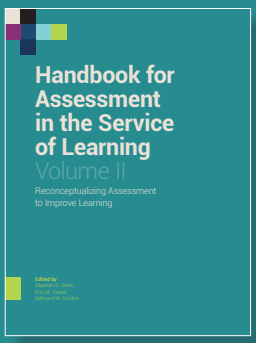
Removing the "Psycho" from Education Metrics

Stephen G. Sireci and Neal Kingston

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Sireci, S. G., & Kingston, N. (2025). Removing the “psycho" from education metrics. In S. G. Sireci, E. M. Tucker, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume II: Reconceptualizing assessment to improve learning.* University of Massachusetts Amherst Libraries.

Removing the "Psycho" from Education Metrics

Stephen G. Sireci and Neal Kingston

Abstract

In this chapter, the authors provide a brief overview of the traditional psychometric concepts of scaling and calibration and then describe more contemporary notions that leverage technology to (a) develop complexity scales for items and tasks, and (b) provide actionable information for teachers and learners. Different methods of reporting assessment results, and how those results can support learning, are also discussed.

The field of "*psychometrics*" has been described as "the measurement of psychological characteristics such as abilities, aptitudes, achievement, personality traits, skills, and knowledge" (American Psychological Association et al., 1985, p. 93). Dissecting the word into its constituent roots, "psycho" refers to the measurement of unobservable characteristics of people, and "metrics" refers to measurement. The process of measurement requires a scale, so what is measured can be quantified. Unfortunately, measuring unobservable attributes is difficult, and proper quantification of those attributes is even more complex. The approximately 150-year history of psychometrics has reflected that complexity, and in doing so may have prevented educational assessments from reaching their full potential to help students learn. How can educational assessments help students learn? We believe one way is by providing clear information about student learning that can help both learners and their teachers know what students know and what to do next. However, providing clear and actionable information from educational tests has not been a strength of psychometrics.

In this chapter, we review some of the history and terminology of psychometrics relevant to educational assessment and illustrate new ways of using assessment information to better serve learners. Some of the new directions we suggest involve removing the layer of complexity that has traditionally come with scaling educational assessments. We posit that the primary purpose of educational achievement testing should be to enhance student learning by providing valuable feedback that can guide both students and teachers. When tests are designed to identify areas of strength and weakness, they help students understand what they have mastered and where they need to focus their efforts. This targeted feedback allows students to take ownership of their learning and make informed decisions about how to improve. For educators, test results can highlight which concepts need to be revisited or taught differently, ensuring instruction is responsive to student needs.

Properly designed educational assessments can foster educational achievement by emphasizing progress and improvement over time. Instead of viewing tests as immutable arbiters of ability, students can see them as opportunities to demonstrate growth and learn from their mistakes. This shift in perspective can reduce test anxiety and encourage a more positive attitude towards learning. By focusing on the *formative* aspects of testing, educators can create a supportive environment where students feel motivated to engage deeply with the material and strive for continuous improvement.

Moreover, when the primary goal of testing is to enhance learning, it aligns assessment practices with the broader educational mission of developing lifelong learners. Tests that are integrated into the learning process, rather than being isolated events, can help students develop critical thinking skills, problem-solving abilities, and a deeper understanding of the subject matter. This approach ensures testing is not just a measure of what students know at a single point in time, but a tool that actively contributes to their ongoing educational journey.

To properly promote new ways of using assessment results to support learning, we first briefly describe the current influence of psychometrics in educational assessment. Thus, we begin with some brief history to illustrate why psychometric perspectives have dominated the educational assessment field. We then introduce some key terms used in reporting the results of educational assessments. The final sections of the chapter describe what we believe to be particularly effective ways to report the results of educational assessments to help learners learn.

Relevance of Psychometrics to Educational Assessment

Why are most, if not all, educational and psychological test results reported on score scales? To answer this question, we must go back to the first experimental psychology laboratory established by Wilhelm Wundt at the University of Leipzig (Germany) in the 1880s. Working in Wundt's lab was Ernst Weber, who measured people's perceptions of physical stimuli such as weight, temperature, and pressure under strictly controlled conditions. He found when physical stimuli were increased or decreased (e.g., when additional weight was added to a scale) participants in his experiments did not always notice the increase or decrease until some threshold (of increase or decrease) was reached. He called this threshold the "just noticeable difference," and modeled the relationship between "stimulus intensity" and "perceived intensity" using a simple ratio. Fechner was highly influenced by this work because he was convinced the mind existed independently of the body and was looking for a way to prove it. He extended Weber's work using a logarithmic formula with the unit of measurement being these "just noticeable differences." We will skip the details here (which can be found in most introductory psychology textbooks or in Sireci et al., 1998), but what is important to note here is Fechner used these "just noticeable differences" to develop the first "psychological scale." It was this scaling process that laid the claim for psychology as a legitimate, quantifiable science.

Thus, the beginning of psychological measurement required a scale to quantify what was being measured. As the science of educational measurement evolved, this notion of scaling was adopted. The earliest forms of large-scale educational tests ordered people along the score scale based on their test performance. These scales were designed to capture "individual differences," and given the strong influence of Charles Darwin's work on variation and genetic evolution around this same time, much of the focus in early psychometrics was using measurement to understand the psychology of individual (and group) differences, rather than on understanding *each* individual. Francis Galton, Darwin's cousin, led that effort, which eventually became known as the eugenics movement (Sireci & Randall, 2021).

Today, the idea of using intelligence tests to categorize and order groups of people has been resoundingly refuted, in large part because the creation of such tests is culturally dependent (Malda et al., 2010). However, the idea of placing people on a continuous score scale based on a test endures via a scaling process called *item response theory* (IRT). In the next section we briefly explain traditional and IRT scales, which are the means for describing learners' performance on educational tests.

What is a Test Score Scale?

There are many different ways students' performance on an educational test can be reported. In most cases, the communication of a student's performance on a test will be reported on some type of numerical scale. In this section, we provide a conceptual overview of traditional test score reporting scales, as well as their strengths and limitations.

How are the Results of Educational Assessments Typically Reported?

The results from educational assessments can be reported in many ways (See for example, Linn & Gronlund, 2005; Zenisky & Hambleton, 2016). The simplest way to report test takers' performance on a test is to provide the number of points earned. For example, if a test is worth 50 points, and a student earned 30 points, the student's test score may be reported as 30. This simple score is often referred to as the *raw score* because there is no transformation of this simple count of the number of points earned. Thus, raw scores are not on a scale per se, and are determined solely by the number of items on test, the number of points each item is worth, and the number of points earned by the student.

Taking into account the maximum number of points that could be earned is often done by reporting the *percent correct score*, which is the percentage of points earned by the test taker divided by the maximum number of points that could possibly be earned. For example, 30 points out of a maximum of 50 points would be reported as a percent correct score of 60%. By accounting for the maximum number of possible points that could be earned, the percent correct score is an intuitive scale that ranges from 0 to 100—the percent correct score scale.

Raw scores and percent correct scores are often criticized for having little meaning because their interpretation is completely bound to the specific set of items on a test. For this reason, raw scores and percent correct scores cannot be compared across different tests, even if the different tests are designed to measure the same knowledge or skills. For example, "90% correct" may seem like a good score on a difficult test, but not such a good score on an easy test. Thus, raw score and percent correct metrics fall short for many testing purposes, such as tracking learners' performance over time, or using different sets of items (test questions) to avoid mistaking memorization for learning. To address this problem, *scale scores* are used on tests designed for more enduring purposes such as monitoring performance over time or generalizing learners' performance to a wider domain of knowledge and skills. Scale scores are test results that are reported on a "common" scale, so when different students take different tests, their performance can be meaningfully compared on the same scale.

The process of creating a score scale for educational tests uses the statistical concept of *deviation*, which reflects differences, or distances, of test takers from a focal point, such as a mean test score, or from one another. In educational assessment, this scaling perspective is known as *norm-referenced* testing, and must be distinguished from *criterion-referenced* testing, which is another way to report the results from educational assessments. Thus, before describing how score scales from educational assessments are typically created, we first describe the difference between these two assessment orientations. Understanding the perspectives of and differences between norm- and criterion-referenced testing is perhaps the most important knowledge needed to properly interpret, understand, and explain learners' performance on educational tests.

Norm-referenced and Criterion-referenced Interpretations of Test Performance

In the 1980s, one of the most popular shows on television was *Cheers*. The setting for the show was a neighborhood bar in Boston called, you guessed it—Cheers. What does this show and this bar have to do with interpreting test scores? There were two diehard customers who were always in the bar—Norm and Cliff[1]. They were the "average Joes" who hung out of the bar, talked to everyone, were nice to everyone, nosey, and so forth. Although they argued with each other they were friends. Norm has the perfect name because he was average at best—average meaning he was a "normal" guy. In the same way, norm-referenced score interpretations are referenced to the average—to the Norm. In statistics or sampling, the norm can refer to a standard normal distribution or some expected value of average. Thus, one way students' test scores can be interpreted is how far they are from this average—from this norm.

Norm-referenced testing grounds interpretation of learners' performance with respect to a *norm group*, which is a sample that sufficiently represents the population to be tested. Using this representative sample, learners' performance on a test is interpreted by how well they did relative to the norm (e.g., did she score above or below the mean?), or against specific groups of people (e.g., did she score in the top 10% of all students who took the test?) The norm group is typically established at a particular point in time. That is, the nationally representative sample tested in the norming year serves as the reference group until the test is re-normed. Percentile ranks and similar derivations of raw scores can be used to report norm-referenced information.

The normal curve is typically used to derive scale scores on educational tests. A *standard normal deviate* scale converts raw scores from a norm group to deviation scores, based on how far a learner's score is from the mean of the norm group, in standard deviation units. The formula for this transformation is,

$$SS = \frac{X-\mu}{\sigma} \quad [1]$$

where SS is the scale score that corresponds to the raw score X, μ is the mean raw score of the norm group, and σ is the standard deviation of raw scores in the norm group. On this scale, a score of -1 indicates a raw score that is one standard

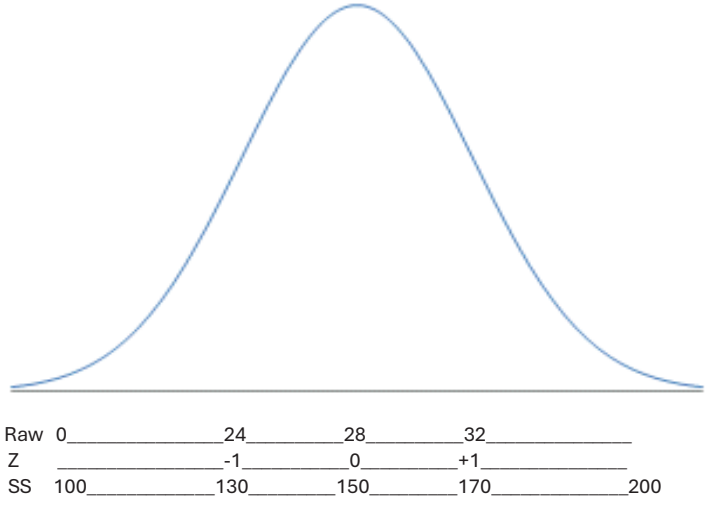
1 Not to be confused with Norman Cliff who wrote an excellent multivariate statistics book in 1987.

deviation below the norm group, and a score of 1 indicates a raw score that is one standard deviation above the norm group. Most score scales transform this normal deviate scale to one that avoids negative numbers and conforms to a scale considered to be more intuitive for interpretation. For example, the math section of the SAT college admissions test ranges from 200 to 800 with a mean of 500 and a standard deviation of 100 (based on a nationally representative sample of high school seniors in 2016—the mean and standard deviation may shift over time). Learners who score 650 on that scale have performed 1.5 standard deviations above the mean of the norm group.

An example of norm-referenced scales derived from the norming of raw scores is presented in Figure 1. This is a fictitious example from a nationally normed geography test on which there were 36 questions on the test form that was normed, and each question was worth one point. Thus, the raw score scale ranged from 0 (no items correct) to 36 (all items correct). Very few students did extremely poorly or extremely well on the assessment, which gave a rough normal distribution to the data. The mean of the norm group on this test was 28 and the standard deviation was 4. Under the raw score scale is the standard normal deviate scale (Z), and under that scale is a transformation of the normal deviate scale to report "geography test scale scores" (SS) that range from 100 to 200 with a mean of 150 and standard deviation of 20[2]

2 Transformations such as this one can be made using the equation of a straight line (i.e., a linear transformation), where SS=$bX+a$, and b the slope of the line) is equal to the ratio of the desired standard deviation (100) to the observed standard deviation (4), and a=$\mu-(b\bar{X})$. In this example $ss = [\frac{100}{1}x] + [150 - [(\frac{100}{1})(0)]]$, given the mean and standard deviation of the standard normal deviates are 0 and 1, respectively. This example illustrates how choice of scale is an arbitrary decision that can be used to report scores in a metric considered most acceptable by the testing agency.

Figure 1.
Illustration of Scale Score Transformations



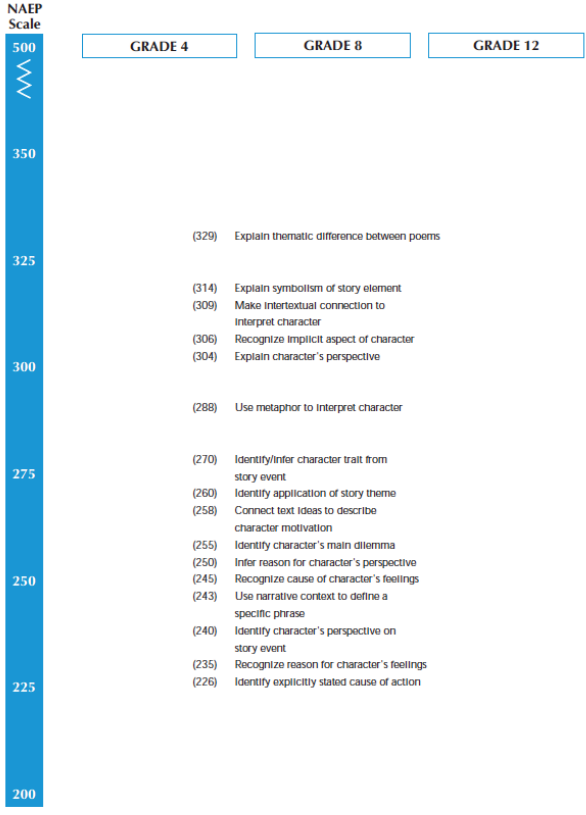
It should be noted the distribution of raw scores does not need to have a normal distribution for this transformation, and there is often an equating step involved when there is more than one test form involved. Those details are beyond the scope of this chapter and so readers who are interested in learning more about the psychometric activities of scaling and equating are referred to Angoff (1984) or Kolen (2006).

Item response theory (IRT) is an advanced probabilistic model for scaling educational tests that relates the probability of a learner correctly answering an item to the learner's location on the theoretical scale being measured. IRT also involves calibrating items onto the scale, which allows for learners to be placed on the same scale, even when they respond to different items. Readers interested in the details of IRT can refer to Hambleton, Swaminathan, and Rogers (1991) or Thissen and Steinberg (2020). We classify this method of scoring learners and placing them on a scale as norm-referenced, because to define the IRT scale, either the mean of the test taker population, or the mean of the item difficulties,

is arbitrarily set to a specific value, such as zero. However, because IRT locates each test item on the same scale as the test takers, actual items can be used to illustrate the knowledge and skills measured at specific points along the scale. This way of incorporating meaning into the score scale is referred to as *item mapping*. An illustration of item mapping is provided in Figure 2 (from Forsyth, 1998). This item map was used to help explain students' performance on the National Assessment of Educational Progress (NAEP) Reading assessment. Using the content standards measured by the items and the items' locations on the IRT scale, descriptions of the knowledge and skills students have at different points along the scale can be communicated. This procedure is suitable for large-scale tests because of the statistical requirements for IRT scaling. Where applicable, it can be used to understand skills students have mastered, skills they need to work on to achieve mastery, and subsequent skills they should work on next. Thus, although it is primarily a norm-referenced scaling procedure, IRT can be used to provide criterion-referenced information. IRT is also the underpinning of other scaling models, discussed in a later section of this chapter, that focus on providing criterion-referenced information.

Criterion-referenced testing does not involve relating learners' performance on a test to a mean or any other aspect of the population of learners. Instead, the information reported about performance on a test is referenced to the knowledge and skill domain targeted by the test. Criterion-referenced interpretations of test performance include statements like "Carlos mastered 82%" of the material on manipulating fractions" or "Yue achieved "proficient" status on the Grade 4 English Language Arts test." Unlike norm-referenced testing, in criterion-referenced testing, how other people performed on the test is not relevant to the interpretation of any one person's performance.

Figure 2.
Illustration of Item Mapping for Interpreting Students' Performance (from the National Assessment of Educational Progress Reading Test)



Source: Forsyth (1998).

Today, many criterion-referenced tests report learners' results using *achievement levels*. For example, the statewide summative (i.e., end of year summary) tests in Maryland classify students into one of the following four performance levels: beginning learner, developing learner, proficient learner, or distinguished learner. The "proficient learner" level represents what the state considers to be proficient for that grade level[3]. Theoretically, all students could be placed in the proficient, or any other, level. Other examples of achievement levels are "basic," "proficient," and "advanced," which are used on the NAEP (also known as the Nation's Report Card; see Loomis & Bourque, 2001).

The process of reporting learners' results in terms of achievement levels involves determining a level of performance on the test that is associated with each level. This process is described as setting "cut-scores" on the test or "standard-setting"; the latter term referring to the establishment of performance standards on the test associated with each achievement level (See Cizek & Ernest, 2016) for information on the process of standard-setting). The distinguishing feature of criterion-referenced information is it references learners' performance to a well-defined knowledge and skill domain, rather than to each other. Linn and Gronlund (2005) provide several guidelines for facilitating criterion-referenced information from educational tests. These guidelines focus on clearly defining the objectives of instruction and assessment, and ensuring the assessment provides sufficient information for determining whether students have mastered the objectives. They also suggest using item formats other than selected-response item formats (e.g., multiple-choice items) because students may correctly answer items by guessing.

It should be noted that tests designed to provide criterion-referenced information can also provide norm-referenced information. For example, in addition to reporting an achievement level for each student in Maryland, scale scores are also reported, and the mean performance of students in the school, district, and state is provided for parents, teachers, and others, to compare students' performance to these local and state averages.

3 The No Child Left Behind legislation and its extension the Every Student Succeeds Act requires all states receiving Federal funding to test students in reading, math, and science in several grades and to establish at least three achievement levels on each test, one of which must be "proficient" in that grade level.

The information provided by a test should match the intended purpose of the test. Therefore, the degree to which norm-referenced information from a criterion-referenced test is useful or holds substantive meaning would need to be established via research. There are also dangers, or potential negative effects, of reporting both norm-referenced and criterion-referenced types of information from educational tests. Norm-referenced information can reduce the academic self-concept of students who perform relatively worse than their peers, and some states use deficit-language in assigning achievement labels to students. For example, O'Donnell and Sireci (2021) reviewed all 50 states statewide summative testing score reporting practices and found that labels for the (same) lowest-performing level ranged from "inadequate" to "beginning learner." We were shocked to learn children were receiving score reports with such derogatory descriptions as "inadequate" and we support the work of Maryland and other states to promote more asset-based score reporting practices, which is a topic we return to in a subsequent section.

Table 1 summarizes the more traditional reporting practices for current educational tests. This summary provides a helpful baseline for us to compare the more innovative practices that are specifically designed to use educational assessments to promote learning.

Table 1.
Summary of Traditional Score Reporting Metrics for Educational Tests

Information Type	Reporting Metric	Explanation and Examples
Norm-referenced	Scale score	Distance of a student's score from norm group mean group in "standard deviation units" (e.g., Stanines (1–9 scale; mean=5); T-Scores (20–80 scale; mean=50)
	Percentiles	Percentage of students in norm group who score at or below student's score
	Age equivalent	Age at which the test score is "average" (e.g., 8.1= test performance is that of the average for a student who is a little over 8 years old).
	Grade equivalent (GE)	Grade in which the test score is "average" (e.g., a GE score of 4.5 is the average score of a 4th grader in the 5th month of school).
Criterion-referenced	Percent correct	Total points earned on a test divided by max possible points (e.g., 5-item test, each item worth 1 point, 4 items correct = 80%).
	Mastery score	A specific score signifying a student is proficient in the material tested.
	Achievement levels (Performance Classifications)	Specific levels of performance that describe various levels of mastery/proficiency (e.g., Pass/Fail; Basic, Proficient, Advanced)
	Subscores	Raw, percent correct, or scale scores for specific skill or content areas (e.g., persuasive writing; computations, applications, etc.)
	Item performance	Presents the items students took, the students' response, and the answer/scoring rubric.
Other	Gain scores	Difference between test scores taken at different time points (e.g., Spring score–fall score)

New Ways of Reporting Assessment Results to Help Learners Learn

The widespread use of commercial educational assessments that report a single score has made it clear that reporting a single score is ineffective in supporting student learning. With a single overall test score the actions an educator can take are limited and not particularly effective. For example, if a student has a low math score, a teacher can (a) continue to teach to the middle of the class and watch that student fall further and further behind, (b) provide that student with extra help, but with no knowledge of what lessons need to be reviewed, or (c) have the student repeat the grade. But as stated by mystery author Rita Mae Brown (1983, p. 68), though usually misattributed to Albert Einstein, "Insanity is doing the same thing over and over again and expecting different results." If a teacher does not know what it is that the student does not know, how can they efficiently and effectively remediate? Too often the primary outcome is disengaging students and reinforcing their belief that they cannot learn. Thus, there is a clear call for providing more instructionally valuable information from educational assessments.

First Came Subscores

The initial response to the need for more actionable educational test scores was subscores: separating out sets of items based on a test's content specifications and creating scores based on only those items (See Table 1). This separation can be done in several ways. For example, the ACT provides a composite scaled score based on all test items, but also provides math, science, English, and reading scores based on non-overlapping subsets of items as well as STEM and ELA scores that are based on combined item subsets.

One problem with subscores is they are based on a smaller number of items than total scores, and thus are less *reliable* (reliability refers to the consistency of test results). Another problem is they tend to correlate highly with each other, making it difficult to interpret score profiles (i.e., to distinguish between the different types of information provided). This correlation is magnified should some of the same items appear in multiple subscores. To minimize overinterpretation, some standards-based testing programs collapse score scale results to categories such as "below standard," "near standard," and "above standard." As discussed earlier another similar approach is to report subscores as either illustrating mastery or not mastery of the materials.

Backward Design to the Rescue

As an alternative to building a test and then deconstructing it to try to produce actionable subscores, *one could design a test with instruction in mind*. This alternative might be done by first designing the score report considered most likely to help teachers better teach students, and only then designing the test that would support that score report. Given that a goal of assessment in the service of learning is to help students learn and teachers teach, then perhaps an overall scaled score is not necessary. Perhaps teachers just need to know what each student has and has not yet mastered. Perhaps there might be a psychometric approach designed to do just that–and there is. We describe that approach next.

Diagnostic Classification Modeling

What is now usually referred to as diagnostic classification models (DCMs), but has also been referred to as cognitive diagnostic assessment, has much of its roots in the mid-20th century with the development of latent class models (Lazarsfeld, 1950; Lazarfeld & Henry, 1968). These models were initially used to identify unobserved subgroups within a population based on observed data. The concept of latent traits, which are not directly measurable, but can be inferred from patterns of responses, laid the groundwork for DCMs. Tatsuoka (1983) and Embretson (1984) specifically looked at decomposing cognitive processes within an IRT model to provide diagnostic information, which led to modern DCM.

DCMs provide a detailed analysis of students' knowledge and skills. Unlike traditional assessments that yield a single overall score, DCMs offer a nuanced profile of a student's strengths and weaknesses across multiple attributes or skills. This multi-dimensional approach allows educators to understand not just whether a student can solve a problem, but also which specific skills they have mastered or need to improve. By categorizing students into mastery or non-mastery for each skill, DCMs provide valuable diagnostic feedback that can be used to tailor instruction and interventions to meet individual learning needs. Figure 4, later in this chapter, provides an example of a score report consistent with a DCM approach.

The application of DCMs is particularly beneficial in formative assessments, where the goal is to monitor student learning and provide ongoing feedback. Teachers can use detailed diagnostic information to adjust their teaching strategies and provide targeted support. In large-scale assessments, such as state or national exams, DCMs offer a comprehensive picture of student performance across different

regions or demographics, informing policy decisions and resource allocation. Additionally, DCMs enhance adaptive testing environments by ensuring that the adaptive algorithm considers multiple skills simultaneously, providing a more accurate measure of student abilities.

Technically, DCMs are based on the assumption that the underlying traits being measured are categorical rather than continuous. This means students are classified into categories such as "mastery" or "non-mastery" for each skill. DCM models use a *Q-matrix* to define the relationship between test items and the underlying skills or attributes. The Q-matrix is a binary matrix that specifies which skills are required to correctly answer each test item. This matrix is crucial for the accurate estimation of students' skill profiles. The estimation process involves complex algorithms that can handle the multivariate nature of the data, often employing techniques from Bayesian statistics to provide robust and reliable classifications (Rupp, Templin, & Hanson, 2010).

Despite their advantages, implementing DCMs can be complex, requiring sophisticated statistical techniques and software. Educators and administrators need training to interpret and use the results effectively. Ensuring the validity and reliability of DCM-based assessments is crucial, involving rigorous testing and validation processes (Thompson, Clark, & Nash, 2021). Moreover, DCMs require detailed data on student performance across multiple items and skills, which can be resource-intensive to collect and manage. Nevertheless, the detailed insights provided by DCMs enable more personalized and effective educational interventions, ultimately supporting better learning outcomes. By offering a comprehensive profile of student abilities, DCMs represent a significant advancement in educational assessment, moving beyond traditional scoring methods to support more informed and targeted educational practices.

DCMs can be based on many approaches that identify skills and their relationship to mastery classifications as long as the result is a Q-matrix that captures those relationships. Skills can be initially identified in a single list or a hierarchy. Another, more recent, approach to identifying skills can be described as *learning maps*, which we discuss next.

Learning maps

Learning maps are one form of an organized learning model (Kingston et al., 2022) They are visual representations that illustrate the relationships among various knowledge, skills, and understandings within a subject area. They comprise an directed acyclic graph consisting of interconnected nodes, each representing a specific concept or skill. These nodes are probabilistically linked with the probability of having mastered a successor node conditionally dependent on having mastered one or more precursor nodes. Thus, the pathways among nodes show how different pieces of knowledge are related and how they build upon each other. Learning maps help educators and learners see the broader context of what is being learned and identify the pathways through which any individual student can best progress in their understanding.

Figure 3 shows a version (the learning map is updated regularly) of the Dynamic Learning Maps Alternate Assessment (DLMAA) mathematics map, which at the time the map was captured consisted of 2,554 nodes and 5,605 connections ranging from developmental infancy (recognizes an object) to advanced high school mathematics. With a map this large it should not be surprising that only the overall structure is captured in the figure. Those interested in how the learning maps for the DLMAA were developed can find more information in Bechard, et al. (2019).

Figure 3.
DLMAA Mathematics Learning Map

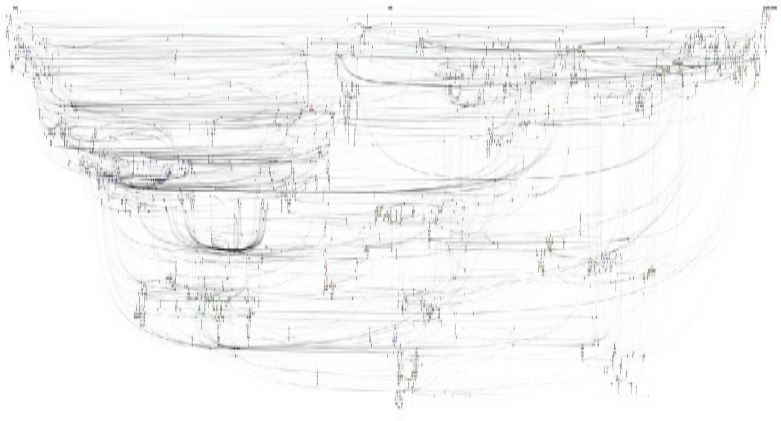
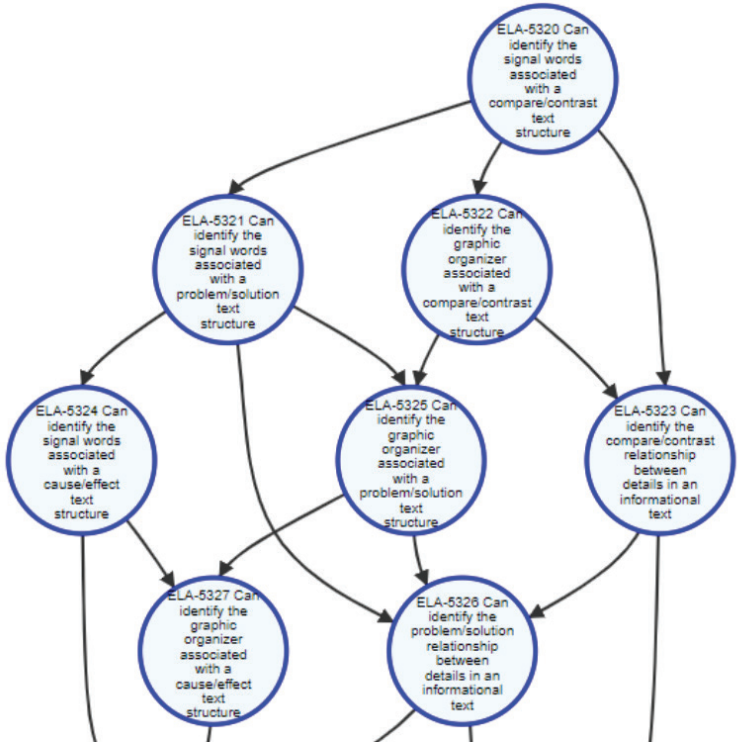


Figure 4 shows a section of an English Language Arts learning map that was developed as part of DLMAA. DLMAA will be described further in the next section of this chapter.

Figure 4.
Section of the DLMAA English Language Arts Learning Map



Because learning maps are directed acyclic graphs, they are amenable to diagnostic classification modeling including, but not limited to, an approach known as Bayesian network analysis. Bayesian networks are particularly useful for modeling complex systems where variables interact in uncertain ways, such as the human mind. They allow for efficient computation of the probabilities of

various outcomes given certain evidence, making them valuable for tasks such as diagnostic reasoning, prediction, and decision-making under uncertainty, such as which content to probe to best understand what a student knows and can do. Bayesian networks can incorporate both prior knowledge and new observed data to update inferences made from assessment data. In the next section, we turn to an example of learning maps applied to a specific situation—alternative assessments for cognitively impaired students.

Dynamic Learning Maps Alternate Assessment

The Dynamic Learning Maps Alternate Assessment (DLMAA) was designed to measure educational progress for students with significant cognitive disabilities. DLMAA was designed with different priorities than other accountability assessments. As stated by Kingston et al. (2014, p. 5), "For the world of educational assessment to better serve students with significant cognitive disabilities, we must begin with a goal for large-scale assessment that helps students learn." To accomplish this goal, six high-level features were chosen to guide the development of the DLM: "(1) fine-grained learning maps that guide instruction and assessment, (2) a subset of particularly important nodes that serve as content standards to provide an organizational structure for teachers, (3) instructionally embedded assessments that reinforce the primacy of instruction, (4) instructionally relevant testlets that model good instruction and reinforce learning, (5) accessibility by design (vs. accommodation) and alternate testlets, and (6) status and growth reporting that is readily actionable."

Learning maps play a crucial role in evaluating the educational progress of students with significant cognitive disabilities. The DLM system uses these maps to plot out individual concepts in nodes, showing the multiple ways that students' knowledge, skills, and understandings develop over time. This model helps educators uncover reasons why a student may be struggling with a particular concept and find possible pathways for students to expand their knowledge and skills.

The DLM assessments are designed to be adaptive and accessible, providing a more accurate measure of student proficiencies by adjusting the difficulty of questions based on the student's responses. The learning map model is integral to this process, as it ensures the assessment is aligned with the specific learning needs and proficiencies of each student. By mapping out the connections between different concepts, the DLM system helps educators identify the most effective

instructional strategies and provide targeted support to enhance student learning outcomes. This approach not only supports better academic performance but also helps in developing personalized educational plans that address the unique needs of each student.

DLMAA score reports

DLMAA reports scores by identifying precursor and successor nodes related to specific educational standards (essential elements in DLM parlance). These nodes are expressed as five levels in an extracted learning progression, illustrated in Figure 5.

Figure 5.
Section of DLMAA Learning Profile

Student's performance in 7th grade English language arts Essential Elements is summarized below. This information is based on all of the DLM tests Student took during Spring 2022. Student was assessed on 13 out of 13 Essential Elements and 4 out of 4 Areas expected in 7th grade.

Demonstrating mastery of a Level during the assessment assumes mastery of all prior Levels in the Essential Element. This table describes what skills your child demonstrated in the assessment and how those skills compare to grade level expectations.

		Estimated Mastery Level				
Area	Essential Element	1	2	3	4 (Target)	5
ELA.C1.1	ELA.EE.RI.7.5	Understand the functions of objects	Identify concrete details in an informational text	Recognize how titles reflect text structure and text purpose	Understand sequencing	Understand how parts of the text affect overall text structure
ELA.C1.2	ELA.EE.RL.7.1	Differentiate between text and pictures	Identify characters, setting, and major events	Identify words that answer explicit questions	Identify where explicit information is stated and where inferences can be drawn	Identify explicit and implicit information
ELA.C1.2	ELA.EE.RL.7.4	Understand words for absent objects and people	Identify definition of words explicitly defined in a sentence	Identify word meaning of multiple-meaning words using context clues	Determine the meaning of idioms and figures of speech	Determine the connotative meaning of words and phrases
ELA.C1.2	ELA.EE.RI.7.2	Match a picture representation with a real object	Identify concrete details in an informational text	Identify the implicit main idea in an informational text	Identify multiple main ideas in an informational text	Summarize a familiar informative text

Levels mastered this year | No evidence of mastery on this Essential Element | Essential Element not tested

This report is intended to serve as one source of evidence in an instructional planning process. Results are based only on item responses from the end of year spring assessment. Because your child may demonstrate knowledge and skills differently across settings, the estimated mastery results shown here may not fully represent what your child knows and can do. For more information, including resources, please visit https://dynamiclearningmaps.org/states.

Page 3 of 5

The score reporting for DLM illustrates one current way for providing actionable information to help students learn. In the next and final section of our chapter, we discuss additional ideas for providing information to support student learning, as well as a research agenda for making progress toward this goal.

Criterion-Referenced Scales

In addition to using DCM to indicate the learning progressions of students, other research has focused on modeling the difficulty (challenge) presented by items and incorporating that understanding into the score scale. For example, Embretson (1993), Sheehan and Mislevy (1990), and other researchers (e.g., Fischer, 1973) have illustrated how the content, cognitive, and other "complexity" features of items can be used to understand and model the degree to which items challenge learners. These researchers have focused on relating item attributes, such as the number and type of calculations required to solve a math problem or the linguistic complexity of a reading passage, to the construct measured. In a sense, this work extends the item mapping described earlier (See Figure 2) to not only locate items on the scale, but to also quantify the specific knowledge and skills required to successfully answer items at different points along the scale. This research suggests the coding of content, cognitive, and other item features can be used to build a "complexity scale" to indicate where students are on the continuum of knowledge and skill measured by a test. The potential benefit of these complexity scales is the scale score reflects the complexity of the challenge presented by the items as they are located along the scale, rather than by how far students are from one another with respect to performance on a set of items. Quantifying score scales in this manner entirely removes norm-referencing from the scaling process and aligns students' performance with the cognitive challenges presented by the items (Feng et al., 2024). Future research is needed to evaluate whether reporting scores in this way will reveal more about the cognitive processes used by students in responding to questions and whether this information will support their learning.

Developing Educational Tests to Support Student Learning: The Path Forward

The inspiration and champion of the Handbook in which this chapter is written, Edmund W. Gordon, has long criticized educational tests as focusing on the status of learning rather than the process of learning. That is, educational tests tend to reflect *whether* students have learned, but have not been good at showing *how* they have learned. We believe the demonstration of performance along a criterion-referenced scale (e.g., item mapping) and subscore reporting based on learning progressions provide helpful information to support student learning, with the latter coming closer to the process of learning. However, to meet the challenge provided

by Gordon (2013, 2020), perhaps our assessing and reporting mechanisms should focus on representing processes, rather than only representing content.

As illustrated in Figure 4, the cognitive process of "identify" is a feature in the DCM for the learning map. Because DCMs map out a progression of skills, they have the potential to also map out processes, which could be a potential way to use assessments to provide information regarding the processes learners use to respond to questions, as well as the determining the likely success of different processes for individual learners. By understanding the processes learners used to answer questions, educators can "redirect" learners' to more effective processes or target the most important cognitive skills learners need to support their learning.

Another important area in need of further work is gathering empirical evidence on specific features of different approaches to assessment for learning, and the effectiveness of those features with the needs of different kinds of learners. Specifically, more research is needed on the best types of feedback to provide regarding student performance. In a meta-analysis of the effect of feedback on student learning (based on 435 studies of over 61,000 learners), Wisniewski, Zierer, and Hattie (2020) found a weighted mean effect size of 0.48, indicating a strong and positive effect of feedback. However, they also noted substantial variability due to context and type of feedback (e.g., reinforcement or punishment, corrective feedback, or high information feedback) and quality of feedback. Direction of feedback (teacher-to-student, student-to-teacher, or student-to-student) had only a moderate explanatory effect. Also, and not unexpectedly, there appears to be a publication bias in the published articles,[4] but none in the 116 dissertations. Dissertations showed a lower effect size of 0.36.

In short, while there is little question that overall feedback makes a positive difference, there is insufficient evidence regarding what *type* of feedback systems work best for which students. Various feedback models (e.g., Butler & Winne, 1995; Hattie & Timperley 2007) and their features need to be validated within educational assessment methods and systems. It is not enough to say tests should provide useful feedback; examples and strategies for providing useful feedback must be provided.

4 Peer-reviewed journals are notorious for their reluctance to publish non-significant findings, which biases the results of systematic reviews.

Another development in psychometrics that deserves more attention for targeting the processes learners use in responding to test items and using that information to serve learners is validity evidence based on response processes. The use of evidence regarding the cognitive processes students use to solve test items was largely championed by Messick (1989) and codified into the last two versions of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999; 2014). This source of validity evidence is used to support the use of a test for a particular purpose by probing whether learners use the intended cognitive processes when responding to items. Examples of such evidence come from focus groups, think-aloud studies, and cognitive interviews where test takers indicate the strategies they used in responding to items and report the thoughts and emotions they had while doing so (Padilla & Benítez, 2014). It would be interesting to not only use this information as validity evidence, but also as evidence to be reported to learners and educators as part of the results of an assessment.

Recently the analysis of "log data" from educational assessments has also been used to provide validity evidence based on response processes. Log data refers to the data a computer or other digital device captures as a test taker responds to assessment items. Like other methods for probing learners' response processes, these data are used in validity studies to confirm the intended cognitive processes are measured on an assessment, to check whether students are motivated to try their best when taking an assessment, and even as part of scoring an assessment. For example, He and von Davier (2016) analyzed log data from the computer-based Programme for International Assessment of Adult Competencies (PIAAC) assessment to understand how test takers solved problems on the assessment. Similarly, He at al. (2021) analyzed log data from a computerized test of problem solving and found it could be used to determine how far the observed processes used by test takers deviated from optimal problem-solving solutions. As another example, Wise et al. (2021) used the amount of time learners take to respond to test items as a measure of their engagement with the test. Chung et al. (this volume) provide other examples of how log data can be leveraged to understand more about the cognitive processes students use to solve test items. Clearly, these developments illustrate how log data from digital assessments can isolate and report the processes learners use. Access to this information can empower both learners and educators to adjust problem-solving behaviors in ways that maximize

success. Reporting information on engagement will also likely be useful for interpreting learners' test performance and for improving assessments that show low levels of engagement.

Another important area of the path forward is ensuring that the primary consumers of the results of educational tests—learners and their educators—clearly understand the results of an assessment. As O'Leary et al. (2017) pointed out,

> *score reports…are fundamental to the process of communication between test developers and their audience. As such, the interpretability of score reports (i.e., how well members of the intended audience are actually interpreting and using scores as reported) is of the utmost significance and is fundamental in claims about validity.* (p. 21)

Clearly, providing information on learners' performance on educational assessments must be done in a way that is understandable to learners and educators. In this chapter, we have argued that we must go beyond traditional score reporting practices to provide actionable information on both the status and process of learning. In keeping with the points of O'Leary et al. (2016), even the best intended extensions will be invalid if the audiences they are intended for cannot interpret them.

As a last suggestion for the path forward we provide an additional thought. If educational assessment in the service of learners requires that "Feedback, adaptation, and other relevant instruction should be linked to assessment experiences" (Principle 5), perhaps the results from educational tests should not so much focus on metrics, but focus on how they can best facilitate *conversations* between learners, teachers, and others. Perhaps even the results can facilitate conversations between learners and test developers—maybe even psychometricians! As described in Heritage and Kingston (2019) there has long been a divide between the approaches to assessment of classroom teachers and psychometricians. More and broader conversations are likely to lead to better tests that do even more to serve learners.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Psychological Association.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Educational Testing Service. (Reprint of chapter In R. L. Thorndike (Ed.). *Educational Measurement* (2nd Edition), American Council on Education, 1971).

Bechard, S., Clark, A., Swinburne Romine, R., Karvonen, M., Kingston, N., & Erickson, K. (2019). Use of Evidence-Centered Design to Develop Learning Maps-Based Assessments. *International Journal of Testing*, *19*(2), 188–205. https://doi.org/10.1080/15305058.2018.1543310

Brown, R. M. (1983). *Sudden Death.* Bantam Books.

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Cizek, G. J., & Earnest, D. S. (2016). Setting performance standards on tests. In S. Lane, T. Haladyna, & M. Raymond (Eds.), *Handbook of test development* (pp. 212–237). National Council on Measurement in Education.

Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S. E. (1984). *A general latent trait model for response processes. Psychometrika, 49,* 175–186.

Feng, W., Tran, P., McNichols, W., Sireci, S. G., & Lan, A. (2023, October). *Using artificial intelligence to scale multiple-choice mathematic items.* Paper presented at the annual meeting of the Northeastern Educational Research Association.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Forsyth, R. A. (1998). NAEP frameworks and achievement levels. In M. L. Bourque (Ed.), *Proceedings of the Achievement Levels Workshop.* National Assessment Governing Board.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17,* 145–221.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81–112. https://doi.org/10.3102/003465430298487

He, Q., Borgonovi, F., & Paccagnellac, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education, 166*. 104170

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development*, (pp. 750–777). Information Science Reference.

Heritage, M., & Kingston, N. M. (2019). Classroom Assessment and Large-Scale Psychometrics: Shall the Twain Meet? (a conversation with Margaret Heritage and Neal Kingston). *Journal of Educational Measurement, 56*(4), 670–685.

Kingston, N. M., Alonzo, A. C., Long, H., & Swinburne Romine, R. (2022). Editorial: The use of organized learning models in assessment. *Frontiers in Education, 7*, 1009446. https://doi.org/10.3389/feduc.2022.1009446

Kingston, N. M; Clark, A. K., Pardos, Z., & Lee, S. Y. (April 2014). *Determining a Reasonable Starting Place for an Instructionally Embedded Dynamic Assessment: Heuristic versus Bayesian Network Analysis.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155–186), Praeger.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis & the interpretation and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–472). Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Houghton Mifflin.

Linn, R. L., & Gronlund, N. E. (2005). *Measurement and assessment in teaching (9th edition).* Upper Saddle River, NJ: Pearson Prentice-Hall.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard-setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives* (pp. 175–217). Mahwah, NJ: Erlbaum.

Malda, M., van de Vijver, F. J. R., & Tamane, Q. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38,* 582–595.

Messick, S. (1989b). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). American Council on Education.

O'Donnell, F., & Sireci, S. G. (2021). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment.* https://doi.org/10.1080/10627197.2021.2016388

O'Leary, T. M., Hattie, J. A. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues and Practice, 36(2),* 16–23.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford.

Ryan, J. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T.M. Haladyna (Eds.), *Handbook of Testing* (pp. 677–710). Lawrence Erlbaum.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement, 27*, 255–272.

Sireci, S. G., & Randall, J. (2021). Evolving notions of fairness in testing in the United States. In M. Bunch & B. Clauser (Eds.), *The history of educational measurement: Key advancements in theory, policy, and practice* (pp. 111–135). Routledge

Sireci, S. G., Wainer, H., & Braun, H. (1998). Psychometrics, overview. In *Encyclopedia of biostatistics*. John Wiley & Sons.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Thissen, D., & Steinberg, L. (2020). An intellectual history of parametric item response theory models in the twentieth century. *Chinese/English Journal of Educational Measurement and Evaluation, 1,*(1), 23–39. https://doi.org/10.59863/GPML7603

Thompson, W. J., Clark, A. K., & Nash, B. (April 2021). *Technical Evidence for Diagnostic Assessments*. Paper presented at the annual meeting of the National Council for Measurement in Education (virtual conference).

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, *26*(3), 163–174. https://doi.org/10.1080/10627197.2021.1956897

Zenisky, A. L., & Hambleton, R. K. (2015). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, and T. Haladyna (Eds.), *Handbook of test development* (2nd ed.), pp. 585–602. Routledge.