edreports.org

Center for Assessment

Guide for Implementation

Interim Assessment Reviews

Grades 3-8

Table of Contents

Overview

The purpose of this document is to outline the process used to evaluate evidence of alignment and technical quality within an interim assessment submitted for an EdReports review. This document provides a structural overview of the evaluation tool, identification of the entities involved in the review process, phases of the review process, and an appendix of document samples.

Structure of the Evaluation Process

The evaluation process includes three Gateways within which the criteria for review are organized. The categories of evidence considered in each of the three gateways are outlined in Figure 1.

- Evidence related to the alignment of the assessment to the expectations of the college- and career-ready standards and fairness and accessibility is evaluated in Gateway 1.
- The quality of evidence supporting the intended score interpretations and uses for each type of information provided on score reports (i.e., achievement, predictive, sub-scores, and growth) is evaluated in Gateway 2.
- Evidence related to score reporting interpretation and use in coordination with supplemental instructional materials and supports is evaluated in Gateway 3.

All three gateways work together to offer a comprehensive profile of the quality and sufficiency of evidence provided to support the interpretation and use of assessment results, as intended.

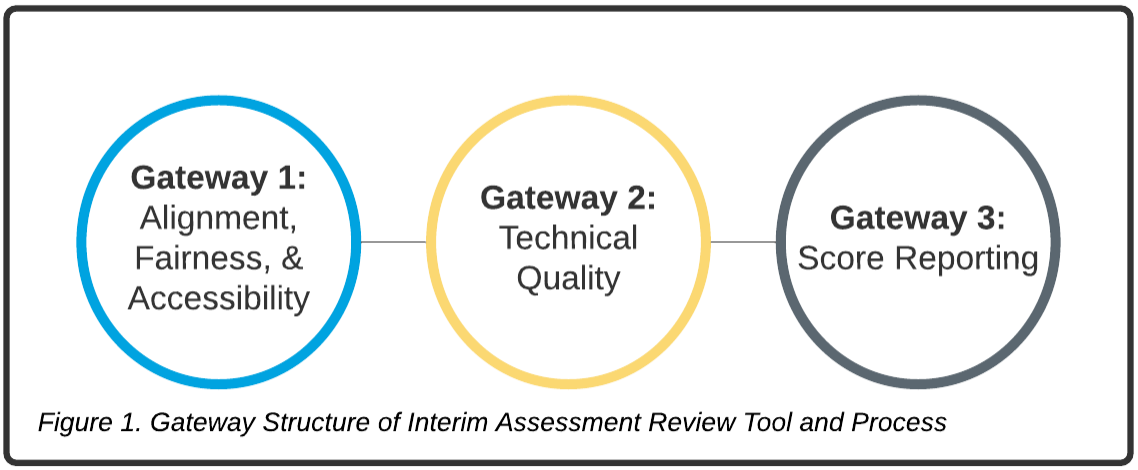
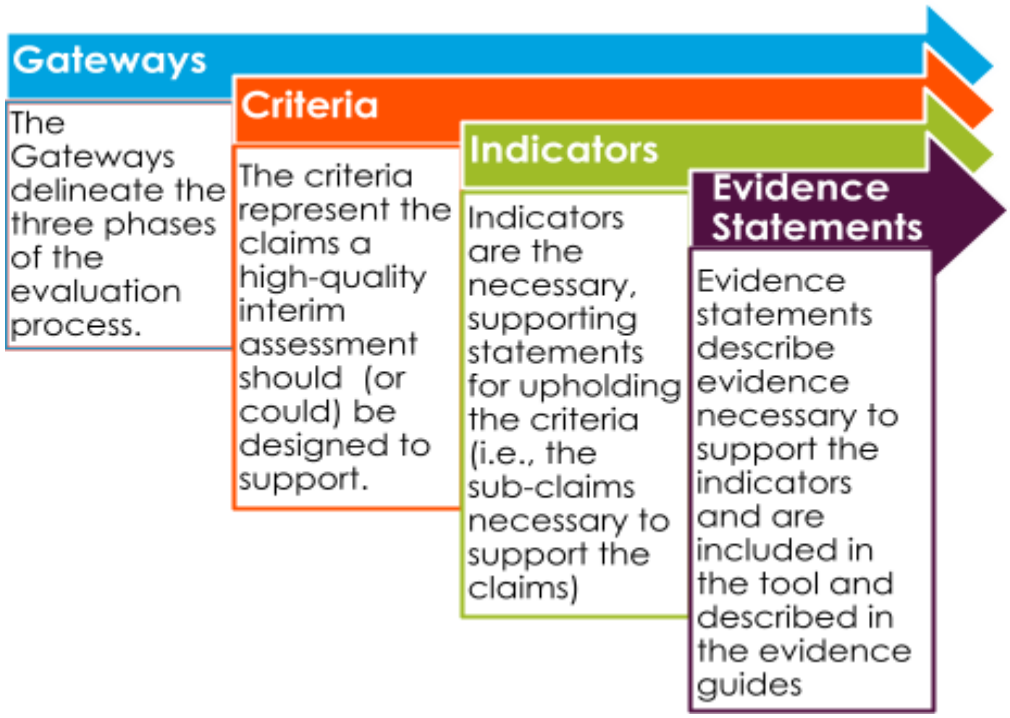


Figure 1. Gateway Structure of Interim Assessment Review Tool and Process

Within each Gateway is a distinct set of criteria, each further described through a series of indicators that more specifically support the general expectation of the identified criterion. Each indicator is further detailed (though not exhaustively itemized) through a series of evidence statements. Figure 2 provides an illustration of the gateway, criteria, indicator, evidence statement structured to support detailed and consistent reviews across interim assessments.

Figure 2. Hierarchy of quality criteria specification



Participants

Throughout this document, five entities are referenced, each with a clearly defined role in the overall review process. For clarity, each group and its expected role is outlined in Table 1.

Table 1. Groups Involved in the Alignment and Technical Evaluation Reviews

Group	Role
EdReports: *Staff* The organization requesting the review process. Responsible for organizing and facilitating the review of evidence of high-quality test forms and reports that demonstrate alignment to both college- and career-ready standards as well as nationally-recognized standards for educational measurement.	• Establishes initial contact with the assessment vendor. • Establishes all vendor contracts necessary to engage in evaluation. • Works with the vendor to establish requirements related to transmission, storage and return/purging of secure content and materials provided to support the review process. • Identifies and contracts with educators to serve as reviewers. • Coordinates and schedules the evaluation activities among the educator reviewers. • Works with the vendor to clarify evidence necessary to support all phases of the review process and verify it is organized in the manner specified. • Provides training to the educator reviewers on the review process, tool, evidence guides, and supporting documents and clarifies the format/location of provided materials. • Facilitates meetings with the educator reviewers with the goal of clarifying consensus feedback, ratings, and scoring. • Supervises calibration of reviewer teams across grade bands for the same assessment. • Supervises the production and editing of the reports written by the reviewer teams.

EdReports: *Educator Reviewers* Independent contractors with educational expertise hired and trained by EdReports to participate in the Interim Assessment Review Process.	• Participate in training facilitated by experts in content-area assessment and standards alignment to prepare for the EdReports review process for Gateways 1 and 3. • Engage in individual review of materials pertaining to the assigned indicators to gather evidence and assign personal ratings prior to the weekly team meeting. • Participate in 1 hour weekly team meetings to review evidence gathered by all reviewers and to reach consensus on scoring for each indicator. • Prepare draft reports to be calibrated by the team and reviewed by the EdReports staff prior to being finalized.
Center for Assessment: *Coordinators* Representatives responsible for organizing and facilitating the evaluation of evidence of technical quality (Gateway 2) and select criteria within Gateways 1 and 3.	• Identify and contact appropriate technical reviewers to act as evaluators • Coordinate and schedule the evaluation activities among the technical reviewers • Work with the vendor to clarify evidence necessary to support evaluation of technical quality and verify that it is organized in the manner specified. • Provide training to technical reviewers on the evaluation process, tool and clarify the format/location of provided materials • Facilitate meetings with technical reviewers with the goal of clarifying consensus feedback, ratings and recommendations. • Draft evaluation report based on technical reviewers' recommendations and revise as needed based on feedback.
Vendor: The primary group/organization responsible for assembling and organizing evidence for evaluation.	• Identify, gather and organize appropriate evidence to inform the evaluation process • Support the development of a general overview/summary of the submitted assessment consistent with the guidance provided in Appendix A. • Address questions posed by EdReports and the Center for Assessment and requests for clarification (as necessary). • Participate in the finalization of reports through the EdReports Errors and Omissions process • Provide EdReports with vendor comments in response to the final report (if desired). Note: these comments are published alongside the final reports on the EdReports website.
Technical Reviewers: The technical reviewers charged with reviewing and evaluating submitted evidence for Gateway 2 and select indicators in Gateways 1 & 3.	• Participate in training and evaluation activities • Review the evidence provided to support evaluation • Provide comments and ratings related to the adequacy of that evidence provided relative to expectations • Discuss thoughts and finding with peers with the goal of coming to consensus. • Comment on and approve the final report

Phases of the Review Process

EdReports and the Center for Assessment will work collaboratively in the review of assessments submitted for evaluation. The process of the review, though iterative in nature, can be summarized in five phases or steps of implementation (see Figure 3). Each phase of the process is detailed in the pages that follow.

Figure 3. Phases of the Review Process

P1: Preparation	• Selection of Review Teams • Collect Vendor Survey • Acquisition of Assessments and Accompanying Documentation for Review • Schedule vendor orientation meeting
P2: Vendor Orientation & Evaluation Training	• EdReports Training of the Educator Reviewer Team • Center for Assessment Training of the Technical Review Team • Vendor-Delivered Assessment Orientation

P3: Evaluation	• Educator Reviewer Evaluation Process
	• Technical Review Team Evaluation Process
	  ○ Independent review of evidence against evaluation criteria and indicators by technical review team members.
	  ○ Technical review team meets to discuss evidence and establish consensus indicator ratings and rationales.
	• Engage in resolution discussions, as needed, to establish consensus ratings for "shared" indicators (i.e. those reviewed by educators and technical reviewers).

P4: Report Generation & Approval	• Draft Executive Summary and Criterion-Level reports based on educators and technical reviewer comments. ○ Technical review team reviews the draft report, and suggests edits around text associated with reviewed criteria/indicators. • Update draft report based on provided comments/edits.
P5: Errors/ Omissions Process & Vendor Response	• Transmission of draft reports to vendors • Vendor Review & Submission of Counter Evidence • Deliberation of Review Teams • Transmission of finalized reports

Phase 1. Preparation

Selection of Review Teams

Educator Reviewers

Highly qualified and extensively trained educator review teams are the heart of EdReports reviews. Each interim assessment review team, comprised of four to five educator reviewers, focuses their attention on a single commercial assessment. Each team consists of a lead reviewer, a writer who synthesizes the findings of the overall review team, and up to three general reviewers.

- Team leads are EdReports reviewers with extensive experience in the field. Most leads have also guided previous materials reviews and all are held in the highest professional esteem by the EdReports staff.
- Team writers are also experienced and highly respected EdReports reviewers with expertise in conveying review findings in the printed word.
- General reviewers are selected from among those who successfully complete a rigorous application process and are screened to assure they are free from any conflicts of interest.
    - Candidates complete an extensive professional survey filtering applicants by work history, content area expertise, educational leadership, and educational experience, etc. to assure all reviewers have strong expertise in both their content area as well as expertise in appropriate methods for assessment within their content area and an understanding of test construction and design.
    - Qualified candidates selected from the survey pool are invited to complete a rigorous performance task simulating aspects of an actual interim assessment review. Candidates are asked to analyze discrete aspects of an abridged assessment and justify their analysis with evidence and explanation.
    - EdReports staff of content specialists and project managers review and score performance tasks. In collaboration, they determine the candidates who will move forward into the interview phase of reviewer selection.
    - Content specialists and project managers conduct real-time interviews and make final selections for seating interim assessment review teams.
- EdReports strives for a balanced representation of educational expertise on each team. Teams consist of a mix of assessment specialists, district or state administrators, building level administrators, and classroom teachers.

Technical Review Team

The evidence necessary to support the evaluation of a given assessment depends on the purpose(s) of the assessment and the score-based interpretations necessary to use test results as intended. For this reason, those selected to conduct the evaluation must have not only a deep understanding of applied psychometric issues, but also how they interact with contextual factors to influence decisions regarding the quality, relevance and sufficiency of evidence.

The Center for Assessment will identify experts to engage in the technical review of interim assessments. Factors considered when selecting evaluators and making assessment review assignments include:

- Proven applied and technical expertise in the field of educational measurement and assessment.
- Understanding of operational and technical issues impacting assessment design, implementation and validation.
- Appropriately independent from the assessment to be evaluated: In no case should an evaluator be associated with the test under review in a manner that would make him/her feel inclined or obligated to defend (or discount) the evidence provided for personal or professional reasons.

- Although each evaluator should meet the qualifications outlined above, any focal areas of expertise (e.g., assessment of students with disabilities, Equating/Scaling; Value-Added Models; Validation, etc.) should be distributed across evaluators and take into account such factors as the intended student population and manner in which assessment results will be used.

Vendor Survey

Once a vendor decides to participate in an interim assessment review they will be asked to complete a vendor survey. The vendor survey serves as an efficient way to collect general information about the assessment that informs evidence collection, review and reporting. The vendor survey poses questions about the design of the assessment, the types of scores and information reported, and the manner in which results are intended to be interpreted and used. The responses to this survey will allow for the EdReports team to identify which criteria will and will not apply for this assessment, and to work with the vendor to ensure the right type/range of evidence is provided to support the review process.

Acquisition of Assessments and Accompanying Documentation for Review

Substantial and diverse evidence is necessary to support a comprehensive review of an educational assessment. The complete body of evidence for any given assessment will take various forms collected from a variety of different sources within or outside the vendor's organization, (e.g., test design. development, and administration specifications, test forms, test items, technical reports, research studies, meeting minutes).

Given the detailed nature of the required documentation, EdReports and the Center for Assessment have created several resources to support the evidence collection, organization and review process.

Appendix A provides examples of the types of evidence that may be submitted to inform the evaluation of each criterion and its associated indicators across all Gateways for ELA and Math. It is important to note that the examples are only for illustrative purposes and documentation may vary among assessment programs. Review tools and evidence guides are provided to vendors, in part, to assist them in identifying evidence that will best support each indicator.

The vendor is responsible for determining which pieces of evidence are necessary to support the evaluation of each criterion at the indicator level. In doing so, vendors should strive to *identify the minimum amount of documentation necessary to allow for qualified educator reviewers and technical reviewers to evaluate the extent to which the claim underlying each indicator and its associated criterion has been met.* In other words, the evidence provided should be detailed and comprehensive, but it must not be a "data dump" that requires the reviewers to sift through piles of marginally relevant materials to determine what is important. Guidelines to support the collection and organization of evidence for evaluation are provided in Appendix B.

Although the primary responsibility of determining what materials are necessary and how they should be presented falls to the vendor, EdReports and the Center for Assessment are ultimately responsible for ensuring an efficient and effective product review. EdReports and the Center for Assessment will verify that all evidence obtained from the vendor is clear, appropriate and accessible and pose questions or requests for additional information, as needed.

To help improve the efficiency of the evaluation process, EdReports and the Center for Assessment will confirm that all submitted materials are organized and indexed in a way that facilitates accessibility to educator reviewers and technical reviewers.

Vendors should delineate evidence using an Evidence Log and Evidence List (see examples in Appendix B) to clearly document and prioritize the evidence submitted in support of each indicator by evidence statement.

Secure Assembly of Evaluation Materials

Some of the evidence provided to support assessment review may be considered by the vendor to be confidential or proprietary, (e.g., test items, forms, draft procedural documentation, security protocols). To ensure confidentiality and security are maintained, all procedures related to the storage, delivery, and removal of secure evidence will be established and maintained by EdReports and the vendor as part of the initial request for information.

The Center for Assessment will ensure any secure materials submitted to support the technical evaluation (i.e., that are not part of what is routinely provided to organizations that purchase the assessment) are kept secure and access provided only to those who need it. It is the responsibility of each member of the technical review team to maintain the confidentiality of materials provided.

Evidence provided to support the evaluation process should remain available to EdReports and the Center for Assessment until the final report is published.
At that time, access to provided materials can be restricted or limited by the vendor as agreed upon by each participating organization.

Phase 2. Vendor Orientation and Evaluator Training

The primary activities associated with Phase 2 of the Interim Assessment Review include evaluator training and the vendor-delivered orientation of the evidence submitted to support the review process. The actual order in which some of the training will occur is dependent upon the timing for delivery of assessment products in the alignment and technical review. These two activities, training and orientation, are essential steps in the five-phase review process culminating in the independent review and evaluation of all submitted evidence. There will be a separate vendor orientation for the educators and the technical reviewers.

Vendor-Delivered Assessment Orientation

The vendor assessment orientation is *not* intended to be deeply technical or serve to defend/extend upon the evidence requested and provided for review. Rather, the orientation is a brief, but comprehensive presentation, focusing on background and contextual information relevant to the evaluation process. To that end, vendors should present information useful to the review and evaluation process, including background of the assessment's history, the targeted test-taking population, standards addressed, assessment design and rationale, and other aspects of the assessment pertinent to design and administration. All aspects of the vendor-delivered presentation should be consistent with the information provided in the Vendor Survey and evidence submitted via the evidence log.

The vendor's orientation should allow time and opportunity to answer questions about the assessment design and administration for clarity, noting that the evaluation will be based solely upon the evidence and materials submitted for review.

Information that we encourage vendors to provide *for the purposes of the orientation* are provided below in Table 3.

Table 3. Contextual Elements to Include in the Vendor-Delivered Assessment Orientation

Encouraged	Discouraged
• The intended purpose of the assessment and uses of assessment results • Summary of the assessment design • Intended test taking population and size • Recommended time and frequency of administration (i.e., spring, fall, etc.) • Description of how students and teachers interact with the assessment • Number and type of accommodated forms available for use • Key terminology and acronyms • List of submitted score reports • Overview of organization of materials and evidence provided for review	• Anecdotal evidence of technical quality or validity • Plans for future research and analysis • Technical procedures/ methods used to ensure quality, comparability, etc. related to intended use that should be reflected in submitted evidence • Outside opinions related to the quality of the assessment or its utility • Results or findings from other external evaluations of alignment or technical quality.

Although there will not be time to walk through every piece of evidence submitted for review, the vendor should select one or two examples to illustrate how the structure and format of the evidence repository align with the structure of the evaluation tool, focusing on any information that reviewers will need to be able to navigate the repository as a whole. If there are pieces of evidence that do not exist electronically, these should be clearly indexed and provided to EdReports or the Center for Assessment prior to reviewer training.

Following the assessment orientation, review teams will discuss the design of the assessment and organization of the evidence provided for review to gain a shared understanding of the materials. Additionally, review teams will review the information provided in the vendor request form relative to their review tasks: EdReports in relation to Criterion 1.1, Criterion 1.2, Criterion 1.3 and Gateway 3; the Center for Assessment in relation to Criterion 1.3, Gateway 2, and Gateway 3.

The review teams will determine which, if any criteria or indicators, will not be reviewed as part of this process. Reviewers will also use this information to flag additional uses suggested by score reports, technical materials or documentation that are not identified in the evidence submitted for this review. If any final questions arise during this time, EdReports or the Center for Assessment will work with the vendor, as appropriate, to get clarification or additional information.

Training of the Educator Reviewer Team

EdReports educator reviewers have amassed hundreds of collective years in operational, research and academic training in the area of educational assessment. Still, knowing the importance of interrater reliability to the overall validity of the Interim Assessment Tool, EdReports conducts extensive training prior to the review process and targeted training during the review process.

- Prior to the review process, several training sessions are required of the educator reviewers.
    - EdReports Orientation provides reviewer training on the unique responsibilities and roles of the EdReports interim assessment reviewer.
    - Functions of the Interim Assessment Tool provides reviewer training on the features and scoring mechanisms within Gateways 1 and Gateway 3 of the interim assessment tool.
    - Using the EdReports Evidence Guide backgrounds the reviewers in applying a focused, step-by-step process to the weekly review.

- Understanding the Criterion introduces the Evidence Guides to enrich reviewer understanding and knowledge of how evidence statements are used to evaluate each of the criterion indicators.
- Applying the Criterion leads educator reviewers through an analytic process simulating an operationalized review of an assessment framework and design aligned to the Criterion 1.1.
- As the review moves from Criterion 1.1 into the criteria that follow, i.e., 1.2, 1.3, 3.1, and 3.2, educator reviewers will receive topic-specific training before the actual review of the assigned commercial product.

Training of the Technical Review Team

Prior to the first scheduled assessment review, the Center for Assessment will meet with the technical review team to introduce and discuss the expectations reflected in the evidence guides and the evaluation process.

The Center for Assessment will ensure that technical reviewers:

- Understand that the tool and evidence guides are intended to guide evaluation, but the evidence statements underlying each indicator are not intended to be comprehensive or exhaustive
- Build shared understanding of indicators and address clarifications, extensions or definitions, as necessary
- Address proposed modifications to the tool if necessary
- Gather notes and comments of additional information/evidence to add to the evidence guide for consideration in evaluating a particular indicator or criterion.

To make the training as efficient as possible the tools and evidence guides will be provided in advance. This initial orientation to the process and evidence guides will occur prior to the first assessment review.

In order to calibrate on the evaluation process and criteria, the first assessment reviewed will be simultaneously evaluated by all members of the technical review team. Each subsequent review will be conducted by two technical reviewers from the larger team. Further, prior to independent review (described below) the Center for Assessment will have the technical reviewers discuss and rate one or two indicators together in order to align their thinking about the rating process and expectations.

Phase 3: Evaluation

The goal of the EdReports Assessment Evaluation is to provide districts/schools with knowledge about specific interim assessments that will inform selection and purchasing decisions.
EdReports and the Center for Assessment will provide educator reviewers and technical reviewers, with unstructured time to consider the submitted evidence against the expectations outlined for each indicator.
Across all reviewers, educator reviewers and technical reviewers, is a shared set of review tenets:

- Prior knowledge about, or experience with, the assessment should not be considered when making determinations about the degree to which expectations are met.
- If important evidence that should have been provided and is *known* to exist but does not appear in the submitted body of evidence, an evaluator may go through appropriate channels, i.e., EdReports or the Center for Assessment to request the documentation for review by the entire team.

Distribution of Criteria and Indicators in the Review Process

Educators will review evidence for all of the criteria in Gateways 1 and 3. Technical Reviewers will review evidence for all of the criteria in Gateways 2 and 3 and a subset of the criteria in Gateway 1 (See Table 2).

Table 2. Distribution of review responsibilities

Gateway 1	Criterion 1.1	Educator Reviewers (all indicators) Technical Reviewers (1.1.a)
	Criterion 1.2	Educator Reviewers
	Criterion 1.3	Educator Reviewers and Technical Reviewers
Gateway 2	All Criteria	Technical Reviewers
Gateway 3	All Criteria*	Educator Reviewers and Technical Reviewers

*Evidence statements may be divided among educator reviewers and technical reviewers where appropriate.

Establishing procedures for "shared" indicators

In Gateway 1 there are several indicators that are reviewed by both educators, and technical reviewers. These include the following:

- 1.3.a Items and test forms are developed and reviewed using procedures that ensure fairness.
- 1.3.b Appropriate accommodations and support 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.
- 1.3.c The range and types of technology provided within the assessment support the validity of assessment outcomes.

In Gateway 3 each criterion and its associated indicators are reviewed by both educators and technical reviewers. The criterion statements are provided below:

- 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.
- 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.
- 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.
- 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.

For criterion 1.3, educator ratings and feedback will be provided to the technical review team to consider when establishing their consensus indicator and criterion-level rating.

For Gateway 3, educators and technical reviewers will independently establish consensus indicator and criterion level ratings. The Center for Assessment and EdReports will jointly review these ratings and the associated rationales/comments. If the ratings differ, representatives from each group will work together to come to a shared resolution.

In both cases, the criterion-level report will include comments and feedback provided by each group so the different perspectives are clear.

Educator Reviewer Evaluation Process (EdReports)

Educator reviewers are assigned to an EdReports cluster, a group of four to five reviewers assigned to a single commercial assessment: one lead, one writer, two or three general reviewers. Each cluster is provided with a detailed action plan outlining the weekly goals at the indicator level. Having received training in the application and use of the Interim Assessment Tool, educators work independently through the week to review and interrogate vendor-provided assessment materials. Once each week, the cluster meets for a one-hour discussion of the independent findings. During the weekly meeting, the lead focuses the discussion on the specific indicator and the writer logs the discussion.

Independent Educator Reviewer Weekly Process

1. Educator reviewers are typically assigned a single indicator each week. Occasionally, indicators may be close enough in content that the week's review will target two indicators or an indicator may be expansive enough to require multiple weeks.
2. Reviewers follow the Evidence Collection Guidelines to gather documentation and data relevant to the assigned indicator.
3. Reviewers interrogate the assessment evidence applying guiding questions based on the evidence statements associated with each specific indicator.
4. Reviewers note examples of the assessment's adherence or alignment to the evidence statements as well as any absence of evidence to support the inquiry. Notations, which are kept in a secure repository, are specific and cite document titles, pages numbers, sheets, and cells wherein the data is found.
5. Reviewers also note evidence of any contradictions to the expectations of a high quality assessment as indicated by the evidence statements or guiding questions.
6. Reviewers independently rate the degree to which the provided evidence meets, partially meets, or does not meet the scoring criteria.
7. Reviewers participate in a weekly meeting to discuss and calibrate scoring, coming to a consensus on the assigned score for the weekly indicator in relation to the expectations of the interim assessment tool.
8. Review leads develop a weekly agenda to propel discussion and focus the discussion on the elements of the evidence provided. Review leads engage all general reviewers in sharing the specifics of their independent reviews as a means to deepen the analytic process and draw attention to the strongest evidence.
9. Reviewers replicate the process as the review moves from Criterion 1.1 into criterion that follow, i.e., 1.2, 1.3, 3.1, and 3.2.

Technical Review Evaluation Process (Center for Assessment)

Process for Implementing the Independent Review

The Center for Assessment is responsible for articulating a coherent, efficient evaluation strategy that ensures the expectations for evaluation are consistent and understood across reviewers. The steps associated with the independent review process are outlined below.

1. Review the Evidence Log.
2. Review the educator ratings provided for all criteria associated with Gateway 1.
3. Review and evaluate the evidence submitted for each of the indicators in Criterion evidence submitted for Criterion 2.1.
4. Review the remaining criteria and indicators within Gateway 2 (2.2-2.4), as applicable.
5. Review and evaluate the evidence submitted for each of the indicators within Criterion 1.3, considering consensus ratings and feedback provided by educators.
6. Review the criteria and indicators associated with Gateway 3.
7. Submit forms with ratings and comments to the Center for Assessment.

This process flow will be discussed with the Technical Review Team during the initial training, so that expected activities and associated materials are clearly understood. Each of these steps is discussed in detail in the section that follows:

1. **Review the vendor request survey and the range of evidence submitted for review.** Consider the evidence submitted for each criterion and indicator. Identify any core or primary pieces of evidence (i.e., that which informs multiple criteria) that should be reviewed prior to conducting your indicator-level evaluation (e.g., documents that summarize the Theory of Action, history, purpose/goals/uses and/or design of the assessment.). Note: the Center for Assessment or vendor may recommend particular documents for initial review if appropriate.
2. **Review the consensus educator ratings and comments associated with Gateway 1.** Going into Gateway 2, it is important for the reviewers to understand the degree to which the assessment content and specifications were perceived as aligned to the expectations of the standards. Understanding what was evaluated in GW 1 also ensures that the technical review appropriately extends the validity feedback collected in GW 1.
3. **Review the evidence associated with Criteria 2.1**
    a. Indicator-Level Ratings: For each indicator (2.1a - 2.1d), consider the body of evidence provided to make an overall, holistic determination about the extent to which the evidence for that indicator meets expectations given the description provided in the evidence guide and any relevant contextual considerations provided by the vendor for review. Use an Evaluator Rating Sheet to record key comments and concerns, make a preliminary indicator rating (e.g., Does Not Meet, Partially Meets, or Meets) and provide a written rationale for that rating.

    Indicator-level determinations will depend on reviewer judgment regarding the appropriateness of the evidence provided. To support consistency in the rating process across technical reviewers, an operational definition of each rating category is provided below.

Table 3. Indicator-Level Rating Descriptors for Gateway 2 and Gateway 3

Rating	
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.
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.
Does Not Meet	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.

During the independent review process evaluators will take comprehensive notes so they can share their thinking with the technical review team and the Center for Assessment during group discussion. Specifically if there were:

- particular pieces of evidence that weighed more heavily than others when evaluating a given indicator;
- additional pieces of evidence considered during review (I.e., provided by the vendor outside those listed for this indicator on the evidence log);
- contextual factors that influenced the indicator rating.

If the evidence submitted goes against or violates the expectations defined for this indicator, this should be clearly indicated in the comment section so it can be discussed with the team of technical experts.

b. Criterion-Level Performance: After evaluating the evidence associated with each indicator the evaluator should briefly summarize his/her thoughts related to the sufficiency of the body of evidence provided to support the overall criterion, and assign an overall criterion level rating of Does not Meet, Partially Meets or Meets Expectations.

   The technical reviewers should be prepared to discuss their rationale for their rating, including the degree to which different indicators held different weight in supporting their decision.

4. Review the remaining criteria and indicators within Gateway 2 (2.2-2.4), as applicable.
5. Review and evaluate the evidence submitted for each of the indicators within Criterion 1.3, considering consensus ratings and feedback provided by the educator review team. Rate each indicator and the criterion using steps 3a and 3b provided above.
6. Review Gateway 3 criteria and indicators using steps 3a and 3b.
7. Submit your Independent Evaluator Rating forms and any additional review comments to the Center for Assessment.

Calibration and Consensus Process for Technical Reviewer Findings

Shortly after the completion of independent review, the technical reviewers meet. The goal of this meeting is three-fold: 1) to establish a consensus rating on the degree to which the evidence provided for a given indicator met expectations 2) to establish a consensus rating on the strength of the body evidence presented for each criterion and 3) articulate the components of an evidence-based argument for each rating that references the evidence reviewed in relation to the expectations.

To allow ample time for these activities, it should be assumed that this meeting will take two days; however, the Center for Assessment may decide to shorten or extend this meeting as appropriate given the scope of the materials reviewed and the degree of agreement observed in the evaluators ratings.

The Center for Assessment is responsible for facilitating the discussion as well as taking detailed notes. Clearly, much more is said than can make it into the final reports. It is the responsibility of the Center for Assessment to capture and summarize consensus opinions and comments for inclusion in the final reports. To this end, the Center for Assessment may decide to record the large group discussion so that it can be referenced as needed during report generation. Throughout the discussion process, comments regarding how/why the evidence reviewed met expectations and specific areas of strength and weakness should be documented.

Phase 4: Report Generation & Approval

Two types of reports are produced from the overall EdReports review: 1) an executive summary report and a 2) a criterion-level report. The first report documents the consensus rating for each criterion along with an executive summary of the rationale for those ratings. The second report will give criterion-level alignment information that shares key evidence used to make the final score determination. The first report provides a narrative describing the scores while the second report provides analytic details for individual criterion scores.

The first report displays the criterion scores in context. Since each criterion consists of multiple indicators, these criterion-level scores will aggregate the calibrated scores determined in the EdReports reviewer teams and through consensus within the technical reviewer teams. The report will describe whether the assessment “Meets”, “Partially Meets”, or “Does Not Meet” each criterion in the Interim Assessment Tool using the

aggregated criterion-level score. For the Gateway 2 criteria, ratings are assigned in consideration of the consensus performance levels provided in step 3b of the evaluation process.

The second report provides the performance level assigned to each criterion as well as comments summarizing the quality of evidence reviewed at the indicator level. This report is intended for a more technical audience and is designed to supplement and reinforce the evaluation results summarized in the executive report. The Center for Assessment should begin drafting this report after the completion of independent technical review (i.e., Phase 2) and prior to the large group discussion (Phase 3). The writer on the EdReports teams creates aspects of this report during the weekly review cycles but only finalizes sections of this report as Phase 3 concludes for a specific criterion. This phased writing approach will allow the Evaluation Team from both organizations to review and modify the preliminary report during Phase 3. Additionally, the report can serve as a starting point for conversation. After Phase 3, EdReports and the Center for Assessment will revise the initial draft of the evaluation reports on their respective reports based on feedback and comments provided through group discussion with the technical reviewers.

The degree of initial consensus among evaluators in making indicator-level ratings will not be necessary to include in the comprehensive report, as long as evaluators were able to come to a shared understanding about the body of evidence and agree upon criterion-level ratings. The weekly conversations within the EdReports teams will ensure educator reviewers have a shared understanding of why their group decided on each score at the indicator and criterion levels. Upon completion of the intermediate draft, the evaluation report is provided to the technical reviewers and EdReports leadership for review and comment before finalization. Once the report has been completed, the vendor is given the opportunity to review and, if desired, provide a brief explanation or set of dissenting comments for consideration in an appendix to the report.

Phase 5: Errors & Omissions Process and Vendor Response

Timeline and Process for the Errors & Omissions Process:

EdReports educator review teams and Center for Assessment's technical reviewers strive to accurately and fairly evaluate materials. In order to ensure the highest quality reports, EdReports provides the Vendor with the opportunity to note any errors/omissions to bring back to the review teams. This occurs before finalizing the review as well as Vendor's Response that is published alongside the final reports.

Errors & Omissions (if desired)

The final phase of the evaluation process takes place once the reports have been drafted by the review teams and reviewed by the EdReports and Center for Assessment staff. During the Errors and Omissions process, draft reports are sent to the vendor for them to review for any factual errors, misinterpretations, or omissions of provided evidence. Vendors have one week to respond indicating their *intention* to provide counterevidence. Counterevidence is due within two weeks of the transmission of the draft reports.

Once counterevidence has been received, it is reviewed and disseminated to the appropriate review personnel (i.e., educator reviewers and technical reviewers) to review alongside previously collected evidence and ratings. A decision on final ratings is made by the review teams and is sent back to the vendor from EdReports within two weeks of receipt along with copies of the final reports. The final reports will include information as to changes that were deemed warranted and any final comments from the review teams.

EdReports will only accept one submission of errors and omissions.

Vendor Response, and Background Information:

The Vendor's submission can include both errors and omissions believed to be present in the draft report. An *error* is a citation that is factually inaccurate or misplaced. An *omission* is a citation or place in the materials that the Vendor believes was overlooked by the educator review team or technical reviewers that would merit a different score. Concerns about the application of criteria or the review criteria themselves can be included in the vendor response. EdReports is happy to discuss this process with the Vendor and answer any clarifying questions regarding the report or counter evidence submission.

Submitting a response (if desired)

The Vendor may provide a response to our review of up to 1500 words. EdReports will post the Vendor response at the same time the report is published. If the Vendor would like to change this response after the publication date to reflect any revisions to the assessment or materials, or to provide additional information, they are welcome to contact EdReports at any time. Our expectation is to post the Vendor response verbatim. However, EdReports retains the right to omit any factual inaccuracies about our process and/or unprofessional language.

Background information (if desired)

EdReports welcomes the Vendor to submit an additional 1500-word background piece to provide users of the EdReports website more background information about the assessment in the following three areas: description and detail regarding the program that informed the development of the assessment materials, evidence of efficacy, and supplemental services provided by the Vendor to support the implementation of the assessment. EdReports will post this information at the same time the reports are published. The Vendor is welcome to submit and/or revise this optional document at any time as long as it meets our criteria for length and is within the scope of three areas mentioned above.

Appendices

A. Examples of Evidence that May be Provided for Evaluation

B. Templates to Support the Collection and Organization of Evidence

C. Independent Evaluation Rating Sheet

Appendix A: Examples of Evidence by Indicator ELA & Math

This appendix is designed to provide examples of the type of evidence (e.g., materials, reports, documents) that may be submitted to support the evaluation of each criterion, its indicators, and the evidence statements contained therein..
Provided below is the full list of criteria and indicators associated with the interim assessment evaluation process. Column one states the indicator and column two provides examples of evidence or documentation that could support the review. Appendix B contains a sample prioritized evidence list that is to be completed by the vendor to clarify exactly where the educator reviewers and technical reviewers should look to identify the information submitted for each evidence statement (e.g., page numbers, tables within technical reports, select slides within a training deck). This will ensure a focus on the most important evidence during the review, rather than a divided focus as a result of having to find this information within one or more source documents.

Please note, the examples are only for illustrative purposes and documentation may vary among assessment programs. Review tools and evidence guides are provided to vendors, in part, to assist vendors in selecting the appropriate evidence they have to best support each indicator. Documents may be used to support more than one indicator.

Examples or requirements that appear in multiple indicators are provided in black font for the first instance and gray font for subsequent instances.

ELA

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.1	Assessment development guidelines and assessment blueprints align to the expectations of the college- and career-ready standards.
Indicator	Examples of Supporting Evidence
1.1.a Assessment development documentation provides clear expectations and detailed guidance to support the development of high-quality standards-aligned materials.	✔ Theory of Action/Assessment Design Rationale ✔ Description of the target assessment domain/construct ✔ Documentation of test development process/rationale ✔ Item development documentation and specifications ✔ Item writing training materials ✔ Scoring guides, rubric/s, or policy documentation for polytomously-scored items ✔ Guidelines, rationales, and processes for passage review and selection ✔ Guidelines and review processes to ensure content accuracy, technical accuracy, and editorial accuracy
1.1.b Test blueprints and/or other specifications reflect an appropriate distribution of	✔ Test blueprints, specifications, or other form development documents delineating distribution of content, item types, and cognitive demand of items within each form. ✔ Scoring matrices, guidelines, rationales, etc.

content and related score points, item types, and cognitive demand within forms.	✔ Rationales, justifications, taxonomies establishing levels of cognitive demand ✔ Information related to the framework for cognitive demand used for developing and evaluating items. ✔ Standards alignment tables or other documentation ✔ If a CAT or multistage testing is to be used, formal test assembly specifications are provided that clarify content and other related constraints, and the rules and associated rationales for the selection and administration of test items by the CAT/blocking algorithm (at item, testlet, and test level) including guidelines for determining starting points, termination conditions, and details related to exposure control, where applicable)

Criterion 1.2	Text passages, assessment items, and resulting test forms align to the expectations of the ELA domains as delineated by the college- and career-ready (CCR) standards.
Indicator	Examples of Supporting Evidence
1.2.a The text passages are of high-quality and aligned with the expectations of the CCR standards and aligned to test development documentation or blueprints.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment ✔ Actual representation of all text passages, including media or other formats, associated with reading and writing items ✔ Metadata indicating blueprint alignment ✔ Complexity reports including quantitative and qualitative analyses supporting text selection ✔ Analyses of text alignment and grade level placement ✔ Meta-data related to text type, authorship, grade level placement, complexity, etc. ✔ Listing of assessment text titles, authorship, publication status, etc.
1.2.b Test items and test item sets (e.g., EBSR) are written to elicit evidence of learning relative to one or more of the college and career ready standards and aligned to test development documentation or blueprints.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment ✔ For CAT assessments, the forms should represent the full spectrum of students along the achievement continuum ✔ Item meta-data indicating standards alignment for each item within provided test forms ✔ Item metadata indicating blueprint alignment
1.2.c The range of item types and cognitive demand within each form is sufficient to strategically assess the depth and complexity of the standards being addressed and	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment ✔ For CAT assessments, the forms should represent the full spectrum of students along the achievement continuum. ✔ Associated item meta-data indicating the level of cognitive demand associated with each item appearing on the provided sample of test forms/events

aligned to test development documentation or blueprints.	✔ Matrix indicating the range of cognitive complexity within and across testing forms ✔ Item metadata indicating blueprint alignment
1.2.d The assessment is aligned to the reading expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment
1.2.e The assessment is aligned to the writing, research, and language expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment
1.2.f The assessment is aligned to the speaking and listening expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment
1.2.g The assessment is aligned to the reading standards for foundational skills expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment

Criterion 1.3	The interim assessment is fair and accessible for all students in the intended test taking population.
Indicator	Examples of Supporting Evidence
1.3.a Items and test forms are developed and reviewed using procedures that ensure fairness.	✔ Test and item development documentation adherence to the core principles of universal design. ✔ Test and item rendering specifications ✔ Documentation related to any relevant bias/sensitivity reviews ✔ Business rules and associated rationales for evaluating and mitigating differential item and test functioning ✔ Any validity studies related to the appropriateness of the reported scores for sub-groups of students ✔ Sample items or released items representative of the assessment
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 students with disabilities and	✔ Definitions of the intended test-taking population ✔ List of provided and/or supported accommodations ✔ Evidence that the provided and/or supported accommodations are appropriate for the intended test-taking population ✔ Administration manual or relevant documentation supporting test administration ✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment

students with limited English proficiency.	
1.3.c The range and types of technology provided within the assessment support the validity of assessment outcomes.	✔ Documentation supporting test administration on the supported testing platforms (e.g., browsers, adaptive technology, operating systems) ✔ Access to any auditory supports that may be available for the provided sample forms ✔ Access to any digital tools (e.g., dictionaries, highlighters) that may be available for the provided sample forms ✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment

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	**Examples of Supporting Evidence**
2.1.a Item and form development procedures result in high-quality test events.	✔ Item development specifications (including task models and scoring rubrics) and processes ✔ Qualitative and quantitative item review, modification, and piloting procedures ✔ Item-level summary statistics for all items appearing on the sample forms provided ✔ Test development and review procedures, including documentation related to the development of test blueprints and or adaptive specifications ✔ Form-level statistics and summary data for all of the sample forms provided, including item maps if available ✔ For CAT assessments, any available summary data indicating the level of fidelity of test events to the test blueprint
2.1.b Achievement scores are reliable.	✔ Item development/review specifications ✔ Item selection criteria that may influence overall score reliability ✔ Procedures for calculating and evaluating score reliability ✔ Observed reliability estimates for the provided sample forms ✔ Summaries of reliabilities across the score continuum and at any relevant cut scores (if applicable)
2.1.c Achievement scores support intended interpretations of student performance.	✔ Documentation articulating the intended interpretations for the achievement scores. ✔ Procedures used to establish the scaled score metric and the characteristics of the scale. ✔ Performance/Achievement level descriptors (if applicable) and any applicable standard setting reports ✔ Procedures used to establish or calculate reported norms including details related to the norm group. ✔ Equating/linking procedures and results

	✔ Research agenda and/or empirical evidence supporting the validity of overall achievement scores as measures of the intended knowledge and skills
2.1.d Achievement scores are appropriate for supporting their intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information. ✔ Documentation that shows the supported uses of the achievement scores are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the achievement scores

Criterion 2.2	The interim assessment provides valid information regarding predicted student performance on a state summative assessment or other measure(s).
Indicator	Examples of Supporting Evidence
2.2.a The design of the interim assessment supports its use in predicting performance on one or more external measures	✔ Test blueprints that clearly show construct representation ✔ Rationale detailing the appropriateness of the assessment for making predictions on the intended criterion measure(s)
2.2.b Predicted results are reliable.	✔ Procedures used for calculating and evaluating the reliability of predicted scores/classifications ✔ Observed reliability estimates for the reported predictions on the student-level score reports for the provided sample of test events (forms) ✔ Summaries of studies evaluating classification accuracies
2.2.c Predicted results (e.g., expected scaled scores, performance levels, passing status, etc.) reflect a student's likely 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. ✔ Process and data used to establish the cut scores associated with predicted performance (if applicable). ✔ Descriptions and procedures for setting norms, if applicable. ✔ Procedures and results for predictive validity studies for every criterion measure associated with a prediction.
2.2.d Predicted results are appropriate for supporting their intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information ✔ Documentation that shows the supported uses of the predicted results are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the predicted results

Criterion 2.3	The interim assessment provides for valid inferences about a student's specific areas of strength and need.
Indicator	Examples of Supporting Evidence

2.3.a Test forms are designed to provide specific information about a student's areas of strength and need in the content domain.	✔ ✔ ✔	Test blueprints Test development documentation which outlines the minimum number of items/points necessary to report sub-scores at each level of granularity for which they are provided Meta-data on the provided sample forms that indicate which items are used to calculate/inform each of the reported sub-scores
2.3.b Reported sub-scores are reliable.	✔ ✔	Procedures for calculating reliability/precision indices associated with the reported sub-scores. Observed reliability estimates for the sub-scores associated with the provided sample test forms
2.3.c Reported sub-scores support intended interpretations of student performance in defined sub-skill areas.	✔ ✔ ✔ ✔ ✔	Procedures used to calculate reported sub-scores. Procedures used for establishing cut-scores, if applicable (e.g., sub-scores are reported in two or more categories rather than raw or transformed scores) Procedures for calculating and defining any applicable norm groups for norm-referenced sub-scores Dimensionality or correlational studies examining the relationships among the reported sub-scores Validity studies that provide evidence of the appropriateness of the reported sub-scores for reflecting achievement in the intended sub-domain areas
2.3.d Reported sub-scores are appropriate for supporting their intended uses.	✔ ✔ ✔	Marketing materials that outline the specific intended uses of provided score/information Documentation that shows the supported uses of the sub-scores are clearly and consistently articulated to users Validity evidence supporting the intended uses of the sub-scores

Criterion 2.4	The interim assessment provides valid information regarding student growth in the content domain.
Indicator	**Examples of Supporting Evidence**
2.4.a The interim assessment is designed to support reported measures of growth.	✔ Documentation describing the procedures of on-going monitoring of the properties of the score scale
2.4.b Student growth scores are reliable.	✔ Procedures for estimating standard errors around the reported growth information ✔ Observed reliability estimates for the reported growth information associated with the provided sample forms for multiple students along the achievement continuum
2.4.c Student growth scores support the intended interpretations.	✔ Procedures for calculating reporting student growth information ✔ When appropriate, procedures and business rules for calculating aggregate growth scores ✔ Technical reports documenting any significant programmatic changes and how they may affect growth calculations and interpretations ✔ Studies summarizing any validity studies that have been conducted to support the use of the reported growth scores for making inferences about student progress in the content domain

2.4.d Student growth scores are appropriate for supporting the intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information ✔ Documentation that shows the supported uses of the growth scores are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the growth scores

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	**Examples of Supporting Evidence**
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).	✔ Documentation describing information provided on various score reports as applicable to various audiences. ✔ User guides and interpretive materials intended to support appropriate use and interpretation of reported scores. ✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.
3.1.b Score reports include information about the degree of error associated with the achievement score.	✔ ✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores. ✔ ✔
3.1.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. ✔ User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. ✔ Access to instructional or curricular supports. ✔ Technical manual documentation connecting data analysis, score interpretation, and intended uses. ✔

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	Examples of Supporting Evidence
3.2.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).	✔ Documentation describing information provided on various score reports as applicable to various audiences. ✔ User guides and interpretive materials intended to support appropriate use and interpretation of reported scores. ✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.
3.2.b Score reports include information about the degree of error associated with the predicted performance score.	✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores.
3.2.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. ✔ User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. ✔ Access to instructional or curricular supports. ✔ Technical manual documentation connecting data analysis, score interpretation, and intended uses.

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	Examples of Supporting Evidence
3.3.a The design of the score reports and supporting materials (e.g., user manuals	✔ Documentation describing information provided on various score reports as applicable to various audiences.

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).	✔	User guides and interpretive materials intended to support appropriate use and interpretation of reported scores.
	✔	Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators).
	✔	Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised.
	✔	Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison.
	✔	Warnings of how to avoid misuse of scores and resultant score reports.
3.3.b Score reports include information about the degree of error associated with sub-scores.	✔	Information regarding degree of error associated with predicted performance and its interpretation.
	✔	User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores.
3.3.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔	Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators.
	✔	User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance.
	✔	Access to instructional or curricular supports.
	✔	Technical manual documentation connecting data analysis, score interpretation, and intended uses.

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	Examples of Supporting Evidence	
3.4.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).	✔	Documentation describing information provided on various score reports as applicable to various audiences.
	✔	User guides and interpretive materials intended to support appropriate use and interpretation of reported scores.
	✔	Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators).
	✔	Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised.
	✔	Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison.

	✔	Warnings of how to avoid misuse of scores and resultant score reports.
3.4.b Score reports include information about the degree of error associated with student progress scores.	✔ ✔	Information regarding degree of error associated with predicted performance and its interpretation. User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores.
3.4.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔ ✔ ✔ ✔	Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. Access to instructional or curricular supports. Technical manual documentation connecting data analysis, score interpretation, and intended uses.

.

MATH

Gateway 1: Alignment, Fairness, & Accessibility

Criterion 1.1	Assessment development guidelines and assessment blueprints align to the expectations of the college-and career-ready standards.
Indicator	Examples of Supporting Evidence
1.1.a Assessment development documentation provides clear expectations and detailed guidance to support the development of high-quality standards-aligned assessments.	✔ Theory of Action/Assessment Design Rationale ✔ Description of the target assessment domain/construct ✔ Documentation of test development process/rationale ✔ Item development documentation and specifications ✔ Item writing training materials ✔ Scoring guides, rubric/s, or policy documentation for polytomously-scored items ✔ Guidelines, rationales, and processes for passage review and selection ✔ Guidelines and review processes to ensure content accuracy, technical accuracy, and editorial accuracy ✔ Information related to the framework for cognitive demand used for developing and evaluating items.
1.1.b Test blueprints and/or other specifications focus strongly on the content that is most important for students to master by reflecting an appropriate distribution of content and related score points.	✔ Test blueprints, specifications, or other form development documents delineating distribution of content, items, and item points within each form For CAT assessments, test blueprint documentation may also include: ✔ If a CAT or multistage testing is to be used, formal test assembly specifications are provided that clarify content and other related constraints, and the rules and associated rationales for the selection and administration of test items by the CAT/blocking algorithm (at item, testlet, and test level) including guidelines for determining starting points, termination conditions, and details related to exposure control, where applicable). ✔ Simulations studies with content alignment data

Criterion 1.2	Assessment items and resulting test forms align to the expectations of the Math standards as delineated by the college- and career-ready standards.
Indicator	Examples of Supporting Evidence

1.2.a Test forms delivered to students reflect an appropriate distribution of content and related score points and item types within forms.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) ✔ Associated item meta-data indicating standards alignment and answer keys associated with each item appearing on the provided sample of test forms.
1.2.b Test items are written to elicit evidence of learning relative to one or more of the college- and career-ready standards and aligned to test development documentation and/or blueprints.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) ✔ Associated item meta-data indicating the content alignment associated with each item appearing on the provided sample of test forms.
1.2.c The range of item types and cognitive demand within each series of assessments is sufficient to strategically assess the full intent and complexity of the standards being addressed and aligned to test development documentation and/or blueprints.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) ✔ Associated item meta-data indicating the level of cognitive demand associated with each item appearing on the provided sample of test forms. ✔ Matrix or blueprint indicating range of cognitive demand across testing forms.
1.2.d The assessment is aligned to the procedural skill and fluency expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) In addition, vendors may choose to provide documentation may like: ✔ Item specifications, particularly for the standards requiring procedural skills and fluencies ✔ Released items and models given to assessment writers that target procedural skills and fluencies

1.2.e The assessment is aligned to the conceptual understanding expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) In addition, vendors may choose to provide documentation may like: ✔ Item specifications, particularly for the standards requiring conceptual understanding ✔ Released items and models given to assessment writers that target conceptual understanding
1.2.f The assessment is aligned to the application expectations of the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) In addition, vendors may choose to provide documentation may like: ✔ Item specifications, particularly for the standards requiring application ✔ Released items and models given to assessment writers that target application
1.2.g The assessment includes mathematical practices as described in the CCR standards.	✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile (as described in Section 2) In addition, vendors may choose to provide documentation may like: ✔ Item specifications, particularly for the standards requiring mathematical practices ✔ Released items and models given to assessment writers that target specific mathematical practices

Criterion 1.3	The interim assessment is fair and accessible for all students in the intended test taking population.
Indicator	Examples of Supporting Evidence
1.3.a Items and test forms are developed and reviewed	✔ Test and item development documentation adherence to the core principles of universal design. ✔ Test and item rendering specifications.

using procedures that ensure fairness.	✔ Documentation related to any relevant bias/sensitivity reviews. ✔ Business rules and associated rationales for evaluating and mitigating differential item and test functioning. ✔ Any validity studies related to the appropriateness of the reported scores for sub-groups of students. ✔ Sample items or released items representative of the assessment.
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 students with disabilities and students with limited English proficiency.	✔ Definitions of the intended test-taking population. ✔ List of provided and/or supported accommodations. ✔ Evidence that the provided and/or supported accommodations are appropriate for the intended test-taking population. ✔ Administration manual or relevant documentation supporting test administration. ✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment
1.3.c The range and types of technology provided within the assessment support the validity of assessment outcomes.	✔ Documentation supporting test administration on the supported testing platforms (e.g., browsers, adaptive technology, operating systems). ✔ Access to any auditory supports that may be available for the provided sample forms. ✔ Access to any digital tools (e.g., dictionaries, highlighters) that may be available for the provided sample forms. ✔ 3-test events at each assessed grade: 25th percentile, 50th percentile, and 75th percentile ✔ Reviewer completed simulated assessment

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	Examples of Supporting Evidence
2.1.a Item and form development procedures result in high-quality test events.	✔ Item development specifications (including task models and scoring rubrics) and processes ✔ Qualitative and quantitative item review, modification, and piloting procedures ✔ Item-level summary statistics for all items appearing on the sample forms provided ✔ Test development and review procedures, including documentation related to the development of test blueprints and or adaptive specifications ✔ Form-level statistics and summary data for all of the sample forms provided, including item maps if available

	✔ For CAT assessments, any available summary data indicating the level of fidelity of test events to the test blueprint
2.1.b Achievement scores are reliable.	✔ Item development/review specifications ✔ Item selection criteria that may influence overall score reliability ✔ Procedures for calculating and evaluating score reliability ✔ Observed reliability estimates for the provided sample forms ✔ Summaries of reliabilities across the score continuum and at any relevant cut scores (if applicable)
2.1.c Achievement scores support intended interpretations of student performance.	✔ Documentation articulating the intended interpretations for the achievement scores ✔ Procedures used to establish the scaled score metric and the characteristics of the scale ✔ Performance/Achievement level descriptors (if applicable) and any applicable standard setting reports ✔ Procedures used to establish or calculate reported norms including details related to the norm group. ✔ Equating/linking procedures and results ✔ Research agenda and/or empirical evidence supporting the validity of overall achievement scores as measures of the intended knowledge and skills
2.1.d Achievement scores are appropriate for supporting their intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information ✔ Documentation that shows the supported uses of the achievement scores are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the achievement scores

Criterion 2.2	The interim assessment provides valid information regarding predicted student performance on a state summative assessment or other measure(s).
Indicator	Examples of Supporting Evidence
2.2.a The design of the interim assessment supports its use in predicting performance on one or more external measures	✔ Test blueprints that clearly show construct representation ✔ Rationale detailing the appropriateness of the assessment for making predictions on the intended criterion measure(s)
2.2.b Predicted results are reliable.	✔ Procedures used for calculating and evaluating the reliability of predicted scores/classifications ✔ Observed reliability estimates for the reported predictions on the student-level score reports for the provided sample of test events (forms) ✔ Summaries of studies evaluating classification accuracies
2.2.c Predicted results (e.g., expected scaled scores, performance levels, passing	✔ The data and procedures used to establish and evaluate the predictive relationship for a given test taking sample

status, etc.) reflect a student's likely performance on the state summative assessment or other intended criterion measure(s).	✔ Process and data used to establish the cut scores associated with predicted performance (if applicable) ✔ Descriptions and procedures for setting norms, if applicable ✔ Procedures and results for predictive validity studies for every criterion measure associated with a prediction
2.2.d Predicted results are appropriate for supporting their intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information ✔ Documentation that shows the supported uses of the predicted results are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the predicted results

Criterion 2.3	The interim assessment provides for valid inferences about a student's specific areas of strength and need.
Indicator	Examples of Supporting Evidence
2.3.a Test forms are designed to provide specific information about a student's areas of strength and need in the content domain.	✔ Test blueprints ✔ Test development documentation which outlines the minimum number of items/points necessary to report sub-scores at each level of granularity for which they are provided ✔ Meta-data on the provided sample forms that indicate which items are used to calculate/inform each of the reported sub-scores
2.3.b Reported sub-scores are reliable.	✔ Procedures for calculating reliability/precision indices associated with the reported sub-scores ✔ Observed reliability estimates for the sub-scores associated with the provided sample test forms
2.3.c Reported sub-scores support intended interpretations of student performance in defined sub-skill areas.	✔ Procedures used to calculate reported sub-scores ✔ Procedures used for establishing cut-scores, if applicable (e.g., sub-scores are reported in two or more categories rather than raw or transformed scores) ✔ Procedures for calculating and defining any applicable norm groups for norm-referenced sub-scores ✔ Dimensionality or correlational studies examining the relationships among the reported sub-scores ✔ Validity studies that provide evidence of the appropriateness of the reported sub-scores for reflecting achievement in the intended sub-domain areas
2.3.d Reported sub-scores are appropriate for supporting their intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information ✔ Documentation that shows the supported uses of the sub-scores are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the sub-scores

Criterion 2.4	The interim assessment provides valid information regarding student growth in the content domain.

Indicator	Examples of Supporting Evidence
2.4.a The interim assessment is designed to support reported measures of growth.	✔ Documentation describing the procedures of on-going monitoring of the properties of the score scale
2.4.b Student growth scores are reliable.	✔ Procedures for estimating standard errors around the reported growth information ✔ Observed reliability estimates for the reported growth information associated with the provided sample forms for multiple students along the achievement continuum
2.4.c Student growth scores support the intended interpretations.	✔ Procedures for calculating reporting student growth information ✔ When appropriate, procedures and business rules for calculating aggregate growth scores ✔ Technical reports documenting any significant programmatic changes and how they may affect growth calculations and interpretations ✔ Studies summarizing any validity studies that have been conducted to support the use of the reported growth scores for making inferences about student progress in the content domain.
2.4.d Student growth scores are appropriate for supporting the intended uses.	✔ Marketing materials that outline the specific intended uses of provided score/information ✔ Documentation that shows the supported uses of the growth scores are clearly and consistently articulated to users ✔ Validity evidence supporting the intended uses of the growth scores

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. Score reports support accurate and appropriate interpretations of student performance.
Indicator	**Examples of Supporting Evidence**
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).Multiple versions of score reports are available and effectively designed for use	✔ Documentation describing information provided on various score reports as applicable to various audiences. ✔ User guides and interpretive materials intended to support appropriate use and interpretation of reported scores. ✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.

by students/parents, teachers, and administrators in the manner intended.	
3.1.b Score reports include information about the degree of error associated with the achievement score.Score reports and other resources (e.g., user's manual/interpretive guides) are developed to ensure overall achievement scores are interpreted and used appropriately to support decision-making.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores. ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.
3.1.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.Score reports and other resources (e.g., user's manual/interpretive guides) are developed to ensure information about predicted performance is interpreted and used appropriately to support decision-making.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. ✔ User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. ✔ Access to instructional or curricular supports. ✔ Technical manual documentation connecting data analysis, score interpretation, and intended uses. ✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Score reports generated from reviewer taken assessments. ✔ Warnings of how to avoid misuse of scores and resultant score reports.

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	Examples of Supporting Evidence
3.2.a The design of the score reports and supporting materials (e.g., user manuals and interpretive guides) and	✔ Documentation describing information provided on various score reports as applicable to various audiences. ✔ User guides and interpretive materials intended to support appropriate use and interpretation of reported scores.

the types of information provided are consistent with the intended interpretations and uses for specific users (e.g., educators, parents, students, or administrators).	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.
3.2.b Score reports include information about the degree of error associated with the predicted performance score.	✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores.
3.2.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. ✔ User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. ✔ Access to instructional or curricular supports. ✔ Technical manual documentation connecting data analysis, score interpretation, and intended uses.

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	Examples of Supporting Evidence
3.3.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).	✔ Documentation describing information provided on various score reports as applicable to various audiences. ✔ User guides and interpretive materials intended to support appropriate use and interpretation of reported scores. ✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.
3.3.b Score reports include information about the degree	✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores.

of error associated with sub-scores.	
3.3.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. ✔ User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. ✔ Access to instructional or curricular supports. ✔ Technical manual documentation connecting data analysis, score interpretation, and intended uses.

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	Examples of Supporting Evidence
3.4.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).	✔ Documentation describing information provided on various score reports as applicable to various audiences. ✔ User guides and interpretive materials intended to support appropriate use and interpretation of reported scores. ✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences, (e.g., students, teachers, administrators). ✔ Procedures to identify and flag students for which the integrity of the intended interpretations may be compromised. ✔ Alerts to test design and/or administration factors that may threaten fair use of achievement scores to make intended comparison. ✔ Warnings of how to avoid misuse of scores and resultant score reports.
3.4.b Score reports include information about the degree of error associated with student progress scores.	✔ Information regarding degree of error associated with predicted performance and its interpretation. ✔ User guides and interpretive materials supporting appropriate interpretation and actionable use of reported achievement scores.
3.4.c Sufficient and appropriate guidance (e.g., instructional or curricular supports) is provided to support the intended interpretations and uses, when needed.	✔ Copies of real score reports generated from test events using the sample forms for each of the intended audiences,e.g., students, teachers, administrators. ✔ User guides and interpretive materials to support appropriate, actionable use and interpretation of predicted performance. ✔ Access to instructional or curricular supports. ✔ Technical manual documentation connecting data analysis, score interpretation, and intended uses.

Appendix B: Guidelines and Templates to Support the Collection and Organization of Evidence

This appendix is designed to provide guidelines to support the identification and organization of evidence, to support evaluation. Those charged with assembling the evidence provided for review are responsible for determining what (and how much) evidence is necessary and appropriate to support the evaluation of each indicator and criterion. In doing so they should strive to identify *the minimum amount of evidence necessary to allow for qualified technical reviewers to make informed and reliable determinations regarding the extent to which a given expectation is supported.* Some general guidelines to support this process are provided below:

Comments and Guidelines to Support the Identification and Organization of Evidence

- Put yourself in the evaluators' shoes. What information do they need to understand and interpret the evidence provided? For example, are there program-specific terminologies or acronyms that require definition? Information that is interesting (and only marginally related), but not necessary should be considered extraneous.
- If the same *procedures, methods and statistical criteria* are generally used across multiple grades and content areas (e.g., those related to test development/review, scaling and equating procedures, standard setting, etc.) it is not necessary to provide documentation of these procedures (and the accuracy with which they were implemented) multiple times. A limited sample of evidence can be provided with a comment around the grades/content areas to which it generalizes.
- If, there are *important* procedural differences across grades/subjects that are *relevant to the evaluation of a given criterion/indicator* (e.g., differences in the type of external data provided to support standard setting at the high-school vs. the elementary school level) the information and context necessary to support evaluation of both procedures should be clearly identified in the evidence list.
- When indicators *require the review of key outputs that would vary across grade spans or subjects* (e.g., validity evidence supporting on grade or on track inferences; reliability coefficients; score reports), relevant results and documentation should be provided for all grades/content areas.
- If an indicator is asking about the manner in which a particular type of data is represented, described, formatted or presented; exemplary samples of that output, rather than all instances should be sufficient to support evaluation.
- Documentation related to the endorsement of evidence by an external review committee should be detailed enough so that evaluators will know what the committee reviewed, the nature of the discussion surrounding the relevant material and the recommendations that resulted. An agenda for a meeting does not meet this requirement.
- To assure accuracy and efficiency in the evaluation process, in the narratives directing evaluators to relevant evidence for a given indicator, evaluators should be provided with clear document references, page numbers, and paragraph/table/figure citations to guide them directly to the relevant evidence. The documents themselves should be marked up to highlight the relevant evidence (e.g., drawing red boxes around the

relevant evidence, highlighting in yellow, annotating which claim(s) the marked up content addresses). If the context of an entire document is not deemed relevant, an excerpt may be provided for brevity, but a full reference to the document should be provided.
- To ensure access to the evidence, all documents should be provided in commonly used formats, such as PDF and Microsoft Office.

If a specific type of evidence is unavailable or believed not to be applicable given the goals of the test, the Vendor should include comments explaining why this is the case.

When collecting and organization evidence it is extremely important that those charged with organizing the set of materials that will be submitted to EdReports by the Vendor for evaluation note the following:

1. Vendors will **not** be given multiple opportunities to provide evidence to Technical Reviewers once the evaluation process has begun (i.e., the Evaluation Team has been provided with the materials for review). Therefore the Vendor must meet all requirements for evidence the *first time around*. The provision of data/evidence is not an iterative process.
2. There are several places in the evidence guide where process-based documentation must be supplemented by evidence that the process actually occurred. Vendors should ensure that, when necessary, evidence that a process occurred as intended is provided to the evaluation team for review (e.g., outcomes of technical advisory meetings, information gained through external review of procedures/materials, etc.).
3. Evaluators are looking for a coherent validity argument in the evidence provided. They are not planning to have to construct one on their own! When multiple pieces of evidence must be considered simultaneously in support of a given indicator or criterion, the Vendor is responsible for reflecting this in their submission. It should not be assumed that the evaluator will put the pieces together as needed to support the evaluation.

Sample Evidence Log

The table below is an example of an evidence log that includes reports that could be used in the evaluation of an assessment program. This document is to be completed by the Vendor. This example is provided to help the organizations submitting their own documents and to illustrate the type of descriptions that would be appropriate.

Document Number	Document Name	Brief Description
1	Technical Report (2018-2019)	Technical Report – includes grades 3-8 Math, Reading, Writing and Science
2	Sample of Score Reports	Sample student, school, roster, and teacher reports
3	Score Report Interpretive Guide	Materials provided to support test users in interpreting each score report
4		
5		
6		
.		

When submitting the evidence log, for each piece of evidence submitted, enter the document name (as it appears in the directory or electronic evidence file) and provide a brief description of its contents. If there is a set of interdependent documents that will always be viewed together, these can be concatenated and assigned one document number.

Prioritized Evidence List

In the table below, you will see an example of a prioritized evidence list that includes reports that could be used in the evaluation of a hypothetical assessment program. In this example, each row of the Prioritized Evidence List corresponds with a specific *indicator* and *associated evidence statement*.

Each piece of evidence should be referenced by the document number assigned on the E-log (or something similar) and, when appropriate, those page numbers (chapters, appendices, etc.) most relevant to evaluating a particular indicator should be noted. If the relevance of a particular piece of evidence is not readily apparent or additional background information is necessary to support its review, this should be noted in the comments section. Similarly, if two pieces of evidence are linked, or should be jointly considered in service to a given indicator, this should also be noted with comments. It is assumed that for most assessments there will be at least *some* evidence provided to support each indicator. If a specific type of evidence is unavailable or believed to be not applicable given the goals of the test, the Vendor should include comments explaining why this is the case.

Indicator	Evidence statement	Materials to be Reviewed	Comments
2.1a	2.1a Item development, review, and piloting procedures and materials were designed to ensure all newly developed items meet technical quality standards.	#1, Pages 15-20	Includes a summary of the item development process and timeline
2.1a	Blueprints and test development and review procedures ensure forms meet content specifications and statistical quality criteria.	#1, Pages 25-40	
		#1, Pages 50-55	

When submitting your evidence summary, for each evidence statement, enter the document number and name (as it appears in the directory or electronic evidence file) and a brief description of its contents. When supplying this information, provide as much specificity as possible, whether it is specific chapters in a technical manual, or specific pages from the minutes of a TAC meeting. In the comments section, information should be supplied to help the evaluators fully comprehend how the document supports the evidence statement.