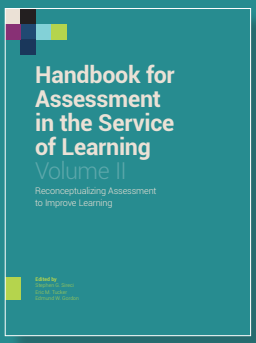
Validity for Assessments Intended to Serve Learners

Stephen G. Sireci and Danielle Crabtree

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., & Crabtree, D. (2025). Validity for assessments intended to serve learners. 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.

Validity for Assessments Intended to Serve Learners

Stephen G. Sireci and Danielle Crabtree

Abstract

Validity has been described as the most important characteristic in educational testing; yet many practitioners are unfamiliar with the terminology and what needs to be done to ensure tests are fulfilling their intended purposes. In this chapter, we define validity and describe the importance of validating educational assessments designed to serve learners. We discuss how a focus on validity and the practice of validation can help assessments in the service of learning accomplish their goals. Such validation involves collaborating with educators, parents, guardians, and students to design assessments that provide value to them and support learning. We point out such collaborations are typically not incorporated into test design and validation, but are essential for educational assessments to reach their potential in serving learners.

Validity for Assessments Intended to Serve Learners

There are many associations of the word *validity*. One might think of a valid argument, which is one that is well reasoned and conclusive because it is based on sound logic and evidence. Others might think of the legitimacy of a certain claim, or of the facts supporting a point someone made in a dispute. These understandings of validity are similar to how the term is used in educational testing. Given that tests are developed and used for specific purposes, the degree to which those purposes are fulfilled is important to evaluate and document. Equally important is evaluating and documenting the degree to which there are unintended side effects associated with the testing process.

Educational assessments designed to serve learners have a common, general purpose to support student learning. Examples of more specific purposes include engaging students in their learning, providing information to students and their teachers about how they learn, reflecting students' strengths, building their academic self-efficacy, and indicating areas in need of improvement. The goals of assessment in the service of learning are consistent with the goals of instruction in the service of learning. Hence, these "educative assessments" can be thought of as pedagogical tools. However, for assessments to truly serve learners they must not only demonstrate evidence of such service, they must abide by *Principles of Assessment in the Service of Learning* (Baker et al., 2025; Volume I of this *Handbook* series) that require assessment practices to be transparent to stakeholders, grounded in high-quality evidence, and geared toward feedback that improves teaching and learning.

Many of the chapters in this *Handbook* describe theories and methods for designing and conducting assessment in the service of learning. One may ask, "Do these theories and methods work?" "Do these assessments actually serve learners?" These evaluative questions are the focus of the present chapter because in educational testing, investigations of "validity" encompass the evaluation of the effectiveness, utility, and consequences of an assessment. In this chapter, we first define validity and the process of *validation*. We then discuss how a focus on validity and the practice of validation can help assessments in the service of learning conspicuously accomplish their goals. Elevating the voices of educators, parents, guardians, and students is a key feature in designing and validating assessments in the service of learning. We believe these voices have been left out

of many prior validation efforts, which has seriously undermined the degree to which educational assessments have reached their potential to serve learners.

What is Validity?

In educational assessment, *validity refers to the degree to which evidence and theory support the use of a test for a particular purpose.* Evidence refers to the results from strategic and comprehensive research designed to evaluate how well a test is fulfilling its intended purposes and avoiding any unintended, harmful consequences. Theory refers to the conceptualizations describing what the test is measuring, and the logic describing how the interpretations and use of test results will fulfill the intended purposes.

This definition implies several attributes of validity in educational testing. Specifically,

- Validity is not a characteristic of a test, but rather a characteristic of the interpretations, actions, and outcomes associated with a testing process.
- The purposes and intended uses of a test must be clearly specified, because they communicate the testing claims to stakeholders, and set the targets for validation.
- If test results are to be used for multiple purposes, each use must be supported by validity evidence.

Note that the second and third bullets build transparency in testing through a shared understanding between those who create or commission the tests, and those who use them. That shared understanding relates to the purpose of the assessment, the specific knowledge and skills being tested, and the intended use of test results and how such use will support learning.

Validity is an overarching standard against which all tests are evaluated. In addition to the attributes described thus far, it is also important to note validity evidence must include both evidence confirming the appropriateness of the use of the test for its intended purposes and *evidence that use of the test does not result in unintended harmful consequences*. This latter point is especially important for assessments in the service of learning because their explicit purposes are to

evoke only beneficial consequences such as increased learning, increased love of learning, and affirming the learners' enormous capacity to learn.

Validity is sometimes quickly joined to another component of assessment quality—reliability. Reliability refers to the consistency of test scores, that is, if students were tested over and over again, would they get the same test scores—just as if we step on a bathroom scale over and over again we would get the same reading of our weight. Reliability is related to validity in that if scores from an assessment are not reliable, that means students can get very different scores because the test itself or the testing process is unstable. Thus, consistency in the information provided by an assessment (i.e., reliability) is a prerequisite for validity (for more on reliability, see Bandalos, 2018).

Before further describing validity with respect to tests intended to support learners, we discuss the process evaluating educational tests, which is called *validation,* or a *validity investigation*.

What is Validation?

Validation is the process of evaluating the validity of the use of a test for a particular purpose. It involves carefully planned and ongoing research for critiquing the theory underlying test development and for gathering validity evidence. Validation research begins with the earliest stages of test development (e.g., defining the testing purpose and the targets of measurement[1]) and continues throughout all phases of the lifecycle of a testing program. The goal of validation is to provide a sound basis for interpreting and using test scores and other information resulting from the testing process. In addition, validation research can be, and should be, used to improve tests and testing processes whenever the evidence suggests improvements are warranted.

In a subsequent section we return to the topic of validation to indicate how evidence can be gathered to evaluate the degree to which a test designed to support learners accomplishes its intended goals. First, however, we discuss how those goals should be articulated, and include some examples of specific purposes of assessments in the service of learning

1 As we subsequently describe, the targets of measurement are often referred to as "constructs," because theories are *constructed* about the unobservable processes that explain test takers' thoughts and behaviors when responding to test items (Sireci, 2020).

Goals, Purposes, and Uses of Assessments in the Service of Learning

The common goal of all assessments for the service of learning is to benefit the learner. Learners can gain direct benefits and indirect benefits, depending on the purpose and use of the assessment. Direct benefits are those the learner experiences before, during, and after the administration of the test. Before the assessment, the learner may take the time they need to study, to inform themselves of the assessment's construct(s), and to hold themselves accountable for their own learning. In this way, the assessment is viewed as an opportunity for development and becomes a crucial part of the learning process. During administration, the learner may benefit from the assessment metacognitively, gaining instantaneous insight into the aspects of the construct(s) they know well, the aspects that still remain uncertain, and the aspects they have yet to learn. After administration, learners gain the benefits of feedback and reflection, allowing them to self-evaluate their learning, to continue to grapple with the concepts they are still trying to comprehend, and providing them control over their own future learning. However, for learners to reap the benefits of reflection, feedback should include free access to the assessment, time to process how the assessment shaped their learning, and clarifying support for confusing concepts.

Indirect benefits are those learners experience due to the assessment-based decisions made by other members of the school community. A common example is a diagnostic assessment triggering an intervention for a learner. In this case, the learner is not necessarily part of the decision-making process (though, they could be), but improvement to their learning is the focal point for all decisions being made. Effective assessments for the service of learning allow educators to pinpoint the learner's misconceptions, provide information that enables the educator to develop personalized instruction to fit the learner's needs, and allow for progress monitoring, to ensure the skills the learner is supposedly developing are actually being developed. Thus, a first step in ensuring and evaluating the validity of an assessment in the service of learning is to clearly state the direct and indirect benefits. These benefits are the goals of the assessment process and so in validation, we are required to look for evidence that they are being realized.

Although the overall goal is to benefit the learner, the purpose of the assessments in the service of learning is to ascertain specific and insightful information into the learner's educational growth and progress. The type of information obtained and

the depth of insight is dependent on the type of assessment used to gather the information. An extensive list of common and emerging models for assessments in the service of learning have been compiled by Gordon (2020). Each model promotes its own purpose so it is important that the purpose for any assessment is transparent and explicitly stated for all those involved in the testing. Therefore, validation of the assessment requires evidence that the stated purpose is in alignment with its intended goals, use, and consequences.

The basic use of assessments in the service of learning is to inform instruction and learning. Assessments as pedagogical tools act as catalysts between instruction and learning, providing the momentum necessary for educational growth. Yet, it is important to remember assessments are just snapshots of time, measuring the learner's knowledge at that moment for a specific purpose. The purpose, administration, and timing of these snapshots are important. Traditionally, assessments used before instruction measure prior knowledge, assessments used during instruction measure the learner's current understanding, and assessments used after instruction are used to measure retention. These traditional approaches to assessment are evaluative, closed-book, and are used to measure "the beginning" or "the end" of learning. Yet, learning is continuous and so the assessments meant to propel learning forward must also be continuous. That is not to say that learners need to be caught in a constant cycle of evaluation. Rather, assessments in the service of learning are used to promote continuous reflection, allowing one to assess what has been learned and what still needs to be learned at a time and in a way that is most informative. This means the purpose, administration, and timing of the assessment is done in collaboration between the learner and those who are providing the instruction.

Moreover, the use of assessments in the service of learning involves planned intended consequences. If an assessment is used to inform instructional decisions, then the consequences of those instructional decisions must also be considered and stated explicitly prior to the assessment. To be clear, consequences in this circumstance do not imply positive or negative outcomes (which are subjective), but rather the consequences that are necessary to benefit the learner. Additionally, planned consequences do not only pertain to the learner. For example, a change in instructor may be deemed an appropriate consequence that would benefit the learner. Once again, planning intended consequences should be a collaborative effort, ensuring all stakeholders are in agreement that the consequences align

with the appropriate instructional decisions, and are for the learner's best interest, as supported by the evidence collected through validation. As clearly stated in the *Principles for Assessment in the Service of Learning* (Baker et al., 2025), helpful and supportive feedback is a critical intended consequence for assessments to serve learners. The next steps suggested by feedback should help support the intended consequence of learning, including the socioemotional aspects such as improved academic self concept.

Validating Assessments in the Service of Learning

Our prior discussion of how we can articulate the purposes of specific assessments designed to serve learning has set us up for discussing how we can evaluate their intended and unintended consequences. The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014) specifies five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). These five sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) testing consequences. We believe these five sources are helpful in evaluating assessments in the service of learning, but we add an additional source of validity evidence that comes from engaging with traditionally forgotten stakeholders in the assessment process—teachers, students, parents, and guardians. Borrowing from O'Leary et al. (2017), we refer to this source of validity evidence as validity evidence based on analysis of test score interpretations.[2]

Sources of Validity Evidence

The word "evidence" has been used numerous times in our discussion of validity and test validation. In this section, we describe specific sources of validity evidence that are helpful for building a validity argument and for demonstrating assessments are of sufficient quality for fulfilling their intended purposes. These sources have been described in the past as different types of validity, but characterizing them in that way is a misnomer. As described earlier, validity refers to the degree to which evidence and theory support the use of a test for a particular purpose. Thus, we cannot divide this concept into types or subtypes. However, we can distinguish

2 At the time of this writing, the AERA et al. Standards are under revision and it is likely this sixth source of validity evidence may be added into the next version.

among the different sources of evidence that can be used to evaluate validity, and represent important aspects of the quality of an educational assessment.

What follows are general descriptions of six sources of evidence that may be useful for evaluating the degree to which assessments designed to serve learners fulfill their goals and do not lead to unintended, negative consequences. These descriptions are intended to describe specific sources of validity evidence that have historically been used to develop validity arguments for testing programs. However, they are not considered representative of all potential sources, and one or more sources may not be applicable to a particular testing situation.

Validity evidence based on test content.
Validity evidence based on test content is used to confirm the content of the test is representative of the knowledge and skill domain the test is designed to measure and is consistent with the testing purpose. The content of a test includes directions, items (tasks), and the stimuli associated with items. In educational testing, validity evidence based on test content is most often gathered by using subject matter experts to review the targets of instruction (e.g., curriculum standards, objectives associated with a lesson, etc.) and to review the test specifications and test items. The subject matter experts should be external to the test development process so their appraisals can be considered independent and unbiased. These experts are trained, and asked to rate test items regarding their degree of alignment to intended objectives, to provide input regarding the degree to which the test represents the targeted knowledge and skill domain, and to evaluate the degree to which students' responses to test items will provide the information intended within the testing purpose.

There are many methods for structuring experts' reviews of tests and items. Traditional methods based on content validity indices have been reviewed by Crocker et al. (1989) and Sireci (1998). These methods typically require experts to indicate the content areas and cognitive levels measured by test items, or to rate how well test items measure specific objectives. Newer methods based on test-curriculum "alignment" evaluate the link between curriculum frameworks, testing, and instruction (Bhola et al., 2003; Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014).

Validity evidence based on test content is particularly relevant to assessment in the service of learning because these assessments should specifically target instructional goals. It is important teachers see the tasks and items that comprise an assessment as reflecting their goals. The tasks presented to learners on these assessments represent the content to be independently evaluated by subject matter experts, who should be experienced with the goals of instruction, and different from the authors of the assessment. These external subject matter experts need to be trained to provide the types of judgments most helpful for determining assessment-instruction alignment. For example, experts can rate the degree to which assessment tasks appropriately measure specific content or cognitive objectives, or capture common misconceptions. An example of a rating task presented to subject matter experts is presented in Figure 1. In this rating task, experts were informed of the content standard (College and Career Readiness Standard for Adult Education—CCRSAE) and cognitive level targeted by each assessment item. They were asked to review each item and complete the four ratings illustrated in Figure 1. These data can be aggregated across raters and across items to provide overall information regarding the content quality of the assessment with respect to instructional alignment. It should be noted the four questions illustrated in Figure 1 are just one example of how to gather data from subject matter experts regarding content quality, and other methods can be used that involve more open-ended questioning, or targeting larger grain size elements of knowledge and skills such as a lesson or curricular unit.

Figure 1.

Example of Item-Level Ratings Gathered to Provide Validity Evidence Based on Test Content

	Not well at all	Slightly well	Moderately well	Very well	Extremely well
How well does the item measure the CCRSAE Standard?	○	○	○	○	○
How well does the item measure the cognitive level?	○	○	○	○	○
How well does the item meet your standards for a high quality item?	○	○	○	○	○
How well do the distractors (incorrect response options) represent realistic mistakes?	○	○	○	○	○

Note: CCRSAE=College and Career Readiness Standards for Adult Education (Pimentel, 2013).

Other evaluation tasks could be presented to subject matter experts to have them gauge the degree to which the tasks presented on the assessment meet the test development targets and are likely to provide the intended information regarding student learning. The degree to which the assessment represents the instructional goals (content representation) and to which all items are relevant to those goals (content relevance) are important evaluation criteria for which data should be gathered. Many methods exist for gathering subject matter experts' perceptions of test content including those based on traditional indices of content validity (e.g., Crocker et al., 1989; Sireci & Faulkner-Bond, 2014) as well as those based on assessment-curriculum alignment (Bhola et al., 2003; Martone & Sireci, 2009). Minimally, teachers and other test developers should get outside opinions from peers regarding the content quality of their assessments intended to serve learners.

One further area in which validity evidence based on test content can be extremely important for assessments in the service of learning is ensuring the content of the test is "culturally responsive," meaning "assessments that take into account the background characteristics of the students; their beliefs, values, and ethics; their lived experiences; and everything that affects how they learn and behave and communicate" (Walker et al., 2023, p. 1). Interrogating test content with respect to the degree to which it is appropriate for the diversity of learners who will interact with the assessment can be conducted using subject matter experts who come from diverse backgrounds and are familiar with the various cultures from which the learners will draw from when interacting with test content. However, culturally responsive assessment is an emerging concept and likely requires more than a finite group of diverse subject matter experts; it will require engaging with test takers themselves and the communities within which they operate (Sireci et al., 2025; Walker et al., 2023).

Validity evidence based on response processes.

Validity evidence based on response processes refers to "evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by test takers" (AERA et al., 2014, p. 15). Such evidence can include interviewing test takers about their responses to test questions, systematic observations of test response behavior, evaluation of the criteria used by judges when scoring performance tasks, analysis of item response time data, and evaluation of the reasoning processes students use when solving test items (Embretson, 1983; Messick, 1989; Mislevy, 2009). Such evidence is helpful for confirming assessments in the service of learning are measuring the cognitive skills they intend to measure, and that students are using the targeted skills to respond to the test items.

Given that the overall goal of assessments in the service of learning is to benefit the learner, assessment must go beyond confirming what a learner "knows" or "does not yet know" and provide information on learners' cognitive processes in responding to test items. For example, if a learner consistently answers questions correctly regarding a specific learning objective, we assume the learner has shown proficiency with respect to that objective. However, as educators are well aware, very little if any insight about learning can be gained from a correct answer. Assessments in the service of learning should be designed to probe

deeper into the learner's thinking, not only allowing them to reflect on their own understanding of the construct, but to also see if they can verbalize their thinking and essentially instruct others about the topic. Observations, cognitive interviewing, and open-ended questions are just a few ways to collect this type of information. These qualitative data collected from the assessment, in combination with the quantitative data, provide the evidence needed to make informed decisions regarding instruction and how to further a learner's learning.

Given the importance of validity evidence based on response processes for assessments in the service of learning, we recommend the use of think-aloud and other cognitive probes (See Padilla & Benitez, 2014) to engage learners in the design and revision of these assessments. This participatory approach will reveal how learners perceive what they are asked to do to solve the problems presented on an assessment, what thinking processes they use to solve them, and what misconceptions they may have. The data gathered from these studies can then be compared to the goals of the assessments. When learners confirm the processes they used to solve items are consistent with the same ones the assessment was designed to target, this evidence supports use of the test for its intended purpose.

In addition to direct measures of learners' thought processes when solving test items, log data from digital assessments can be used as an indirect measure of their cognitive behaviors. These log data include the amount of time learners took to respond to an item, what help features they used, the order in which they completed steps to solve a problem, whether they were engaged in trying to solve the item, and whether they skipped or returned to items (Araneda et al., 2022; He et al., 2023; Wise & Kong, 2005).

Validity evidence based on internal structure.

Validity evidence based on internal structure refers to statistical analyses of students' responses to test items. For example, if the items on a test are all designed to measure a single concept or skill, statistical analysis of students' responses should reflect a single "dimension" or "factor" being measured. On the other hand, if an assessment is designed to provide information about multiple skills, then statistical analysis of students' response data should reveal separate, multiple dimensions. Procedures for gathering validity evidence based on internal structure are technical from a statistical perspective and include methods such as "factor analysis" (both exploratory and confirmatory) and "multi-dimensional

scaling." Internal structure evidence also evaluates the "strength" or "salience" of the major dimensions underlying an assessment and so would also include indices of measurement precision such as reliability estimates, decision consistency (i.e., consistency of decisions or classifications made if the students were retested), and other measures of precision (e.g., standard errors of measurement, test information, etc.).

Validity evidence based on internal structure can also be used to evaluate if the test is operating in the same way across different types of students. Analyses of potential item bias across different types of students (i.e., differential item functioning) or of differences in dimensionality across different types of students (e.g., invariance of test structure) also fall under the category of internal structure validity evidence. Rios and Wells (2014) provide excellent examples of conducting such "invariance" of structure analyses.

Validity evidence based on test structure requires a large amount of student response data and so is less applicable to classroom assessment data. In addition, the structure of an assessment will likely differ based on testing students who have or have not yet had the opportunity to learn the material tested. However, if commercial tests are used and administered to students who have already experienced the intended instruction, such evidence should be provided to ensure the types of information provided by the assessment are supported by the dimensionality and precision analyses.

Validity evidence based on relations to other variables.
Validity evidence based on relations to other variables refers to analyses that involve students' test scores as one variable in the analysis along with other variables designed to (a) confirm or disconfirm the test is measuring what it intends to measure, or (b) determine the degree to which test scores are predictive of certain outcomes of interest. An example of confirmatory evidence is the degree to which reading test scores correlate positively with teachers' ratings of students' reading proficiency. In this example, students' test scores are one variable in the analysis and teachers' ratings are the other variable. As an example of disconfirming evidence, a study could be done to evaluate the relationship between students' reading test scores and a measure of test anxiety. The validity of the assessment would be supported if the test scores showed no or little relationship to anxiety. An example of prediction would be studying the degree to which students'

test scores at the midterm of a semester predict their performance on an Advanced Placement exam at the end of the semester. Validity evidence based on relations to other variables typically use the statistical techniques of correlation and multiple regression analyses to demonstrate and evaluate the strength of relationships between test scores and other variables (See Bandalos, 2018; and Sireci & Benitez, 2023, for examples).

In considering how this type of validity evidence can improve or confirm assessment in the service of learning, other indicators of student learning need to be gathered or available. Examples of relevant external variables include the quality and amount of instruction provided; or the amount of time learners interact with assignments, educational games, or other learning tools. Different groups of learners, or different time periods, could also be used as the "other" variable in these analyses. Investigating how test scores differ across students with different learning styles is likely to provide informative validity evidence. Studying relations of other variables to test scores requires test developers, educators, and researchers to be creative in imagining how evidence of student learning could be reflected outside of the assessment context, gathering that evidence, and then analyzing its relationships with test scores.

Validity evidence based on analysis of score interpretations.
Educational tests are developed with explicit intended interpretations and uses of test results in mind. Thus, those developing the tests and those using test results should have a shared understanding of what is represented by a test score, how that score should be interpreted, and what types of actions it can inform. Thus, the actual interpretations made on the basis of test scores should be the same as those that motivated test development. Thus, evidence that test scores are appropriately interpreted by those who use them is a key component of a validity investigation. This source of evidence evaluates the congruence between intended interpretation and uses of scores and actual interpretations and uses.

Gathering validity evidence based on the appropriateness of score interpretation and uses requires that first, these interpretations and uses are clearly defined, and then investigators engage with teachers, students, and other users of test results to study the degree to which they are interpreting test results as intended. Such research is recommended at both the end of a testing process to confirm the intended interpretations are realized, and also early in the test development process

when designing and developing test score reports. By understanding the aspects of test reports that are confusing to or misinterpreted by teachers, students, and other stakeholders, improvements can be made to the communication of test results to facilitate valid interpretation.

Validity evidence based on the consequences of testing.

Evidence based on the consequences of testing refers to evaluation of the intended and unintended consequences associated with a testing program. Intended consequences of testing originate from the explicit use and intended interpretations of the test scores established by the test developer. For example, an intended consequence of a test used for the service of learning may be the placement of a learner into a proper instructional group. Evidence based on this consequence would show support that the test accurately placed the learner into an instructional group that promotes the learner's academic growth and potential. Unintended consequences of testing can originate from various sources, such as culturally irrelevant test items and/or claims that extend beyond the intended interpretations and use of the test scores. For example, a test intended to place a learner into a proper instructional group may have some items that are unintentionally culturally irrelevant to the learner (Solano-Flores, 2019; Randall, 2021). In this case, evidence based on this consequence may show misalignment between learners' true abilities, their performance on the test, and their achievement in the instructional group. Extending this example, the learner's performance on the placement test could also lead to the misinterpretation of the learner's interests or motivation. Evidence based on this consequence may show a negative association between the learners achievement in the instructional group and their actual interest or motivation in what is being taught.

It is important that the evidence based on consequences reflects the reality of the testing situation. Prior experiences with testing can inform ways that potential consequences can be anticipated and addressed. Evidence based on the consequences of testing are going to be unique for every testing situation. However, it is important that the intended and unintended consequences associated with testing are thoroughly evaluated to ensure the learner does not experience any harmful consequences and only receives consequences that are beneficial to their learning.

There are many ways to gather validity evidence based on testing consequences. In some contexts where the results of testing lead to personal rewards such as selection into a competitive school or class, or a credential, the degree to which the test-based rewards differ across demographic groups will be of interest. Such "adverse impact" studies (Sireci & Geisinger, 1998) evaluate the selection or passing rates across groups and compare them to expected, equitable outcomes. Lane (2014) described other studies that can be done to evaluate the consequences of educational tests such as analysis of classroom artifacts, changes in teaching practices, and surveys and interviews of teachers and students.

Synthesizing and Documenting Validity Evidence

Educators who develop assessments for their students may not feel the need to demonstrate a strong body of evidence to support their use, but it is important to provide such evidence to confirm their assessments are doing what they intend to do and demonstrate a level of quality commensurate with their intended uses. Only with such evidence can educators feel confident in the information provided and the decisions they make based on that evidence. For tests and assessment systems produced on a larger scale, clear and sufficient evidence is needed to justify use of the test. In these cases, the body of evidence and theory that support the use of a test for a particular purpose is called a *validity argument* (AERA et al., 2014; Kane, 1992, 2006, 2013). The word "argument" is not intended to infer a fight or a heated debate, but rather a compelling and cohesive synthesis of theoretical analysis and research results that can be used to justify or dismiss the use of a test for a particular purpose. Thus, a validity argument clearly states the purposes and goals of a testing program and provides an informative summary of the research-based evidence gathered to evaluate how well the test (a) measures the intended constructs, (b) provides the information required to fulfill the testing purposes, (c) promotes accurate interpretations of that information, and (d) does not result in harmful unanticipated consequences for individuals or society.

One may wonder when the evidence in a validity argument is sufficient to justify the use of an assessment in the service of learning. Although it will always be a judgment call, the goal is to provide a compelling argument that would satisfy even a reluctant critic. The AERA et al. (2014) *Standards* acknowledge this subjectivity, but note, "at some point validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible" (p. 22).

Concluding Remarks: Valid Assessments in the Service of Learning

In this chapter, we presented a basic overview of the concept of validity in educational testing, the process of test validation, and how they relate to assessments in the service of learning. There is much written about the theory of validity and the process of test validation, and we have provided many references to further reading in those areas. What we believe is most important to keep in mind with respect to the validity and validation of assessment in the service of learning is (a) confirming the assessments are appropriate and beneficial for the learners being assessed, (b) ensuring the assessment model and content represents and measures the intended learning objectives, (c) the assessment properly informs instructional decisions (and the consequences associated with those decisions), and (d) outcomes of the assessment are interpreted as intended, provide valuable insight into the learner's learning processes, and do not result in any undesirable harmful consequences.

With respect to (a) the degree to which assessments in the service of learning are culturally responsive will deserve specific attention. The learners we assess are diverse with respect to culture, language, socioeconomic status, gender identity, neurodiversity, and many other factors. Valid assessments that support learning should provide both mirrors into learners' own cultures as well as "windows" into other cultural groups (Randall et al., 2022).

The other goals of valid assessment can be summarized by stating validity evidence should be provided that assessments in the service of learning appropriately and sufficiently represent their intended learning goals, and do not unintentionally measure other factors. By vigilantly validating our assessments, we not only uphold their effectiveness, but also fulfill the promise of assessment in the service of learning–making our assessment practices more transparent, more high-quality, and more directly informative for every learner's next steps.

We hope the descriptions of validity we provided in this chapter, as well as the examples of validity evidence that can be gathered to evaluate an assessment, are helpful to those who strive to use assessment in the service of learning to help learners learn. Like all endeavors in assessment in the service of learning, facilitating and evaluating validity is a journey in which we continually learn about how we can improve assessments and the assessment validation process. As assessments in the service of learning evolve, we anticipate more examples of validation studies that provide new insights into how we can best use assessments to support student learning.

References

Araneda, S., Lee, D., Lewis, J., Sireci, S. G., Moon, J. A., Lehman, B., Arslan, B., & Keehner, M. (2022). Exploring relationships among test takers' behaviors and performance using response process data. *Education Sciences, 12*, 104. https://doi.org/10.3390/educsci12020104

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

Bandalos, D. L. (2018). *Measurement theory and application for the social sciences.* Guilford Press.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21–29.

Crocker, L. M., Miller, D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–194.

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

Gordon, E. W. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39*(3), 72–78.

He, Q., Borgonovi, F., & Suárez-Álvarez, J. (2023). Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method. *Journal of Computer Assisted Learning, 39*(3), 719–736. https://doi.org/10.1111/jcal.12748.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*,527–535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp.. 17–64). American Council on Education/Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessments, and instruction,*Review of Educational Research, 79*(4), 1332–1361.

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

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 83–108). Charlotte, NC: Information Age Publishing Inc.

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.

Padilla, J., & Benítez, I, (2014). Validity evidence based on response processes. *Psicothema*, 26, 136–144.

Pimentel, S. (2013). *College and career readiness standards for adult education.* U.S. Department of Education, Office of Vocational and Adult Education. http://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82–90.

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: toward a justice-oriented, antiracist validity framework. *Educational Assessment*, https://doi.org/10.1080/10627197.2022.2042682

Rios, J., & Wells, C. S. (2014). Validity evidence based on internal structure. *Psicothema, 26*(*1*), 108–116.

Sireci, S. G. (2020). "De-"Constructing" Test Validation," *Chinese/English Journal of Educational Measurement and Evaluation, 1.* 教育测量与评估双语季刊: https://www.ce-jeme.org/journal/vol1/iss1/3

Sireci, S. G., & Benitez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217–226. https://www.psicothema.com/pdf/4805.pdf

Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321.

Sireci, S. G., Crespo Cruz, E., Suárez-Álvarez, J., & Rodriguez, G. (2025). Understanding UNDERSTANDardization research. In *Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy* (R. Bennett, L. Darling-Hammond, & A. Badrinarayan, Eds., pp. 415–433). Routledge.

Sireci, S. G., & Faulkner-Bond, K. (2014). Validity evidence based on test content. *Psicothema.* https://doi.org/10.7334/psicothema2013.256

Sireci, S. G., & Geisinger, K. F. (1998). Equity issues in employment testing. In J. H. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.), *Test interpretation and diversity* (pp. 105–140). American Psychological Association: Washington, D.C.

Solano-Flores G (2019). *Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. Frontiers in Education* 4:43. https://doi.org/10.3389/feduc.2019.00043

Walker, M. E., Olivera-Aguilar, M., Lehman, B., Laitusis, C., Guzman-Orth, D., & Gholson, M. (2023). Culturally responsive assessment: Provisional principles. *ETS Research Report Series, 2023*(1), 1–24.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2