Psicothema

Psicothema 2014, Vol. 26, No. 1, 100-107
doi: 10.7334/psicothema2013.256

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Validity evidence based on test content

Stephen Sireci and Molly Faulkner-Bond
University of Massachusetts Amherst (USA)

Abstract

Background: Validity evidence based on test content is one of the five forms of validity evidence stipulated in the Standards for Educational and Psychological Testing developed by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. In this paper, we describe the logic and theory underlying such evidence and describe traditional and modern methods for gathering and analyzing content validity data. **Method:** A comprehensive review of the literature and of the aforementioned Standards is presented. **Results:** For educational tests and other assessments targeting knowledge and skill possessed by examinees, validity evidence based on test content is necessary for building a validity argument to support the use of a test for a particular purpose. **Conclusions:** By following the methods described in this article, practitioners have a wide arsenal of tools available for determining how well the content of an assessment is congruent with and appropriate for the specific testing purposes.

Keywords: Testing standards, validity, alignment, content validity, test development, validation.

Resumen

***Evidencia de validez basada en el contenido del test*. Antecedentes:** la evidencia de validez basada en el contenido del test es una de las cinco formas de evidencias de validez estipuladas en los *Standards for Educational and Psychological Testing* de la *American Educational Research Association*. En este artículo describimos la lógica y teoría que subyace a tal fuente de evidencia, junto a métodos tradicionales y modernos para obtener y analizar los datos de validez de contenido. **Método:** una revisión comprehensiva de la bibliografía y de los mencionados Standards. **Resultados:** para los tests educativos y otras evaluaciones de los conocimientos y habilidades que poseen los examinados, la evidencia de validez basada en el contenido del test es necesaria para elaborar un argumento de validez que apoye el uso de un test para un objetivo particular. **Conclusiones:** siguiendo los métodos descritos en este artículo, los profesionales tienen un amplio arsenal de herramientas disponibles para determinar en qué medida el contenido de una evaluación es congruente y apropiado para los objetivos específicos de la evaluación.

Palabras clave: examinando los Standards, validez, alineamiento, validez de contenido, elaboración de tests, validación.

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). With respect to evidence, they specify five "sources of evidence that might be used in evaluating a proposed interpretation of test scores for particular purposes" (p. 11). These five sources are validity evidence based on test content, response processes, internal structure, relations to other variables, and testing consequences. In this article, we describe validity evidence based on test content. Our goals are to describe (a) content validity evidence, (b) methods for gathering content validity data, and (c) methods for analyzing and summarizing content validity data. Our intent is to inform readers of these important areas so they understand how to gather and analyze validity evidence based on test content to evaluate the use of a test for a particular purpose.

Received: August 27, 2013 • Accepted: October 2, 2013
Corresponding author: Stephen Sireci
Center for Educational Assessment, Hills South
University of Massachusetts Amherst
01003 Amherst (USA)
e-mail: sireci@acad.umass.edu

Defining testing purposes

As is evident from the AERA et al. (1999) definition, tests cannot be considered inherently valid or invalid because what is to be validated is not the test itself, but rather the *use* of a test for a particular *purpose*. Therefore, the first step in validation, and in test development in general, is to specify the intended uses and interpretations of test scores. Thus, gathering validity evidence based on test content, like all other forms of validity evidence, must focus on supporting or evaluating intended testing purposes.

Validity Evidence Based on Test Content versus Content Validity

In the previous versions of the *Standards* (i.e., APA, AERA, & NCME, 1954, 1966, 1974, 1985), validity evidence based on test content was described as "content validity," and this term was also common in the psychometric literature. Lennon (1956) provided an early definition of content validity as "the extent to which a subject's responses to the items of a test may be considered to be a representative sample of his responses to a real or hypothetical universe of situations which together constitute the area of concern to the person interpreting the test" (p. 295). Sireci (1998b) provided a broader definition that included aspects of test development. He described content validity as pertaining to four elements of

test quality: domain definition, domain representation, domain relevance, and appropriateness of the test development process.

Although there is long consensus that these four elements are important for evaluating the use of a test for a particular purpose, many validity theorists claimed "content validity" was not a technically correct term because validity refers to interpretations of test scores and not to the content of an assessment (e.g., Messick, 1989). We see the theoretical logic in that argument; however we, like Ebel (1956, 1977) and Yallow and Popham (1983) believe the term "content validity" is useful for practitioners and lay audiences and effectively communicates an important aspect of the quality of test scores. We define content validity as the degree to which the content of a test is congruent with testing purposes. In addition, we use the terms "validity evidence based on test content" and "content validity evidence" interchangeably. Essentially, the "debate" over the term content validity is one of nomenclature and is likely to persevere in academic circles. However, what will also persevere is the importance of affirming that the content of a test represents its intended construct and is appropriate for accomplishing the testing purposes.

Evaluating test content

The four elements of content validity described by Sireci (1998b)—domain definition, domain representation, domain relevance, and appropriateness of test construction procedures—give us a framework for evaluating test content. Domain definition refers to how the "construct" measured by a test is operationally defined. A *construct* is the theoretical attribute measured by a test, or as Cronbach and Meehl (1955) described *"some postulated attribute of people, assumed to be reflected in test performance"* (p. 283).

A domain definition provides the details regarding what the test measures and so it transforms the theoretical construct to a more concrete content domain. For educational tests, defining the domain measured is typically accomplished by providing (a) detailed descriptions of the content areas and cognitive abilities the test is designed to measure, (b) test specifications that list the specific content "strands" (sub-areas), as well as the cognitive levels measured, and (c) specific content standards, curricular objectives, or abilities that are contained within the various content strands and cognitive levels. For achievement testing in elementary, middle, and secondary schools, the content and cognitive elements of the test specifications are typically drawn from curriculum frameworks that guide instruction. For licensure and certification tests, they are typically drawn from comprehensive practice analyses (Raymond, 2001). Newer methods for defining the domain include "evidence-centered design" (Mislevy, 2009; Mislevy & Riconscente, 2006) or "principled assessment design" (Luecht, 2011), which require the specification of "task models" that will generate the types of information specified in a testing purpose.

Evaluating domain definition involves acquiring external consensus that the operational definition underlying the test is congruent with prevailing notions of the domain held by experts in the field. This is typically accomplished by convening independent expert panels to help develop and evaluate the test specifications. The degree to which important aspects of the construct, curriculum, or job domain are *not* represented in the test specifications is an important criterion for evaluating domain definition. In some cases, it is difficult to measure all aspects of a domain and so the domain definition will explicitly acknowledge those aspects of the domain the test does not measure.

Domain representation refers to the degree to which a test adequately represents and measures the domain as defined in the test specifications. To evaluate domain representation, external and independent "subject matter experts" (SMEs) are recruited and trained to review and rate all the items on a test (Crocker, Miller, & Franks, 1989; Sireci, 1998a). Essentially, their task is to determine if the items fully and sufficiently represent the targeted domain. Sometimes, as in the case of state-mandated testing in public schools, SMEs judge the extent to which test items are congruent with the curriculum framework. These studies of domain representation have recently been characterized within the realm of test *alignment* research (Bhola, Impara, & Buckendahl, 2003). Alignment methods and other strategies for gathering and analyzing content validity data are described later.

Domain relevance addresses the extent to which each item on a test is relevant to the targeted domain. An item may be considered to measure an important aspect of a content domain and so it would receive high ratings with respect to domain representation. However, if it were only tangentially related to the domain, it would receive low ratings with respect to relevance. For this reason, studies of content validity may ask subject matter experts to rate the degree to which each test item is relevant to specific aspects of the test specifications, and then aggregate those ratings within each content strand to determine domain representation (Sireci, 1998a). Taken together, studies of domain representation and relevance can help evaluate whether (a) all important aspects of the content domain are measured by the test, and (b) whether the test contains trivial or irrelevant content. As Messick (1989) described "Tests are imperfect measures of constructs because they either leave out something that should be included… or else include something that should be left out, or both" (p. 34). A thorough study of content validity, prior to assembling tests, protects against these potential imperfections.

The fourth aspect of content validity, *appropriateness of the test development process*, refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material. The content validity of a test can be supported if there are strong quality control procedures in place during test development, and if there is a strong rationale for the specific item formats used on the test. Examples of quality control procedures that support content validity include (a) reviews of test items by content experts to ensure their technical accuracy, (b) reviews of items by measurement experts to determine how well the items conform to standard principles of quality item writing (Haladyna & Downing, 1989), (c) sensitivity review of items and intact test forms to ensure the test is free of construct-irrelevant material that may offend, advantage, or disadvantage, members of particular sub-groups of examinees (Ramsey, 1993), (d) pilot-testing of items followed by statistical item analyses to select the most appropriate items for operational use, and (e) analysis of differential item functioning to flag items that may be disproportionally harder for some groups of examinees than for others (Holland & Wainer, 1993).

With respect to evaluating test content, the AERA et al. (1999) *Standards* state

> "Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance

> of the content domain to the proposed interpretation of test scores. Evidence based on test content can also come from expert judgments of the relationship between parts of the test and the construct" (p. 11).

In the next sections, we describe studies that can be conducted to evaluate these aspects of domain definition, domain representation, and domain relevance.

Methods for conducting content validity and alignment studies

There are a variety of methods that could be used to evaluate the degree to which the content of an assessment is congruent with the testing purposes. Some methods are based on traditional notions of content validity, while others are based on newer notions of test-curriculum alignment. Almost all methods involve SMEs. The differences among the methods essentially stem from (a) the tasks presented to the SMEs, (b) how their data are analyzed, (c) the grain size of the content domain that is the focus of the analysis, and (d) how the data are summarized. Given that all methods involve SMEs, the selection, qualifications, and training of the SMEs essentially determines the quality of a content validity study. All SMEs should be thoroughly knowledgeable with respect to the knowledge and skills being tested, and should be properly trained to complete any item reviews and other tasks. Based on the literature (e.g., O'Neil et al., 2004; Penfield & Miller, 2004), we recommend at least 10 SMEs be used for a content validity or alignment study.

Traditional content validity studies

The most common methods for gathering validity evidence based on test content require SMEs to either (a) match test items to their intended targets, (b) rate the degree to which items adequately represent their intended content and cognitive specifications, or (c) rate the degree to which items are relevant to the domain tested. These studies typically use a "matching task" or Likert-type rating scales to measure the congruence between each item and whatever aspects of the content domain the SMEs are being asked to consider. An example of a "matching" task is presented in Table 1, and an example of how the data from such a study could be summarized is presented in Table 2.

From the matching approach (Table 1), we can see how these data can inform us about the degree to which the items represent their targeted content areas and cognitive levels. For example, the summary of the matching data presented in Table 2 illustrates that the SMEs perceived the content areas measured by the items to be relatively more congruent with the test specifications than the cognitive levels. In particular, the "Analysis, Synthesis, and Evaluation" items were rated less congruent than items from the other areas and levels. The results from this type of study can be used to eliminate or revise particular items, create new items that better represent the areas perceived to be less congruent, or reconsider how these content areas and cognitive levels are defined in the test specifications.

Although the matching approach is useful for evaluating domain representation in a general sense, it does not give us information

Table 1
Example of item matching task for a hypothetical Math Achievement Test

Item #	Content area (select one)			Cognitive level (select one)		
	Number relations	**Patterns, functions, & Algebra**	**Geometry & measurement**	**Knowledge, Comprehension**	**Application**	**Analysis, Synthesis, Evaluation**
1						
2						
3						
4						
5						
…						
100						

Directions: Please review each item and indicate (a) the Content area, and (b) Cognitive level you think the item is measuring. Please be sure to make two selections for each item—one for content area, and one for cognitive level.

Table 2
Example of a summary of item congruence data from a matching task

Content area/Cognitive level	# of items	% of items classified correctly by all SMEs	% of items classified correctly by at least 70% of SMEs
Number sense	25	72	88
Patterns, Functions, Algebra	35	49	94
Geometry/ Measurement	40	55	91
Knowledge & Comprehension	25	48	80
Application	35	49	80
Analysis, Synthesis, Evaluation	40	53	63
Average		54%	83%

Table 3
Example of SME rating task assessing item/objective congruence

Item	Objective	How well does the item measure its objective? (circle one)					
		1 (not at all)	2	3	4	5	6 (very well)
1	Convert units of measure in the same systems						
2	Read values on a bar, line, or circle graph						
3	Find the average (mean) and range for a data set						
4	Find the perimeter of rectangles						
5	Infer meaning from gaps, clusters and comparisons of data						
6	Directly measure and compare the radius, diameter, and circumference of a circle						
8	Read and understand positive and negative numbers as showing direction and change						
…							
100	Use a number line to represent the counting numbers						

Directions: Please read each item and its associated benchmark. Rate how well the item measures its objective using the rating scale provided. Be sure to circle one rating for each item

about *how well* the items measure their associated achievement target. Rating scale approaches are more typically used to gather that type of evidence. An example of an item-objective congruence rating scale approach is presented in Table 3.

Using the rating scale approach we can get an idea of how well specific items, and the group of items measuring a specific objective, adequately measure the intended objective. These data can be summarized at more general levels of the test specifications. For example the objectives within a content area can be aggregated to evaluate the content area as a whole, as we illustrate in Table 4. These fictitious results may suggest that the content categories have good representation with respect to the degree to which the items are measuring the content areas. However, it may still be advisable to review items flagged for low ratings and consider revising or deleting them. A similar aggregation could be conducted to evaluate how well the items are measuring the intended cognitive skills.

Table 4
Example summary of results from Item/CCSS congruence study

Item	Content area	Mean	Median	Aiken Index
1	Number Sense	4.2	4.0	.89*
2	Number Sense	5.3	5.0	.91*
3	Number Sense	4.1	4.5	.90*
4	Number Sense	3.5	4.0	.91*
5	Number Sense	4.6	4.0	.93*
6	Number Sense	3.7	4.0	.92*
7	Number Sense	5.2	5.0	.95*
	Average for Content Area	4.2	4.0	.89*
8	Patterns, Functions, Algebra	3.4	3.5	.76*
9	Patterns, Functions, Algebra	4.5	5.0	.90*
10	Patterns, Functions, Algebra	5.6	5.5	.95*
11	Patterns, Functions, Algebra	5.2	5.0	.92*
12	Patterns, Functions, Algebra	5.4	5.5	.94*
13	Patterns, Functions, Algebra	5.3	5.5	.93*
	Average for Content Area	**4.9**	**5.0**	**.90**
…				

Notes: Statistics based on 10 SMEs and rating scale where 1= Not at all, 6 = very well.
* p<.05

Regardless of the method chosen, appropriately summarizing the results of these content validity studies is important. In addition to the descriptive summaries of domain representation, these studies should also compute congruence/alignment statistics. Such statistical summaries range from purely descriptive to those that involve statistical tests. On the descriptive end, Popham (1992) suggested a criterion of 7 of 10 SMEs rating an item congruent with its standard to confirm the fit of an item to its standard. This 70% criterion could be applied to a more aggregate level such as at the content area or cognitive level. On the statistical end, several statistics have been proposed for evaluating content congruence such as Hambleton's (1980) item-objective congruence index and Aiken's (1980) content validity index. Aiken's index (illustrated in Table 4) ranges from zero to one and essentially indicates the proportion of SMEs who rate the item as above the midpoint of the congruence scale. It can also be evaluated for statistical significance using a variation of the z-test for a sample proportion. In addition, Penfield and Miller (2004) established confidence intervals for SMEs' mean ratings of content congruence.

Alignment methods

Up to this point we have discussed test specifications using two very general dimensions—content areas and cognitive levels—and a finer-grain size of content objective. Alignment methods designed for educational tests tend to focus on several levels of test specifications, and some methods even assess the alignment of the assessment with instruction.

While traditional content validity studies tend to focus more on the broader levels of the content domain and its relation to the test design and specifications, alignment studies take a more fine-grained approach and evaluate the degree to which the content of a test appropriately represents its intended domain in terms of various criteria such as depth, breadth, or cognitive complexity. Alignment methods for evaluating test content emerged from state-level educational achievement testing in the U.S. Bhola et al. (2003)

defined alignment as "the degree of agreement between a state's content standards for a specific subject area and the assessment(s) used to measure student achievement of these standards" (p. 21).

In this section we describe three of the most common alignment methods used in educational testing in the U.S.—the Webb, Achieve, and Surveys of Enacted Curriculum methods. After briefly describing these methods, we discuss a proposed framework for evaluating the outcomes of an alignment study.

Alignment methods

In general, all alignment methods share common characteristics. First, all require the existence and use of a clearly articulated set of content standards against which to evaluate a set of test items. Although the exact ways in which the content standards are put to use or evaluated will vary depending on the method used, as we describe later, all three methods described here require at least two levels of articulation. Second, all alignment methods require convening a panel of SMEs with expertise in the area(s) relevant to the testing purpose. The exact tasks these SMEs will carry out varies depending on the specific method used, but all methods typically begin with a comprehensive training session in which panelists discuss and familiarize themselves with the standards and the test.

Webb method

The Webb (1997) alignment method proposes five dimensions from which to evaluate alignment between content standards and assessments. These dimensions are (a) content focus, (b) articulation across grades and ages, (c) equity and fairness, (d) pedagogical implications, and (e) system applicability. In practice, only content focus has been implemented as the basis for gathering validity evidence based on test content. Descriptions of these dimensions are summarized in Table 5, along with Webb's suggested evaluation criteria for each dimension. As will be clear in subsequent discussion of other methods, this articulation of specific criteria is unique to Webb, and useful; on the flip side, however, there may be relevant exceptions or challenges in meeting these criteria, as we discuss further below.

Similar to the matching task described earlier, in the Webb method, SMEs provide ratings that are essentially binary. For example, SMEs select the content standard they feel is the best match for an item without noting the degree or quality of the match. To arrive at final indices for each sub-dimension, the ratings across all panelists are averaged. For example, to determine the overall categorical concurrence rating, one would average the total number of item-objective matches for all SMEs for each standard. These averages are then compared to the criteria provided. This is in contrast to other methodologies that require SMEs to reach consensus in their ratings, and may mask disagreements or direct conflicts across raters. As Martone and Sireci (2009) noted, such conflicts could be problematic in areas such as categorical concurrence, where panelists may identify six unique items per strand as a group, but disagree among themselves about which items actually match to a given strand.

Webb (1997) also noted the potential for trade-offs in evaluating the full set of ratings from an alignment study. Ideally, an assessment should meet the evaluation criteria for all four sub-dimensions in order to be aligned. In reality, however, even with the relatively lax criteria of 50% for some dimensions, this may not be achievable. Where this is the case, test developers and users will need to use judgment to determine whether their findings indicate acceptable alignment given the test's intended use.

Achieve method

The Achieve alignment method (Rothman, Slattery, Vranek, & Resnick, 2002), is designed to answer three questions:

1. Does each assessment measure *only* the content and skills reflected in the standards? In other words, can everything on the test be found in the state standards?
2. Does each assessment fairly and effectively sample the important knowledge and skills in the standards?
3. Overall, is each assessment sufficiently challenging? (Rothman et al., 2002, p. 6).

The Achieve method is summarized in Table 6. As the table shows, it is structured fairly similarly to the Webb method in that it also uses four dimensions, most of which correspond directly to dimensions in the Webb procedure (e.g., Content Centrality and Categorical Concurrence both attend to how well the items match to different levels of the standards in terms of their content).

The Achieve method differs from the Webb method in three key ways. First, it begins by verifying the test blueprint as a foundation for SMEs' subsequent judgments about the test's content. This step is conducted by a "senior reviewer" ahead of the SME panel

Table 5
Description of Webb (1997) Method

Dimension	Description	Evaluation criterion
Categorical concurrence	The extent to which the items on the test correspond to strands* in the content standards	Minimum of six items per strand
Depth of knowledge	The level of consistency between cognitive complexity articulated in objectives** and tested by items	At least 50% of items should be at or above cognitive complexity level articulated in corresponding objectives**
Range of knowledge	The level of consistency between the range of complexity articulated in objectives** and tested by items	At least 50% of objectives** should be measured by at least one assessment item
Balance of representation	The extent to which the test mirrors the standards in terms of relative emphasis on different strands or topics	Index indicating relative proportion of items to objectives** between standards and test approaches 1

* Most general level at which standards or expectations are articulated
** Most specific level at which standards or expectations are articulated

Table 6
Description of achieve method

Dimension	Focus of measurement	Rating system used
Content centrality	Degree and quality of match between content addressed in each item and the objective* it is designed to measure	Four-point scale: 2 = clearly consistent, 1B = partially consistent, 1A = unclear, 0 = not consistent
Performance centrality	Degree and quality of match between cognitive complexity for each item and the objective* it is designed to measure	Four-point scale: 2 = clearly consistent, 1B = partially consistent, 1A = unclear, 0 = not consistent
Challenge	Source – Degree to which source of challenge in each item is construct-relevant (appropriate) or not (inappropriate). Level – Degree to which a set of items is appropriately challenging for intended examinee population	Source – 1 = appropriate, 0 = inappropriate (automatic 0 if 0 for both content and performance centrality) Level – Written evaluation based on global judgment
Balance and range	Balance – Degree to which a set of items matches the objectives* it is designed to measure in terms of emphasis and representation Range – Fraction of total objectives* within a standard** that are assessed by at least one item	Balance – Written evaluation based on group discussion Range – Indices between 0.50-0.66 are acceptable; index > 0.67 is good

* Most specific level at which standards or expectations are articulated
** Most general level at which standards or expectations are articulated

meeting, and includes identifying "discrepant" items that fail to map correctly to any standards or objectives. Discrepant items are either reassigned or eliminated prior to the SME panel's review, and the senior reviewer documents such decisions by summarizing the observed problem(s) for each. Second, raters use scales for some dimensions that allow them to indicate the degree to which an item matches its standards beyond a simple yes/no judgment. For some ratings, reviewers may also choose primary and secondary matches, which allow for the possibility of noting adjacent ratings across reviewers. Third, the Achieve method requires SMEs to reach consensus on their judgments, in contrast to the averaging procedures used in the Webb method.

The Achieve method also requires panelists to consider items both individually and as sets. These different considerations are phased hierarchically such that the SMEs first make judgments about the content centrality, performance centrality, and source of challenge for individual items and then, having made these judgments, consider the level of challenge, balance and range for sets of items aggregated by strand or test form. These discussions are not systematically built into the Webb method.

Surveys of Enacted Curriculum method

The Surveys of Enacted Curriculum (SEC) method (CCSSO SEC Collaborative Project, 2005; Porter & Smithson, 2001) differs from Achieve and Webb in that its model for evaluating alignment between standards and assessment also considers curriculum and instruction as part of the overall alignment picture. By collecting and incorporating information about what teachers teach in their classrooms and how they use standards in their instruction, it produces information and displays that support action and improvement for teachers, in addition to serving as documentation for validity evidence based on test content.

The SEC method involves the use of a two-dimensional content matrix that crosses pre-specified content topics (specific to the SEC method and based on past research of classroom instruction and standards) with expectations for student performance (which are similar to cognitive levels – e.g., memorize, solve non-routine problems, etc.). Table 7 shows a simplified example of an SEC rating matrix.

Different panelists code different components of the system. For example, teachers code for instructional time using a four-point scale to indicate the percentage of time they devote to a given content/performance combination. For test items and content objectives in the standards, panels of SMEs code by placing the relevant unit (i.e., a specific test items or content objective) into the cell they believe best represents the combination of content and cognitive demand for that unit. For items and objectives, Porter (2002) recommended coding at the most specific level available, as findings can be aggregated upwards according to the structure of the standards.

Test developers can draw conclusions about the alignment between any two components of a system (i.e., standards and instruction, assessment and instruction, and standards and assessment) by comparing the corresponding cell results for each. The index proposed by Porter (2002) essentially corresponds to the proportion of overlap between the cells in the two different matrices (e.g., instruction matrix and assessment matrix). Porter does not provide specific criteria or cut-offs denoting acceptable levels for this index.

Evaluating alignment for an alternate assessment

The three methods described above were all developed for use in the context of general education achievement tests that are used for accountability purposes in K-12 settings in the US. For assessments that are used for different purposes or in different

Table 7
Example of SEC sample matrix

	Expectations for student performance				
Content match	**Memorize**	**Communicate**	**Solve**	**Connect**	**Generalize**
Measurement					
Units of measure					
Conversions					
Surface area					
Area, volume					
Etc.					

contexts, these methods would likely require modification, or alternate methods may be used or developed.

One such example is the Links for Academic Learning (LAL) model (Flowers, Wakeman, Browder, & Karvonen, 2009), an alignment methodology designed specifically for use with alternate assessments for students with significant cognitive disabilities. After evaluating the relevance of extant procedures for use with alternate assessments, Flowers et al. (2009) determined that a new model would be necessary to suit the specific concerns and priorities that are unique to the special education context. The LAL method they developed comprises eight alignment criteria that are specifically relevant for alternate assessment, such as the degree to which communication barriers for the student are minimized or the extent to which the assessment focuses on academic (versus functional) content. Other alignment methodologies could be developed in similar fashion to suit other non-K-12 purposes.

Evaluating alignment

In addition to the rating systems described above that are internal to the alignment process, test developers or users may also wish to evaluate the quality of an alignment methodology itself (or a method's outcomes). In other words, having implemented an alignment method and drawn conclusions about an assessment's appropriateness for a particular use, test developers or users may also want or need to defend these conclusions as justified and valid.

Davis-Becker and Buckendahl (2013) proposed a framework for collecting evidence to defend or evaluate the validity of an alignment procedure. Their framework, which is based on a Kane's (1994) framework for evaluating standard setting studies, comprises four categories of evidence: procedural validity, internal validity, external validity, and utility. For each, the authors present important sub-dimensions (e.g., selection and documentation of panelists, method, and process under procedural validity), as well as evaluation questions to consider, and common potential threats to validity that may arise. Implementation of this framework should help test developers and users defend their alignment results to outside scrutiny, as well as identify any weaknesses that may require the collection of additional validity evidence.

Item similarity ratings

There is one drawback to the content validation/alignment methods discussed so far. By informing the SMEs of the content areas, cognitive levels, objectives/content standards measured by an assessment, they may promote "confirmationist bias" or social desirability. That is, the SMEs may unconsciously rate items more favorably than they actually perceive them to be, to please the researchers. One way around this problem is to have SMEs rate the *similarity* among pairs of test items and use multidimensional scaling to analyze their data (D'Agostino, Karpinski, & Welsh, 2011; O'Neil et al., 2004; Sireci & Geisinger, 1992, 1995).

The logic underling having SMEs rate the similarity among pairs of test items is that items specified to measure similar content areas and cognitive skills in the test specifications should be rated more similar than items measuring different areas and skills. To gather these data, all possible pairings of test items are presented to the SMEs, and their task is to provide a similarity rating for each pair along a Likert-type similarity rating scale. This "paired comparisons" procedure is a valuable method for discovering individual SMEs' perceptions of what the items are measuring without informing them of the test specifications. Thus, it is an elegant and simple manner for controlling unwanted sources of bias in content validity or alignment ratings such as social desirability.

SMEs are required to review the item pairs and circle a rating to indicate their perception of the similarity among the items in terms of the knowledge and skills measured. The SMEs' similarity ratings are analyzed using multidimensional scaling (MDS), which portrays the items in multidimensional space. The results of the analysis are visual, in that the congruence between the SMEs' ratings and the test specifications is ascertained by evaluating the degree to which the dimensions of the MDS space correspond to the dimensions in the test specifications, and the degree to which the items cluster together in this space as predicted by the test specifications. Although this approach addresses concerns regarding social desirability and other rater artifacts, it is not very common because it takes more time for SMEs to complete their ratings and it involves more complex data analysis. D'Agostino et al. (2011) used a sorting task that minimizes the number of required comparisons, and O'Neil et al. (2004) used a item sampling approach to reduce the burden on SMEs. Thus, the general approach is becoming more practical, and can be used to provide important validity evidence regarding the content quality of an assessment.

Discussion

In this article, we defined validity evidence based on test content and described several popular methods for gathering and analyzing such validity evidence. The methods include matching tasks, congruence rating scales, alignment studies, and item similarity ratings. Our review illustrates there are several valuable methods for evaluating the degree to which the content of an assessment is appropriate for a given testing purpose. For educational achievement tests, or tests measuring specific knowledge and skills such as those measured by licensure, certification, or employment tests, we believe validity evidence based on test content is critical for validating test scores that are used to make inferences about examinees with respect to the targeted domain. Although other forms of validity evidence may be important for such tests, validity evidence based on content validity will represent the foundation of any validity argument. When considering the relative importance of criterion-related validity and content validity over 50 years ago, Ebel (1956) noted:

> "The fundamental fact is that one cannot escape from the problem of content validity. If we dodge it in constructing the test, it raises its troublesome head when we seek a criterion. For when one attempts to evaluate the validity of a test indirectly, via some quantified criterion measure, he must use the very process he is trying to avoid in order to obtain the criterion measure" (p. 274).

In the 50+ years that have passed since Ebel's statement, the importance of validity evidence based on test content has not diminished. Thankfully though, the number of methods to assist us in gathering such evidence has increased. Thus, at the present time, we have the knowledge and tools to gather and analyze validity evidence based on test content, and such analyses are likely to improve educational testing practices.

We hope the summary presented in this article empowers and encourages test practitioners and researchers to carry out content

validity studies. Of course, validation of educational assessments is likely to necessitate other forms of validity evidence to convincingly argue that a test is valid for a particular purpose. As the AERA et. Al. (1999) *Standards* point out, in addition to validity evidence based on test content, evidence based on response processes (Padilla & Benitez, 2014), internal structure (Rios & Wells, 2014), relations to other variables (Oren, Kennet-Cohen, Turvall, & Allalouf, 2014), and testing consequences (Lane, 2014), should be used.

References

Aiken, L.R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*, 955-959.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51* (2, supplement).

American Psychological Association (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

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

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

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.

CCSSO SEC Collaborative Project (2005). *Surveys of Enacted Curriculum: A Guide for SEC Collaborative State and Local Coordinators*. Washington, D.C.: Council of Chief State School Officers.

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.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing*, *11*(4), 295-307.

Davis-Becker, S.L., & Buckendahl, C.W. (2013). A proposed framework for evaluating alignment studies. *Educational Measurement: Issues and Practice*, *32*(1), 23-33. doi:10.1111/emip.12002.

Ebel, R.L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement, 16*, 269-282.

Ebel, R.L. (1977). Comments on some problems of employment testing. *Personnel Psychology, 30*, 55-63.

Flowers, C., Wakeman, S., Browder, D.M., & Karvonen, M. (2009). Links for Academic Learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice*, *28*(1), 25-37. doi:10.1111/j.1745-3992.2009.01134.x.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*, 127-135.

Lennon, R.T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

Martone, A., & Sireci, S.G. (2009). Evaluating alignment between curriculum, assessment, 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). Washington, D.C.: American Council on Education.

Mislevy, R.J. (2009, February). Validity from the perspective of model-based reasoning. *CRESST report 752*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence-centered assessment design. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 61-90), Mahwah, NJ: Lawrence Erlbaum.

Oren, C., Kennet-Cohen, T., Turvall, E., & Allalouf, A. (2014). Demonstrating the validity of three general scores of PET in predicting higher education achievement in Israel. *Psicothema, 26*, 117-126.

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

Porter, A.C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3-14.

Porter, A.C., & Smithson, J.L. (2001). Defining, developing, and using curriculum indicators. *Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education*. Retrieved from http://www.cpre.org/sites/default/files/researchreport/788_rr48.pdf.

Raymond, M.R. (2001). Job analysis and the specification of content for licensure and certification exams. *Applied Measurement in Education, 14*, 369-415.

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

Rothman, R., Slattery, J.B., Vranek, J.L., & Resnick, L.B. (2002). *Benchmarking and Alignment of Standards and Testing* (Technical Report No. 566). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://www.cse.ucla.edu/products/reports/TR566.pdf.

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

Sireci, S.G. (1998b). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S.G. (2009). Packing and upacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 19-37). Charlotte, NC: Information Age Publishing Inc.

Webb, N. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. National Institute for Science Education Madison, WI. Retrieved from http://facstaff.wceruw.org/normw/WEBBMonograph6criteria.pdf.

Yalow, E.S., & Popham, W.J. (1983). Content validity at the crossroads. *Educational Researcher, 12*, 10-14.