require interpretation and one or more solution steps while others require formulating, interpreting and modeling processes?

Evidence Collection

- Review the items measuring application to ensure they directly assess whether students can interpret a situation, determine the relevant mathematics, and use those procedures/concepts to determine a solution.
- Review the items measuring application to determine whether they represent a range from interpreting a context that is not immediately obvious to solving problems requiring interpretation, formulation, computation, and parts of the modeling process.
- Review the items aligned to standards requiring application to ensure they do not make the mathematics obvious. Items measuring application must require students to interpret a situation and determine the mathematics applicable to the situation.
- Look at metadata for the items to ensure items aligned to standards requiring application actually measure applications.
- Consider whether assessment design specifications ensure an assessment system directly measures an appropriate range of applications.

Cluster Meeting Discussion

- Are there items which directly measure application (interpreting and solving)?
- How often do items aligned to standards requiring application make the required mathematics obvious?
- How well do these items measure a range of complexities from interpreting and solving problems to complex situations requiring interpretation, formulation, computation, and parts of the modeling process (modeling process only applicable in middle and high school)?
- How well do items aligned to standards requiring application measure the indicated mathematics?
- How well do the items among all test events evaluated measure the full range of applications expected by CCR standards?

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the mathematics standards as outlined by college- and career-ready (CCR) standards.
Indicator 1.2.g	The assessment includes mathematical practices as described in CCR standards. • The assessment design specifications addressing mathematical practices are reflected in the assessment items and resulting forms/test events. • Items including mathematical practices should be connected to CCR-aligned mathematics. • If alignments are provided, items aligned to targeted mathematical practice(s) require them to receive full credit. • Test forms/events reflect the distribution of mathematical practices as outlined in the assessment design specifications.

Scoring		
2 points Materials meet expectations of this indicator. • Targeted mathematical practices are thoughtfully integrated into the system of assessment. • 33% or more of item score points among test events require mathematical practices. • Items requiring the targeted mathematical practices are seen by students among test events. • Items aligned to mathematical practices also align to mathematical content standards. • Items aligned to mathematical practices require students to engage in the practice. • For the most part, the targeted mathematical practices among test events reflects the design proposed in the assessment design specifications.	1 point Materials partially meet expectations of this indicator. • Targeted mathematical practices are tacked onto the end of assessments or not explicitly integrated into the system of assessment. • 20-33% of item score points among test events require mathematical practices. • Items requiring most targeted mathematical practices are seen by students among test events. • Most items aligned to mathematical practices also align to mathematical content standards. • Items aligned to mathematical practices require students to engage in the practice more than 80% of the time. • The range of targeted mathematical practices among test events partially reflects the design proposed in the	0 points Materials DO NOT meet expectations of this indicator. • Targeted mathematical practices are not explicitly integrated into the system of assessment. • Less than 20% of item score points among test events require mathematical practices. • Most students do not see items requiring multiple targeted mathematical practices among test events. • Less than 50% items aligned to mathematical practices also align to mathematical content standards. • Items aligned to mathematical practices require students to engage in the practice less than 80% of the time. • The range of targeted mathematical practices among test events rarely reflects the design proposed in the

	assessment design specifications.	assessment design specifications.

About this indicator:

What is the purpose of this Indicator?

CCR standards require thoughtful integration of mathematical practices. Mathematical practices require students to persevere to find solutions and construct viable chains of reasoning. Assessments should provide students the opportunity to engage in each mathematical practice required by CCR standards.

Resources:

- CCSSO Criteria for High-Quality Assessments
  - C.3 Connecting practice to content
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
  - 7. Practice-Content Connections
  - 8. Focus and Coherence via the Practice Standards
  - 9. Careful Attention to Each Practice Standard
  - 10. Emphasis on Mathematical Reasoning
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
  - 5. Practice-Content Connections
  - 6. Focus and Coherence via the Practice Standards
  - 7. Careful Attention to Each Practice Standard
  - 8. Emphasis on Mathematical Reasoning
- Common Core State Standards for Mathematics
  - "Connecting the Standards for Mathematical Practice to the Standards for Mathematical Content" (p. 8)

Indicator 1.2.g Guiding Questions:

- As specified in the assessment design specifications, are mathematical practices assessed?
- Do items that measure mathematical practices also measure content standards?
- Do students have to engage in the targeted mathematical practices to receive full credit in items aligned to the targeted mathematical practices?
- Are mathematical practices represented among the test events?

Evidence Collection

- Review the assessment design specifications for information about how the mathematical practices are included in the assessment system.
- Review the items requiring students to connect mathematical practices with mathematical content. If items are not aligned to mathematical practices, then review all items for the presence of mathematical practices.
- Review the items requiring mathematical practices to ensure students must engage in the practice to earn full credit on each item.
- Review the test events to determine whether mathematical practices are represented as indicated in the assessment design specifications.
- Look at metadata for the items to ensure items aligned to mathematical practices actually require students to use the mathematical practices to earn full credit.

Cluster Meeting Discussion

- Within the test events, are there items which require mathematical practices?
- What percent of items reviewed among test events require mathematical practices?
- How well do items aligned to mathematical practices require students to engage in the practice to earn full points?
- How well do assessment design specifications ensure test events require mathematical practices for each student?
- What percent of items require mathematical practices according to the assessment design specifications?
- Are the mathematical practices integrated into items to reinforce key aspects of the content (e.g., students use structure and repeated reasoning to make sense of operations in the base-ten system)?

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?

Quality mathematics interim assessments reflect mathematics as a coherent topic focused on major topics that assess conceptual understanding, procedural skills, fluencies, applications, and mathematical practices. 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 Mathematics (CCSSM)
- Council of Chief State School Officers (CSSO) Criteria for High-Quality Assessments
- K–8 Publishers' Criteria for the Common Core State Standards for Mathematics
- High School Publishers' Criteria for the Common Core State Standards for Mathematics
- Achieve Framework to Evaluate Cognitive Complexity in Mathematics Assessments
- 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.
- 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.
- Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal Design for Learning: Theory and Practice*. Wakefield, MA: CAST. pp. 73-76.

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. • Item development documentation/procedures clearly demonstrate adherence to the principles of universal design. • Item rendering specifications clearly reflect the principles of universal design. • Item review processes are designed to minimize construct-irrelevant variance. • Items go through a content bias/sensitivity review to make sure they are appropriate and fair for all relevant student groups. • 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 Materials meet expectations of this indicator. • Item development documentation/procedures clearly demonstrate adherence to principles of universal design. • Test and item rendering specifications clearly reflect the principles of universal design. • Item review processes are effectively designed to mitigate construct-irrelevant variance. • Items go through a content bias/sensitivity review to make sure they are appropriate and fair for all relevant student groups • Procedures are in place to evaluate the technical quality and appropriateness of items and test events for student subgroups and students using varied accommodations.	2 points Materials partially meet expectations of this indicator. • Item development documentation/procedures demonstrates adherence to the principles of universal design. • Test and item rendering specifications reflect the principles of universal design. • Item review processes attempt to mitigate construct-irrelevant variance. • Items go through a content bias/sensitivity review to make sure they are appropriate and fair for all relevant student groups • Procedures are in place to evaluate the technical quality and appropriateness of items and test events for student subgroups and students using varied accommodations.	0 points Materials DO NOT meet expectations of this indicator. • Item development guidelines make no reference to universal design or other means to ensure fairness in the test design. • Test and item rendering specifications reflect little or no concern for principles of fairness. • Item review processes give little to no attention to concerns for construct-irrelevant variance within the assessment. • 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. • 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:

- Do test development documentation 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?
- 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 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 practices?

Content Bias/Sensitivity Review

- Is the makeup of the bias/sensitivity committee(s) reasonably representative of the intended student population? Are all major student groups represented?
- Are guidelines for avoiding bias and sensitivity conflicts adequate to ensure items and 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 form)? 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. ● The test-taking population for which the assessment was/was not designed to support is clearly documented. ● The list of accommodations is aligned to the vendor's definition of the assessment's intended uses. ● The list of accommodations is sufficient to serve the needs of the full population of intended test takers. ● 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. ● Evidence is available that supports the quality and appropriateness of provided accommodations. ● The administration manual is clearly worded and supports teachers and other educational personnel in providing an appropriate testing experience for all students. ● Sample forms or released test items are available to stakeholders at each grade level.

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

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

About this indicator:

What is the purpose of this Indicator?

Assessment systems should accommodate all test-takers in the intended test taking 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 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 addressing 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. • Guidance is provided to support accessibility to the assessment system on a variety of platforms. • Auditory supports present stimuli and items in a natural voice and at a cadence that can be adjusted to accommodate the learner. • Overall visual design, including digital tools (e.g., dictionaries, calculators, number lines, 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 Materials meet expectations of this indicator. • Guidance is provided to support accessibility to the assessment system on a variety of platforms. • Auditory supports present stimuli and items in a natural voice and at a cadence that can be adjusted to accommodate the learner. • Overall visual design, including digital tools (e.g., dictionaries, calculators, number lines, and highlighters) enhances the test-taking experience, does not distract or clutter the digital workspace, and can be easily navigated by students.	2 points Materials partially meet expectations of this indicator. • Guidance is provided to support accessibility to the assessment system on more than one platform. • Auditory supports are present. • Digital tools (e.g., dictionaries, calculators, number lines, and highlighters) are included, however, they may distract or clutter the digital workspace.	0 points Materials DO NOT meet expectations of this indicator. • Vague or overly technical guidance is provided to support accessibility to the assessment system. • Auditory supports are not present to accommodate the learner. • 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, calculators, number lines, and highlighters), enhance the test-taking experience?
- Does the overall visual design, including digital tools (e.g., dictionaries, calculators, number lines, and highlighters), distract or clutter the digital workspace?
- Is the overall assessment design, including digital tools (e.g., dictionaries, calculators, number lines, 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?

There are four criteria associated with Gateway 2. 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. ● Item development, review, and piloting procedures and materials are designed to ensure all newly developed items meet technical quality standards. ● Assessment design specifications and test development and review procedures ensure test events meet content and statistical quality criteria.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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 test 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 were decisions 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. ● Item/test development and review procedures facilitate the reliability of test scores. ● Procedures for calculating and evaluating reliability are well-documented and appropriate. ● Obtained reliability indices and estimates of precision are at an appropriate level to support the use of results as intended.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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 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.

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 ability 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 assessed domain.
Indicator 2.1.c	Achievement scores support intended interpretations of student performance. • Evidence is provided to support the intended interpretations of student achievement. • Equating/linking procedures supporting the comparability of achievement scores and score-based inferences across events/administrations are described and reasonable. • 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. • 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 Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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 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 assessed domain.
Indicator 2.1.d	Achievement scores are appropriate for supporting their intended uses. ● The intended uses for the achievement scores are clearly and consistently articulated. ● There is sufficient theoretical and empirical evidence supporting the intended uses of achievement scores.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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?

There are four criteria associated with Gateway 2. 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* **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. • 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 intended use of the interim assessment does not invalidate or contradict its appropriateness for predicting performance on the intended criterion measure(s). • If an interim assessment was designed to predict performance on specific assessment (e.g., ACT, SAT) evidence supporting that claim is provided.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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 its 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* **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to be predictive.*	Predicted results are reliable. ● 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 they were designed to support (e.g., CCR). ● The predictions demonstrate sufficient reliability to support their intended uses.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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* **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). • The data and procedures used to establish and evaluate the predictive relationship for a given test-taking sample are documented and reasonable. • The procedures used to support intended interpretations are clearly articulated. • Studies support the appropriateness of the predicted result as a measure of future performance.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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 "85th 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 score

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.
- ✔ 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.

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

- ✔ 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* **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. • The intended uses for the predicted results are clearly and consistently articulated. • There is sufficient theoretical and empirical evidence to support the appropriateness of the intended uses of predicted results.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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?

There are four criteria associated with Gateway 2. 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* **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. • The assessment design supports the reporting of sub-scores at each level of granularity for which they are provided • The assessment design supports interpretations of students' areas of strength and need in the content domain.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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 specifications 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* **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. ● 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. ● The calculated reliability and precision indices indicate adequate support for the intended interpretations and uses.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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* **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. ● Evidence is provided to support intended interpretations of all reported sub-scores. ● Empirical data suggest sub-scores represent distinct sub-domains and should be reported separately.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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* **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. • The intended uses for the sub-scores are clearly and consistently articulated. • There is sufficient theoretical and empirical evidence supporting the intended uses for the sub-scores.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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?

There are four criteria associated with Gateway 2. 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* **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. ● 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. ● The technical characteristics of the test and reportable scale support the reported growth measure.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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* **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. ● Procedures for estimating standard errors around the growth estimates are appropriate and well documented. ● The reliability of the growth scores have been evaluated for students at different places along the ability scale. ● The calculated reliability and precision indices indicate adequate support for the intended uses of the reported growth scores.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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* **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. • The procedures and measures for calculating student growth are clearly documented and appropriate. • 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. • Empirical evidence confirms that growth scores provide for valid intended inferences about student learning in the content domain.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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* **This indicator may be considered not claimed (N/C) if the assessment was not intentionally designed to support student progress.*	Student growth scores are appropriate for supporting the intended uses. ● The intended uses for the growth scores are clearly and consistently articulated. ● There is sufficient theoretical and empirical evidence supporting the intended uses for the growth scores.

Scoring		
2 points Meets expectations 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 Partially meets expectations 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 Does not meet expectations 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). ● Score reports effectively represent the intended interpretations and uses of overall achievement results. ● The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses. ● Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when needed. ● Evidence (e.g., studies, focus groups) is provided that users are able to effectively interpret and use reports in the manner intended. ● The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences. ● Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly). ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides.

Scoring		
4 points Materials meet expectations of this indicator. ● 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 Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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

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 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. ● For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability. ● 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 Materials meet expectations of this indicator. ● Sufficient, high quality evidence supports the range of expectations associated with information about the degree of error related to the achievement score.	1 point Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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. ● Guidance is aligned to the use. ● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data. ● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes.

Scoring		
4 points Materials meet expectations of this indicator. ● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses.	2 points Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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* **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). ● Score reports effectively represent the intended interpretations and uses of predicted student performance results. ● The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses. ● Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when applicable. ● Evidence is provided that users are able to effectively interpret and use reports in the manner intended. ● The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences. ● Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly). ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides.

Scoring		
4 points Materials meet expectations of this indicator. ● 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 Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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

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* **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. • For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability. • 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 Materials meet expectations of this indicator. • 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 Materials partially meet expectations of this indicator. • 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 Materials DO NOT meet expectations of this indicator. • 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* **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. ● Guidance is aligned to the use. ● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data. ● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes.

Scoring		
4 points Materials meet expectations of this indicator. ● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses.	2 points Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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* **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). ● Score reports effectively represent the intended interpretations and uses of sub-scores. ● The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses. ● Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when applicable. ● Evidence is provided that users are able to effectively interpret and use reports in the manner intended. ● The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences. ● Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly). ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides.

Scoring		
4 points Materials meet expectations of this indicator. ● 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 Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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

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 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* **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. ● For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability. ● 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 Materials meet expectations of this indicator. ● Sufficient, high quality evidence supports the range of expectations associated with information about the degree of error related to sub-scores.	1 point Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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* **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. ● Guidance is aligned to the use. ● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data. ● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes.

Scoring		
4 points Materials meet expectations of this indicator. ● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses.	2 points Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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* **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). ● Score reports effectively represent the intended interpretations and uses of student growth or progress results. ● The type and grain size of the information reported is appropriate for effectively serving the intended interpretations and uses. ● Evidence shows that there was attention to the audience and specific users in the design process, including user-specific versions of reports when applicable. ● Evidence is provided that users are able to effectively interpret and use reports in the manner intended. ● The documentation should include warnings of potential or common misuses of the results that may result in negative, unintended consequences. ● Reports identify and flag students for whom the integrity of the test interpretations may be compromised (e.g., student clicks through rapidly). ○ The conditions which bring about a flag are articulated on reports and/or in interpretive guides.

Scoring		
4 points Materials meet expectations of this indicator. ● 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 Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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

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 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* **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. • For example, confidence intervals, error bands, or probability statements are provided to represent potential score variability. • 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 Materials meet expectations of this indicator. • 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 Materials partially meet expectations of this indicator. • 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 Materials DO NOT meet expectations of this indicator. • 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* **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. ● Guidance is aligned to the use. ● Any guidance provided has a basis in research and/or was created in consultation with educators experienced in using educational data. ● Guidance is provided to support appropriate use for students scoring at the full range of performance outcomes.

Scoring		
4 points Materials meet expectations of this indicator. ● Sufficient, high quality evidence supports the range of expectations associated with the guidance provided to support the intended interpretations and uses.	2 points Materials partially meet expectations of this indicator. ● 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 Materials DO NOT meet expectations of this indicator. ● 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)?

Endnotes

Appendix 1.1: Major Clusters of the Grade

FOCUS COMPONENT 2: MAJOR CLUSTERS OF EACH GRADE					
QUALITY INDICATORS	MAJOR CLUSTERS	ADDITIONAL OR SUPPORTING CLUSTERS OR OTHER[17]	QUALITY INDICATORS	MAJOR CLUSTERS	ADDITIONAL OR SUPPORTING CLUSTERS OR OTHER[18]
Kindergarten	K.CC: A, B, C	K.MD: A, B	**Grade 5**	5.NBT: A, B	5.OA: A, B
	K.OA: A	K.G: A, B		5.NF: A, B	5.MD: A, B
	K.NBT: A			5.MD: C	5.G: A, B
Grade 1	1.OA: A, B, C, D	1.MD: B, C	**Grade 6**	6.RP: A	6.NS: B
	1.NBT: A, B, C	1.G: A		6.NS: A, C	6.G: A
	1.MD: A			6.EE: A, B, C	6.SP: A, B
Grade 2:	2.OA: A, B	2.OA: C	**Grade 7**	7.RP: A	7.G: A, B
	2.NBT: A, B	2.MD: C, D		7.NS: A	7.SP: A, B, C
	2.MD: A, B	2.G: A		7.EE: A, B	OTHER
Grade 3	3.OA: A, B, C, D	3.NBT: A	**Grade 8**	8.EE: A, B, C	8.NS: A
	3.NF: A	3.MD: B, D		8.F: A, B	8.G: C
	3.MD: A, C	3.G: A		8.G: A, B	8.SP: A
Grade 4	4.OA: A	4.OA: B, C			
	4.NBT: A, B	4.MD: A, B, C			
	4.NF: A, B, C	4.G: A			

Appendix 1.2

Table 1. Content From CCSSM Widely Applicable as Prerequisites for a Range of College Majors, Postsecondary Programs and Careers*

Number and Quantity	Algebra	Functions	Geometry	Statistics and Probability	Applying Key Takeaways from Grades 6–8**
N-RN, Real Numbers: Both clusters in this domain contain widely applicable prerequisites. **N-Q*, Quantities**: Every standard in this domain is a widely applicable prerequisite. Note, this domain is especially important in the high school content standards overall as a widely applicable prerequisite.	**Every domain in this category** contains widely applicable prerequisites.° Note, the **A-SSE** domain is especially important in the high school content standards overall as a widely applicable prerequisite.	**F-IF, Interpreting Functions**: Every cluster in this domain contains widely applicable prerequisites.° Additionally, standards **F-BF.1** and **F-LE.1** are relatively important within this category as widely applicable prerequisites.	**The following standards and clusters** are relatively important within this category as widely applicable prerequisites: **G-CO.1** **G-CO.9** **G-CO.10** **G-SRT.B** **G-SRT.C** Note, the above standards in turn have learning prerequisites within the Geometry category, including: G-CO.A G-CO.B G-SRT.A	**The following standards** are relatively important within this category as widely applicable prerequisites: **S-ID.2** **S-ID.7** **S-IC.1** Note, the above standards in turn have learning prerequisites within 6-8.SP.	**Solving problems at a level of sophistication appropriate to high school by:** • Applying ratios and proportional relationships. • Applying percentages and unit conversions, e.g., in the context of complicated measurement problems involving quantities with derived or compound units (such as mg/mL, kg/m^3, acre-feet, etc.). • Applying basic function concepts, e.g., by interpreting the features of a graph in the context of an applied problem. • Applying concepts and skills of geometric measurement e.g., when analyzing a diagram or schematic. • Applying concepts and skills of basic statistics and probability (see 6-8.SP). • Performing rational number arithmetic fluently.

From page 8 of the [High School Publishers' Criteria for the Common Core State Standards for Mathematics]