Technical Report 2603-SUM

Overview of 2026 Series Summarizing easyCBM® Research

Gerald Tindal, PhD – University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Note: This technical report was supported in part by Riverside Insight, the exclusive distributor of easyCBM®. The report does not reflect any endorsement by any of these organizations.

APA Reference: Tindal, G. (2026). *Overview of 2026 Series Summarizing easyCBM® Research (Technical Report 2603-SUM)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Overview of 2026 Series Summarizing easyCBM® Research

The following Technical Reports represent the integration of technical adequacy research on easyCBM® conducted over the past 25 years by researchers at *Behavioral Research and Teaching (BRT* – https://brtprojects.org). By integrating this research, reviewers can more easily make informed judgments on adoption of various measures. Each report summarizes the original research, though each study is also available on the BRT website. A total of seven summaries are available: Two focus on test development in reading and mathematics, a single summary of both reading and mathematics on alignment to standards, and finally two documents for each of reading and mathematics that address reliability and validity.

Tindal, G., & McCaslin, S. (2026c). *Test Development for easyCBM® in Grades K-8: Reading (Technical Report # 2603-TDK8R)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026d). *Test Development for easyCBM® in Grades K–8: Mathematics (Technical Report # 2603-TDK8M).* Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G. (2026). *Alignment of easyCBM® with Standards in Grades K-8: Reading and Mathematics (Technical Report # 2603-A38RM)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026a). *Reliability of easyCBM® in Grades K-8: Mathematics (Technical Report # 2603-RK8M)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026b). *Reliability of easyCBM® in Grades K-8: Reading (Technical Report # 2603-RK8R)*. Eugene, OR.: Behavioral Research and Teaching, University of Oregon.

Tindal, G., & McCaslin, S. (2026e). *Validity Analyses for easyCBM® in Grades K-8: Mathematics (Technical Report # 2603-VK8M)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

Tindal, G., & McCaslin, S. (2026f). *Validity Analyses for easyCBM® in Grades K-8: Reading (Technical Report # 2603-VK8R)*. Eugene, OR: Behavioral Research and Teaching, University of Oregon

These technical reports not only integrate the research conducted to date but also advance interpretation of this research with reference to three approaches for establishing technical adequacy. Of course, the most authoritative reference is the *Standards* publication. The next two institutions (Center for Assessment [CA] and National Center on Intensive Interventions [NCII]) provide current applications of these standards and differ primarily in their breadth. The former (CA) provides a frame of reference closely oriented toward large-scale testing and state accountability programs. The latter (NCII) is specifically oriented to Multi-Tier Systems of Support (MTSS) and the identification of students with disabilities. Importantly, these three organizational frameworks do not present the same references for evaluating tests. In this next section, more specific information is presented to eventually provide a basis for comparison, as well as inform a different view of the validation process.

Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014)

This critical document provides the overarching authority on establishing all technical adequacy documentation and is considered the ultimate reference. The *Standards* not only consider reliability and validity, but also several larger issues that address test administration protocols, attention to various student populations (e.g., students with disabilities and English language learners), fairness and accessibility, training of personal, and consequences from testing programs.

Interim Assessments from the Center on Assessment (Center for Assessment & EdReports.org, 2023b) and (Center for Assessment & EdReports.org, 2023a)

Two documents are published by the Center: (a) Review Criteria for Interim Assessment English Language Arts: Grades 3-8 (Center for Assessment & EdReports.org, 2023a), and (b) Review Criteria Interim Assessment Mathematics: Grades 3-8 (Center for Assessment & EdReports.org, 2023b). Both documents provide descriptions of Gateways (narrative descriptors of necessary ‘big idea’ topics to be addressed) as well as the specific criteria and indicators to be used by reviewers in evaluating tests or testing programs submitted by vendors (typically to state educational agencies).

- Gateway 1: Alignment, Fairness, & Accessibility
- Gateway 2: Technical Quality
- Gateway 3: Score Reports and Interpretive Guides

National Center on Intensive Interventions (NCII) https://intensiveintervention.org/tools-charts/overview

This organization (https://intensiveintervention.org) provides information on a wide range of topics that address databased interventions, training, and evidence-based interventions, as well as evaluations (Tools Charts) for screening and progress monitoring. In this document, we focus primarily on the Screening Tools Charts.

Standards for Educational and Psychological Testing

This brief overview focuses only on reliability and validity as well as their definitions and procedures for establishing them.

Reliability – Definitions and Types

In the *Standards*, reliability is defined as the degree to which test scores for a defined population are consistent across replications of the testing process, and therefore the extent to which scores are free from measurement error. Reliability is not a fixed property of a test itself; it is a property of the scores obtained in a particular context of use and for a particular group. The *Standards* emphasize that reliability is best understood through an error framework: observed scores vary around a true score (or another reference value) because of multiple sources of error, and reliability quantifies the proportion of observed-score variance that is attributable to true-score variance (or, equivalently, how precisely scores can be interpreted).

The *Standards* describe several complementary ways to estimate and report reliability, each linked to a different facet of replication. (1) Internal consistency evidence (e.g., coefficient alpha, omega, split half) reflects the extent to which items or parts of a test function together to measure a common construct and yield consistent scores within a single administration. (2) Test-retest evidence reflects score consistency over time when the same people are tested on separate occasions, supporting interpretations that assume stability across the retest interval. (3) Alternate or parallel-form evidence reflects consistency across different but equivalent forms, supporting comparability when forms are used interchangeably. (4) Inter-rater or inter-judge evidence addresses consistency when human scoring is involved (e.g., essays, performance assessments), including rater agreement and consistency indices. (5) Generalizability theory extends these ideas by modeling multiple sources of error (facets such as items, raters, occasions) simultaneously to estimate how reliably scores generalize across the relevant universe of admissible observations.

Across all approaches, the *Standards* stress clear reporting: the reliability coefficient (or error index), the population and conditions under which it was estimated, the form(s) used, the time interval for retest, and the score uses to which the evidence is intended to apply. They also highlight the role of conditional precision, encouraging use of standard errors of measurement and related indices to show that precision can vary across the score scale and across subgroups.

Validity – Definitions and Sources of Evidence

The *Standards* define validity as the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is not a single statistic; it is an integrated argument that links the intended score interpretations and decisions to a coherent network of empirical evidence. The focus is on the meaning and use of scores, not on the test as an instrument. Accordingly, the same test can have strong validity evidence for one purpose and weak evidence for another. And validation is an ongoing process as uses, populations, and contexts change.

Validity evidence is organized into five sources. (1) Evidence based on test **content** examines the alignment of test tasks, items, and scoring to the construct definition and the domain the test is intended to represent (including content relevance and representativeness). (2) Evidence based on **response processes** investigates whether examinees and scorers engage in the intended cognitive, behavioral, or scoring processes (often using think-aloud, process data, rater analyses, or audits of scoring rubrics). (3) Evidence based on **internal structure** evaluates whether relationships among items, tasks, and subtests match the construct's dimensionality and theoretical

expectations (e.g., factor structure, IRT models, measurement invariance). (4) Evidence based on relations to other variables considers convergent and discriminant relations, **criterion-related** evidence, and predictive/retrodictive relationships with relevant outcomes. (5) Evidence based on **consequences** of testing examines both intended and unintended outcomes of test use, including fairness implications, impact on instruction or access, and whether decisions based on scores lead to appropriate actions.

The *Standards* emphasize that these sources do not function as a checklist; the weight placed on each depends on the claims being made. A strong validity argument explicitly states the proposed interpretations and uses, identifies plausible threats (such as construct-irrelevant variance or construct underrepresentation), and uses evidence to evaluate and refine the argument.

Other Topics Addressed

Beyond reliability and validity, the *Standards* provide guidance for the full life cycle of educational and psychological testing. They address test design and development (defining purpose, construct, target population, and score scale; creating items/tasks; piloting; and documenting technical quality). They outline expectations for scoring, scaling, equating, and linking, including how to support score comparability across forms, administrations, and versions. Finally, the *Standards* discuss test administration, security, accommodations, and accessibility, emphasizing standardized procedures and careful control of conditions that could distort score meaning.

A major set of topics concerns fairness in testing: avoiding bias, evaluating differential performance that reflects construct-irrelevant factors, and ensuring equitable access to valid score interpretations. Related chapters address testing individuals with disabilities and English learners, including appropriate accommodations, alternate assessments where warranted, and evidence that score interpretations remain defensible for these groups. The *Standards* also cover rights and responsibilities of test takers and test users, including informed consent where applicable, privacy and confidentiality, and appropriate score reporting.

The *Standards* provide guidance for uses of tests in selection, placement, diagnosis, and accountability, stressing that decision rules, cut scores, and classification accuracy require explicit rationale and evidence, and that errors of decision should be considered. They address technical documentation and reporting, calling for transparency about intended uses, limitations, and the proper interpretation of scores. Finally, the *Standards* discuss responsibilities of developers, publishers, and users, including ongoing monitoring of test performance, appropriate revisions, and ethical practice consistent with professional standards.

Center for Assessment Gateways with Criteria and Indicators

The review tool organizes evidence about interim assessments into three sequential “gateways” which represent a major aspect of technical and instructional quality that must be addressed before an assessment can be judged appropriate for use. The same gateways-criteria-indicators are provided for both English Language Arts (ELA) (Center for Assessment & EdReports.org, 2023a) and Mathematics (Center for Assessment & EdReports.org, 2023b). The intent is identical across content areas: (a) confirm strong alignment to college- and career-ready standards and equitable access, (b) confirm that scores and claims have sufficient technical quality for their intended interpretations, and (c) confirm that score reports and guidance enable correct, responsible use by educators and other stakeholders.

Within each gateway are criteria (broad requirements) and indicators (specific, observable evidence statements). Indicators are rated, and criteria carry point totals or are “claim-dependent,” meaning points apply only when the publisher claims to provide a particular type of result (e.g., predictions, sub scores, growth). Together, the gateways function like an argument: The assessment must first measure the right constructs (Gateway 1), then produce defensible results (Gateway 2), and finally communicate those results in ways that support sound decisions (Gateway 3)

Gateway 1: Alignment, Fairness, and Accessibility

Gateway 1 evaluates whether the interim assessment is built to measure the intended standards and whether the assessment experience is fair and accessible. In both math and ELA, this gateway looks for a clearly articulated test

design that targets the breadth and depth of the standards, emphasizes the most important content, and reduces construct-irrelevant barriers so that all students have an equitable opportunity to demonstrate what they know.

Criterion 1.1: Test Development Alignment (8 points).

Focus: design specifications and blueprints that drive high-quality, standards-aligned development.

- Indicator 1.1.a (0/2/4): Design specifications provide clear expectations and detailed guidance. Evidence includes a rationale and research foundation; robust item development documentation; clear scoring rules/rubrics across item types; processes to ensure content accuracy and editorial/technical quality; and specified cognitive-demand ranges sufficient to measure the depth of standards.

- Indicator 1.1.b (0/2/4): Blueprints/specifications emphasize the most important content. For Grades K-8 this means score points concentrate on the "major work" of the standards; for high school courses (if applicable) the distribution reflects the content and skills needed for college and careers.

Criterion 1.2: Item and Form Alignment (16 points)

Focus: alignment of actual items and delivered test events to the standards and intended design.

- Indicator 1.2.a (0/2/4): Delivered forms/events reflect an appropriate distribution of content, score points, and item types—strongly focused on major and supporting clusters (or equivalent priority content).

- Indicator 1.2.b (0/1/2): Items elicit evidence of learning relative to one or more standards without measuring irrelevant knowledge/skills; items align to the design specifications; and items are content-accurate and free of technical/editorial flaws.

- Indicator 1.2.c (0/1/2): Item types and cognitive demand across events are sufficient to assess the full intent and complexity of the targeted standards and align to the blueprint/specifications.

- Indicator 1.2.d (0/1/2): Evidence the assessment addresses standards requiring procedural skill and fluency (or analogous skill components), reflected in item documentation and points distributions.

- Indicator 1.2.e (0/1/2): Evidence the assessment addresses standards requiring conceptual understanding, reflected in items and points distributions.

- Indicator 1.2.f (0/1/2): Evidence the assessment addresses standards requiring application; for HS, modeling expectations (when relevant) are administered to all students.

- Indicator 1.2.g (0/1/2): The assessment includes the discipline-specific practices (e.g., mathematical practices; in ELA, comparable practice/skill expectations) as reflected in specifications and delivered forms, with items requiring the practice to earn full credit when aligned.

Criterion 1.3: Fairness and Accessibility (12 points)

Focus: universal design, accommodations, and technology features that protect score meaning for all students.

- Indicator 1.3.a (0/2/4): Development and review procedures ensure fairness. Evidence includes adherence to universal design principles, item rendering specifications aligned to universal design, review processes that minimize construct-irrelevant variance, bias/sensitivity review, and subgroup/accommodation analyses to evaluate technical quality.

- Indicator 1.3.b (0/2/4): Appropriate accommodations and supports are available for intended populations (including students with disabilities and English Learners). Evidence includes clear documentation of intended test-taking populations, accommodations aligned to intended uses, sufficiency of accommodations, and validity/fairness evidence for interpretations under accommodations, plus clear administration guidance and availability of sample forms/items.

- Indicator 1.3.c (0/2/4): Technology features support validity. Evidence includes platform-access guidance, usable auditory supports (natural voice and adjustable cadence), and an overall visual/digital-tool design (e.g., calculators, highlighters) that is navigable and not distracting.

Gateway 2: Technical Quality

Gateway 2 evaluates whether the assessment's results are technically strong enough to support the interpretations and uses claimed by the vendor. This gateway is structured around four possible types of results: overall achievement, predictions of future performance on an external criterion, sub scores for strengths and needs, and progress/growth over time. The framework is intentionally "claim-dependent": If a program does not claim to provide predictions, sub scores, or growth measures, those criteria are treated as not claimed and are not scored; if the program does claim them, it must provide evidence that the corresponding results are designed appropriately, reliable/precise, valid for the intended interpretations, and suitable for intended uses.

Criterion 2.1: Overall Achievement (8 points)

Focus: defensible achievement scores for the target content domain.

- Indicator 2.1.a (0/1/2): Item and form development procedures yield high-quality test events, with review and piloting aligned to content and statistical quality standards.
- Indicator 2.1.b (0/1/2): Achievement scores are reliable. Evidence includes clear procedures for estimating reliability/precision and obtained indices appropriate for intended use.
- Indicator 2.1.c (0/1/2): Achievement scores support intended interpretations. Evidence may include validity studies, equating/linking methods supporting comparability across events, and documentation that promotes consistent presentation/scaffolding and scoring.
- Indicator 2.1.d (0/1/2): Achievement scores are appropriate for intended uses, supported by sufficient theoretical and empirical justification and consistent articulation of use cases.

Criterion 2.2: Predicted Student Performance (claim-dependent)

Focus: predicted results relative to a state summative or other criterion measure (scored only if claimed).

- Indicator 2.2.a (0/1/2; may be N/C): Design supports prediction by demonstrating construct/content similarity to the criterion and providing evidence for specific prediction claims (e.g., to a named test).
- Indicator 2.2.b (0/1/2; may be N/C): Predicted results are reliable; procedures for estimating reliability of predicted scores/classifications are documented and aligned to the inference (e.g., classification reliability).
- Indicator 2.2.c (0/1/2; may be N/C): Predicted results reflect likely future performance; studies document data, samples, methods, and interpretive logic supporting the predictive relationship.
- Indicator 2.2.d (0/1/2; may be N/C): Predicted results are appropriate for intended uses, supported by adequate theoretical/empirical evidence and clear articulation of use.

Criterion 2.3: Sub-scores (claim-dependent)

Focus: strengths-and-need sub scores at reported strands/objectives (scored only if claimed).

- Indicator 2.3.a (0/1/2; may be N/C): Test events are designed to support reporting at each stated level of granularity and to justify interpreting strengths/needs within the content domain.
- Indicator 2.3.b (0/1/2; may be N/C): Reported sub scores are reliable/precise, with defensible, documented estimation methods and adequacy for intended uses.
- Indicator 2.3.c (0/1/2; may be N/C): Sub scores support intended interpretations and represent distinct sub-domains, supported by empirical evidence.
- Indicator 2.3.d (0/1/2; may be N/C): Sub scores are appropriate for intended uses, with sufficient theoretical/empirical support and clear use statements.

Criterion 2.4: Student Progress (claim-dependent)

Focus: progress/growth interpretations across administrations (scored only if claimed).

- Indicator 2.4.a (0/1/2; may be N/C): The assessment is designed to support growth; content and scale characteristics (within/across grades) match the vendor's growth model.

- Indicator 2.4.b (0/1/2; may be N/C): Growth scores are reliable, including appropriate standard errors and evaluation of precision along the score scale.

- Indicator 2.4.c (0/1/2; may be N/C): Growth scores support intended interpretations; methods are documented, and evidence addresses potential disruptions (e.g., redesign/rescaling) and confirms growth inferences.

- Indicator 2.4.d (0/1/2; may be N/C): Growth scores are appropriate for intended uses, supported by adequate evidence and clearly articulated use cases.

Gateway 3: Score Reports and Interpretive Guides

Gateway 3 evaluates whether score reports and supporting materials help users interpret results accurately and use them responsibly. Even technically strong scores can be misused if reports are unclear, omit uncertainty information, or lack guidance connecting results to action. Accordingly, this gateway checks for audience-appropriate design, communication of error, and actionable guidance. As in Gateway 2, criteria related to predictions, sub scores, and progress are claim-dependent and scored only when those results are provided.

Criterion 3.1: Overall Achievement (10 points)

Focus: reporting and guidance that support correct interpretation and use of overall achievement results.

- Indicator 3.1.a (0/2/4): Report design and information are consistent with intended interpretations and users (educators, parents, students, administrators). Evidence includes user-centered design attention, studies/focus groups showing users can interpret and use reports, warnings about common misuses, and flags for compromised test integrity with conditions explained.

- Indicator 3.1.b (0/1/2): Reports communicate score uncertainty (e.g., confidence intervals, error bands, probability statements) and provide supports/examples to interpret error and its practical implications.

- Indicator 3.1.c (0/2/4): Guidance and supports (instructional/curricular or interpretive) are sufficient and appropriate, aligned to intended use, grounded in research or educator consultation, and cover the full performance range.

Criterion 3.2: Predicted Student Performance (claim-dependent)

Focus: reporting and guidance for predicted performance results (scored only if claimed).

- Indicator 3.2.a (0/2/4; may be N/C): Reports and materials match intended uses for predicted results and demonstrate that intended users can interpret them; include misuse warnings and integrity flags when relevant.

- Indicator 3.2.b (0/1/2; may be N/C): Reports include uncertainty around predicted results and supports to interpret that uncertainty.

- Indicator 3.2.c (0/2/4; may be N/C): Guidance is provided to support appropriate use across the performance range, aligned to the predictive use claim.

Criterion 3.3: Sub-scores (claim-dependent)

Focus: reporting and guidance for sub scores (scored only if claimed).

- Indicator 3.3.a (0/2/4; may be N/C): Reports and materials are consistent with intended uses for sub scores and show users can interpret them; include misuse warnings and integrity flags when relevant.

- Indicator 3.3.b (0/1/2; may be N/C): Reports include uncertainty around sub scores (error bands or similar) with interpretive supports.

- Indicator 3.3.c (0/2/4; may be N/C): Guidance supports appropriate sub score use for students across performance levels and is aligned to the sub score purpose.

Criterion 3.4: Student Progress (claim-dependent)

Focus: reporting and guidance for growth/progress results (scored only if claimed).

- Indicator 3.4.a (0/2/4; may be N/C): Reports and materials for progress results match intended uses and audiences and demonstrate interpretable, usable displays; include misuse warnings and integrity flags when relevant.

- Indicator 3.4.b (0/1/2; may be N/C): Reports communicate uncertainty around progress estimates with supports to interpret it.

- Indicator 3.4.c (0/2/4; may be N/C): Guidance supports appropriate use of progress information, aligned to the growth model and covering the full performance range.

An important consideration is that Interim Assessments have become relevant with the shift from summative testing programs initially used as part of No Child Left Behind (NCLB) to a broader conception of accountability programs that nevertheless, follow the guidelines used with peer review. Importantly, because of their use in large-scale assessment programs, the same standards described below, apply to them.

National Center on Intensive Interventions (NCII)

NCII presents the narrowest conception of technical adequacy for instruments (presumably one of them used in a Multi-Tier Support System – MTSS) that is focused primarily on (a) reliability using correlations between two 'forms' (b) criterion validity, and (c) classification analyses. The following text represents quotations from the web site for the tools charts.

Reliability

Full Bubble: Either (a) a model-based approach to reliability was reported or (b) at least two other types of reliability were reported appropriate for the purpose of the tool and drawn from at least two samples that are representative of students across all performance levels. And for each type of reliability reported the lower bound of the confidence interval around the median estimate met or exceeded 0.70.
Half Bubble: Either (a) a model-based approach to reliability was reported or (b) at least two other types of reliability were reported appropriate for the purpose of the tool, drawn from at least one sample that is representative of students across all performance levels. And/or for each type of reliability reported the lower bound of the confidence interval around the median estimate fell below 0.70 but met or exceeded 0.60.
Empty Bubble: Does not meet full or half bubble.
Dash: Reliability data were not provided.

Validity

Full Bubble: There are at least two types of appropriately justified validity analyses[1] from a sample representative of students across all performance levels *and* the lower bound of the confidence interval around each standardized estimate met or exceeded 0.60 (or if not, within an acceptable range given the expected relationship with the criterion measure(s)).
Half Bubble: Analyses, measures, and sample were appropriate, but evidence was mixed, with one or more estimate(s) either not meeting or exceeding 0.60 or not within an acceptable range given the expected relationship with the criterion measure(s)
Empty Bubble: Does not meet full or half bubble.
Dash: Validity data were not provided

Classification Accuracy

Note: Classification Accuracy is rated separately for each criterion measure and time of year for the administration (e.g., Fall, Winter, Spring). Ratings will be provided for up to two different criterion measures and up to three different time points. Data for additional criterion measures or administration times may be reported but will not be rated.
Full Bubble: All of Q1 – Q3 (below) rated as YES and the lower bound of the confidence interval around the Area Under the Curve (AUC) estimate ≥ 0.80 and Sensitivity ≥ 0.80 and Specificity ≥ 0.80.
Half Bubble: All of Q1-Q3 rated as YES (below) and either (a) the lower bound of the confidence interval around the AUC estimate ≥ 0.70 but < 0.80 or (b) Sensitivity ≥ 0.70 and Specificity ≥ 0.70.
Empty Bubble: Does not meet full or half bubble.
Q1. Was an appropriate external measure of academic performance used as an outcome?
Q2. Was risk adequately defined within an RTI approach to screening (i.e., 10^{th}- 20^{th} percentile)?
Q3. Were the classification analyses and cut-points adequately performed?
Area Under the Curve (AUC) Statistic: an overall indication of the diagnostic accuracy of a Receiver Operating Characteristic (ROC) curve. ROC curves are a generalization of the set of potential combinations of sensitivity and specificity possible for predictors. AUC values closer to 1 indicate the screening measure reliably distinguishes among students with satisfactory and unsatisfactory reading performance, whereas values at 0.50 indicate the predictor is no better than chance.

Note that **usability information** is also presented for each assessment but not evaluated.

Conceptions of Technical Adequacy (Reliability and Validity)

In this next section, examples of differences are described, but should only be considered illustrative. Using the *Standards* as a frame of reference, it is possible to compare interim assessments with the Center for Assessment Gateway criteria and indicators with MTSS documents (submitted to be compliant with NCII tools charts for screening).

Illustrative Comparison between Interim and MTSS Assessments

Example Issue	Topic	Center for Assessment	NCII for MTSS
Assessment Content	Content Validity	Item Blueprints	Not evaluated
Reliability	Conceptual definition with range of metrics, including regression analyses	Traditional criteria used with the Standards	Limited to Pearson Correlations between two ‘forms’
Interpretation of Validity	Four broad sources of evidence (see Figure 1 below).	Traditional criteria used with the *Standards* and addition of reporting and decision making	Limited to Criterion-related evidence and Classification Analyses

In summary, the *Standards* set the stage for state and local educational agencies (SEAs and LEAs, respectively) to adopt and deploy testing programs ranging in use from accountability systems (Center for Assessment and Peer Review) to Multi-tier Support Systems (MTSS).

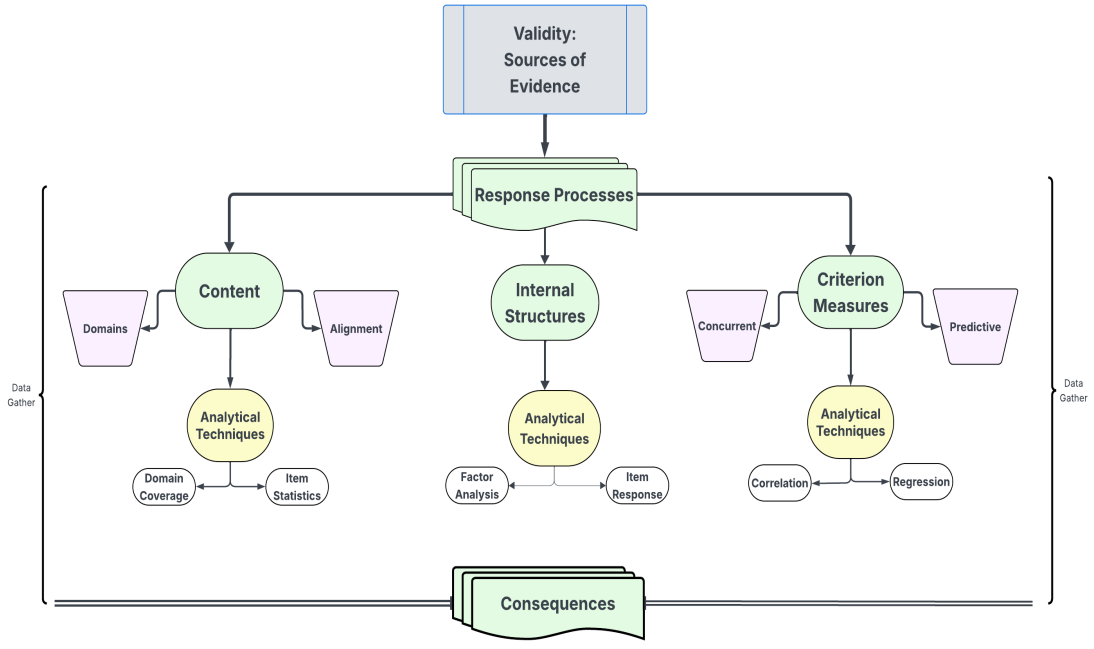
As of 2026, numerous vendors have been listed with NCII with evaluations of their MTSS measurement programs, most of them listed in a publication by the (Oregon Department of Education, 2025) in their survey of academic tests (the District Assessment Inventory, or DAI) that Oregon school districts required schools to administer…”Based on this survey's findings, ODE has provided recommendations and best practices that include observations of areas where current practices are to be commended and suggestions for improvement. The recommendations and best practices are organized in 10 categories:

1. Clarify assessment purpose and goals.
2. Align tests with learning objectives.
3. Use a variety of assessment methods.
4. Integrate assessment into instruction.
5. Leverage staff resources and technology.
6. Prioritize essential tests.
7. Optimize assessment timing.
8. Collaborate with students, families, and educators.
9. Provide professional development.
10. Regularly evaluate and adjust.

Immediately following publication of this inventory, the Oregon state legislature passed House Bill 141 (Oregeon Legislature: 83rd OREGON LEGISLATIVE ASSEMBLY–2025 Regular Session, 2025) requiring all districts to adopt one of four interim assessment that had been reviewed by a panel of measurement experts. The evaluation was based on the Gateways-Criteria-Indicators published by the Center for Assessment. In the end, the assessment landscape in Oregon is now cluttered with multiple assessments being used for both MTSS and state accountability (which also includes the state summative test used to pass peer review, as noted above). We argue, however, that many (most) of these MTSS assessment programs are quite applicable as Interim Assessments, given that they can satisfy the criteria promulgated by the Center for Assessment. At the same time, it is possible for an interim assessment to be used in the context of MTSS. However, this equivalence in function depends upon changing the validation process. And this transformation in validation needs to mover from a traditional perspective of validity to a more comprehensive view.

In a traditional view of validity, measurement systems are evaluated rather holistically and categorically. For example, a test is developed with specific content and administered under standardized conditions. Thus, two major sources of evidence are considered, though the response processes are rarely overtly considered in the context of universal design and accommodations (Tindal, 2025). Then, other sources of evidence are used to establish that the measure in question has not only integrity as a measure (with coherent internal structures) but also relates to other measures. In the end, this series of validation inquiries provide a holistic view of consequential validity.

Figure 1. Diagram of Traditional View of Reliability and Validity



This traditional view, however, is insufficient in expanding the view of validity to the full range of differentiated decision-making for MTSSs. Not only do the NCII criteria (as they currently exist) fail to satisfy a deep validation of MTSS but they also fail to satisfy the broader criteria developed for interim assessments. Furthermore, the criteria for interim assessments also fail to satisfy the criteria needed for a comprehensive validity argument applied to a Multi-Tiered System of Supports (MTSS) (Oregon Department of Education, 2025). In this last section, the *Standards* are applied in a multi-variate manner across the range of measures used, from the initial benchmarks to the eventual progress measures. We use easyCBM® as an example.

Application of a Comprehensive Validity Argument

easyCBM® reports both an Overall Outcome in each subject area (Reading, Spanish literacy, and Math) as well as a matrix reflecting Multi-Skill Multi Method (MS-MM) approach: multiple skills are assessed using multiple methods (selection and production responses as well as computer-based and paper-pencil with a variety of accommodations). By focusing on skill specific distributions, and the use of normative performance at three benchmark periods, information on student performance quickly bridges overall achievement with sub score analyses (determination of risk) and appropriate progress monitoring (where required). Errors in interpretation are also addressed with the convergent and discriminate information that may be contrary to expectations, allowing teachers to follow up with more refined analyses (e.g., using different assessment methods). A *User Manual* provides teachers extensive information on how to document this process.

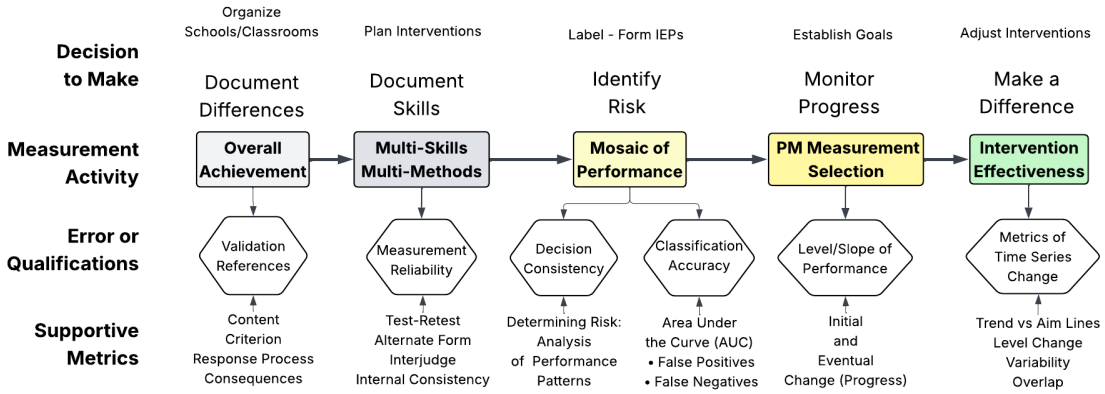
A **Comprehensive Validity Argument** is now possible by **beginning with a decision to be made** and then associating it with an appropriate measurement activity, which also carries with it qualifications or potential sources of error. First, easyCBM® documents Overall Achievement in reading, Spanish, and math that is supported with reference to multiple skill performance documented through multiple measurement methods (MS-MM). Overall achievement and multiple skills are interpreted as relative performance, using percentile rank to note individual differences. Further consideration is given to this MS-MM by analyzing this mosaic of separate sub-test performances and classification accuracy to determine risk of failing to succeed and providing students supports (Tiers 2 and 3). The decision-making process then turns from *documenting individual differences* to *making an individual difference*. The focus now is on selecting appropriate measures for progress monitoring (PM) and finally in determining intervention effectiveness, through interrupted time series graphic displays for each student.

In **Figure 4** below, the top row displays the decision to make; the second row presents the measurement activity (often the most visible event that requires teachers' time and training); in the third row is potential error (or qualification) associated with the measurement activity; finally, in the bottom row is a description of supportive metrics, that lead back to the measurement activity and decision being made. Notice the colors change from neutral (gray) to success with hesitation (yellow) to success (green). This complete validation argument fits within the larger science of education in which conjectures are affirmed only after dubitation is quieted. In the end, all scientific findings are tentative ad infinitum—never absolute (Attribution to Dr. Ed Kame'enui).[1]

To complete this comprehensive validity argument, data need to be collected and used to vindicate or modify interpretations and decisions made with attention to consequential validity. "Evidence for Validity and Consequences of Testing: Some consequences of test use follow directly from the interpretation of test scores for uses intended by the test developer. The validation process involves gathering evidence to evaluate the soundness of these proposed interpretations for their intended uses. Other consequences may also be part of a claim that extends beyond the interpretation or use of scores intended by the test developer" (American Educational Research Association et al., 2014).

[1] Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge – 2nd Edition*. New York: Routledge.

Figure 2. Comprehensive Validity Argument in easyCBM®



In supporting the claims, evidence would address the differences among students in their overall achievement. This evidence ideally would attend to not only averages but perhaps more importantly, the variance among different aggregates and would be documented at all levels. Teachers would use this information to group students so instruction could be efficient. Principals could use this information to streamline staffing and allocation of support resources (like instructional assistants and parent volunteers). Central office personnel (curriculum coordinators and elementary/secondary coordinators as well as school psychologists and teachers on special assignment serving as consultants) could use this information to organize their schedules and address targeted support.

Overall achievement, however, is unlikely to be sufficient in developing actionable interventions. Rather, a more fine-grained analysis is warranted using the MS-MM Matrix. Again, at all levels, information across and within subject domains (e.g., reading and mathematics) would warrant a refined proactive analysis and reactive response that is targeted. Such analyses can then be skill-specific and timely. Importantly, these analyses set up a responsive time-series mechanism for interventions being implemented 'just in time'.

Given the assumption that diagnoses are limited and not always correctly made, the next set of interpretations and decisions simply increase the stakes. In a MS-MM system for identifying risk, two classes of information are necessary. The first is at the *teacher* level, individually or as part of a Response-to-Intervention (RTI) team, as well as at the *individual* student level. Risk can be operationalized not only in terms of ranking students on a normative basis overall (low percentile ranks), but also in the pattern of skill deficits. Given that all skills are not equal, attention can be devoted to necessary pre-requisite skills as the basis for classifying students in need of support. Ideally, the measurement system has established validation data using classification accuracy, which can serve as the basis for judging sensitivity and specificity.

Once students have been grouped into more intensive tiers (Tier 2 and Tier 3), the somewhat ambiguous task remains to monitor progress with the appropriate measures. Assuming the previous interpretations and decisions have been carefully wrought, measures can be selected as 'a bird in the mine shaft' where generalizations can be made to the larger constructs of grade appropriate reading and mathematics problem-solving. Verification can be determined by attending to initial level of performance and immediate gains made, with changes made before too much time has been taken. Obviously, the skills for monitoring progress would also map tightly into the interventions being deployed.

The final interpretations and decisions focus on the effects of interventions and answering the following question: Is change in performance and progress being made? Answers to this critical question can be more specifically answered by addressing the following four questions.

1. Is level of performance sufficient, reflecting neither a floor nor ceiling effect?
2. Is change over time (slope or rise over run) being made?
3. Is variation of performance minimal, particularly relative to the slope?
4. Are goals being met, particularly relative to the slope?

If these basic questions are not affirmatively answered, individually and collectively, intervention changes are warranted. At this point, the validation process moves to an interrupted time-series design and the student's graph is punctuated with a vertical line that separates the data series into pre- and post. Three new questions can then be answered:

5. Does a change in level occur (a comparison between the last data value of the previous intervention and the first data value of the new intervention)?
6. Does the slope increase and the variation around it decrease?
7. Is overlap minimal: A horizontal line demarcating the difference between the highest data value in previous intervention and the lowest data value in the subsequent intervention.

In summary, a comprehensive validity argument involves tying claims to evidence and iteratively addressing specific interpretations. And once begun, " it is commonly observed that the validation process never ends, as there is always additional information that can be gathered to more fully understand a test and the inferences that can be drawn from it. In this way an inference of validity is similar to any scientific inference" (American Educational Research Association et al., 2014).

References

American Educational Research Association, American Psychological Association, & Education, N. C. o. M. i. (2014). *Standards for Educational and Psychological Testing*. Author.

Center for Assessment, & EdReports.org. (2023a). *Review Criteria Interim Assessment English Language Arts Grades 3-8 (v1.0, Final 5/2023).* Dover, NH: EdReports.org and Center for Assessment

Center for Assessment, & EdReports.org. (2023b). *Review Criteria: Interim Assessment Mathematics Grades 3-8 (v1.0, Final 5/2023).* Dover, NH: EdReports.org and Center for Assessment

Oregeon Legislature: 83rd OREGON LEGISLATIVE ASSEMBLY–2025 Regular Session. (2025). *House Bill 141*. Salem, OR

Oregon Department of Education. (2025). *HB 4124 Legislative Report: District Assessment Inventory*. Salem, OR: Author

Tindal, G. (2025). *Rethinking "Standardization" for NAEP to Increase Equity and Access (Technical Report 2510)*. Eugene, OR: University of Oregon Behavioral Research and Teaching