Chapter 7
Validity Evidence in the *Journal of Educational Psychology*: Documenting Current Practice and a Comparison with Earlier Practice

Rebecca J. Collie and Bruno D. Zumbo

Validity Evidence in the Journal of Educational Psychology Within Two Time Periods

The current *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] 1999) state that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). According to Goodwin and Leech (2003), this view of validity is significantly different from earlier editions of the Standards (AERA et al. 1985; APA et al. 1966, 1974; APA 1954), due in large part to the evolution (Shepard 1993) or metamorphosis (Geisinger 1992) that has taken place in relation to validity theory over the past 50 years (Jonson and Plake 1998). In spite of this significant evolution, scholars (e.g., Borsboom et al. 2004; Hubley and Zumbo 1996; Messick 1988) have raised concerns over whether validation practice presented in the literature is keeping pace with this evolution in validity theory. The aim of the current study, therefore, was to document current validation practice by examining evidence presented in articles published from 2000 through 2010 in the *Journal of Educational Psychology*. In addition, the study aimed to see whether and how validation practice has changed over the past 50 years by comparing current practice with earlier practice reported in articles published from 1950 through 1960.

R.J. Collie (✉)
School of Education, University of New South Wales, Sydney, NSW 2052, Australia
e-mail: rebecca.collie@unsw.edu.au

B.D. Zumbo, Ph.D. (✉)
Measurement, Evaluation, and Research Methodology (MERM) Program,
Department of Educational and Counseling Psychology, and Special Education,
The University of British Columbia, 2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
e-mail: bruno.zumbo@ubc.ca

B.D. Zumbo and E.K.H. Chan (eds.), *Validity and Validation in Social, Behavioral, and Health Sciences*, Social Indicators Research Series 54, DOI 10.1007/978-3-319-07794-9_7,

The Evolution of Validity Theory

Validity theory has changed greatly from the 1950s until the present. Up to and during the early 1950s, validity was generally considered under a criterion-based model of validity (Kane 2001; however, see Rulon 1946). This view commonly involved correlating a test with an external criterion measure; if the test correlated highly with the criterion, then it was considered valid (Goodwin and Leech 2003; Jonson and Plake 1998). During this time, the validity of the test itself was the primary concern (Goodwin and Leech 2003), and the degree to which a test measured what it was purported to measure was the key to validity (Kane 2001). In addition, "test validity was a singular concept" (Jonson and Plake 1998, p. 737).

In 1952, the *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA Committee on Test Standards) was published. It introduced "four categories of validity: predictive validity, status validity, content validity, and congruent validity" (Sireci 1998, p. 88). By 1954, when the first version of the Standards (called the *Technical Recommendations for Psychological Tests and Diagnostic Techniques*; APA) were published, the names were changed slightly to "'types' or 'attributes' of validity" (Sireci 1998, p. 88) including predictive, concurrent (previously status validity), content, and construct validity (previously congruent validity). According to Sireci (2009), it was with the publication of the 1954 Standards that "the concept of 'construct validity' was born" (p. 24). It was, however, in a seminal paper by Cronbach and Meehl (1955) that this was further elaborated. Indeed, it was in a response to that paper that Loevinger (1957) was the first to argue that construct validity is the whole of validity.

Over the decades since the first Standards (APA 1954) were published, many further changes have taken place. These have included the forgoing of *types of validity* in favor of *types of evidence* under a unitary view of validity (Jonson and Plake 1998), as well as a change in the view of validity from a property of the test to a characteristic of the test scores or inferences (Goodwin and Leech 2003). The current state of validity theory is visible in the current Standards (AERA et al. 1999), which mention five types of validity evidence. The first is evidence based on test content, which addresses "the extent to which the content of a measure represents a specified content domain" (Goodwin and Leech 2003, p. 183). The second is evidence based on response processes, which examines the processes in which participants' engage in order to respond to test questions to understand why certain responses were given by certain groups (AERA et al. 1999) and to see if they correspond with the construct being measured (Goodwin and Leech 2003). The third is evidence based on internal structure, which examines "the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (AERA et al. 1999, p. 13). The fourth is evidence based on relations to other variables, which analyses the "relationship of test scores to variables external to the test" (AERA et al. 1999, p. 13) and refers to traditional criterion-related validity and traditional aspects of

construct validity such as convergent and discriminant validity (Goodwin and Leech 2003). The final type is evidence based on testing consequences, which refers to the intended and unintended consequences that impact validity through construct-irrelevant variance and construct underrepresentation (AERA et al. 1999; Hubley and Zumbo 2011).

The Current Study

Despite great changes in validity theory over the past 60 years, questions remain as to whether validation practice presented in published articles has kept pace with these changes. The current study, therefore, examined articles published recently in the Journal of Educational Psychology (JEP) in an attempt to document current practice. More specifically, validation practice presented in articles published from 2000 through 2010 was examined. The current study also investigated whether and how validation practice has changed over time. To do that, comparisons were made between past and current practice to identify the degree to which practice has (or has not) changed over time. This type of examination is important given changes to the Standards (e.g., AERA et al. 1999), as well as changes in validity theory (Sireci 2009; Zumbo 2007) over the past 50 years. To do this, a second time period was also included in analyses. Validation practice reported in articles published from 1950 through 1960 was examined in addition to the contemporary articles. This earlier period was chosen for the comparison because it was a time when long held beliefs about validity (e.g., the focus on criterion-related validity) were being actively questioned and the concept of construct validity was first proposed.

In order to conduct the current study, a framework was used that stems from Cizek et al.'s (2008) study that examined validation practice reported in evaluations of measures in the 16th Mental Measurements Yearbook (Spies and Plake 2005). Two overarching research questions guided the current study:

1. What is the nature of current validation practice presented in articles as it pertains to (a) validity characteristics; (b) different sources of validity evidence; (c) number of different sources of validity evidence; and (d) justification for and types of criterion-related predictive, criterion-related concurrent, convergent, or discriminant variables?
2. To what degree has validation practice changed from articles published in 1950–1960 to those published more recently?

Data Source

To obtain data for this study, we conducted a search of articles published in JEP through the online PsycARTICLES database. Issues published between 2000 and 2010, as well as issues published between 1950 and 1960 were included in the

search. Articles that had the term *validity* or *validation* in their abstracts were included in the initial sample. For 2000–2010, this search returned 30 articles. Eleven of these articles had very little to do with validity or validation (e.g., it was mentioned once or twice, but no evidence was provided) or were using a different meaning of the word (e.g., the validity of drawings compared to real life). This left 19 articles as data for the 2000–2010 time period. For 1950–1960, the search returned 29 articles. Again, these were examined based on the content and 13 were excluded because they were theoretical articles about validation or they had very little to do with validity or validation. In total, 35 articles were utilized as data sources. The appendix contains references to all articles examined in the study.

Methods

The articles were examined for their presentation of validity evidence using a similar method to that which Cizek et al. (2008) used in their study. This involved documenting validation practice including the sources of validity evidence provided in the articles, how validity was characterized, as well as several other analyses. However, where Cizek et al. examined reviews of educational and psychological tests in the Mental Measurements Yearbook (Spies and Plake 2005), the current study examined articles on educational and psychological measures published in JEP. Therefore, in order to best answer the research questions, certain categories of examination were excluded (e.g., test type), while additional categories were added (e.g., reference to validity experts, the justification for choice of comparison variables).

Categories of Examination

The first category examined was *validity characteristics*. Four indicators were examined for validity characteristics including whether articles presented a unified or separated view of validity, made reference to validity as a property of a test or property of the inferences of a test, made reference to any version of the Standards (AERA et al. 1985, 1999; APA et al. 1966, 1974; APA 1954), and made reference to experts and/or seminal validity papers. The second category examined was *sources of validity evidence*. Indicators examined in this category were adapted from the current Standards (AERA et al. 1999) and Cizek et al.'s (2008) study. As such, seven sources of evidence were examined: evidence based on response processes, consequence of testing, test content, internal structure, predictive relations to other variables, concurrent relations to other variables, and construct. Two of these sources were further refined as per Cizek et al.'s (2008) study. First, we examined whether internal structure was reported as evidence of reliability, validity, or as reliability evidence that informs validity. Second, we analyzed four

components of construct evidence: whether the article mentioned the term *construct validity*, undertook factor analysis (FA) or structural equation modelling (SEM) to explore constructs, mentioned convergent validity, or mentioned discriminant validity. The third category examined the *number of sources* of validity evidence reported in each article. This simply involved counting how many different sources of evidence each article accurately reported. The final category examined *comparison variables*, which refer to criterion-related predictive, criterion-related concurrent, convergent, and discriminant variables. Two indicators were examined for this category: whether justification was provided for the choice of comparison variables and the types of comparison variables used in each article.

Results

The results are organized by category of examination (e.g., validity characteristics, sources of evidence). The first results reported are for current practice (i.e., articles published in 2000–2010). Following this, comparisons across the two time periods are made (i.e., 1950s versus 2000s). It is important to note that the results are based on accurate validation practice presented in the articles. This means that if articles claimed to provide a certain source of evidence, but did not follow through accurately they were not coded in that category. Among the sample, only two articles fell into this group. Both claimed to provide criterion-related predictive validity evidence, but did not use a criterion variable and/or did not use one that was measured in the future. Thus, they were not coded as presenting criterion-related predictive evidence. In addition to those articles, there were several others that accurately reported a certain source of validity evidence despite not naming it as such. In total seven articles from 1950 to 1960 and eight from 2000 to 2010 were coded as reporting a source of evidence despite not naming it in the article. For example, Mokhtari and Reichard (2002) developed a new measure called the Metacognitive Awareness of Reading Strategies Inventory by obtaining expert opinion from researchers in the field of reading strategies on the "clarity, redundancy, and readability" of the items (p. 251). Thus, although they did not call it evidence based on test content, it was classified as such given that in the current Standards (AERA et al. 1999) evidence based on test content can come from judgments by experts in the area on the relationships between items in the test and the construct.

Documenting Current Practice

The first aim of this study was to document current practice. In order to do that, the articles published between 2000 and 2010 were examined for presentation of the various categories under examination: validity characteristics, sources of validity evidence, number of sources, and comparison variables. These are discussed in turn below.

Validity Characteristics

The first category under examination concerned the characterization of validity presented in the articles (Research Question 1a). As described above, four indicators were used to assess this category. The first indicator examined whether articles presented a unitary or separated view of validity. Analyses revealed that the majority of articles reported multiple types of validity (63 %). Furthermore, 42 % of articles published at this time mentioned types of validity that are not considered in the current or previous Standards (e.g., AERA et al. 1999). These types of validity included face, internal, external, postdictive, ecological, and factorial validity. For example, d'Ailly (2003) mentioned predictive, concurrent, construct, as well as ecological validity. Several articles (26 %) referred to construct validity only and/or other types of validity as a part of construct validity. For example, "...the present study is a construct validity investigation of the [scale] by independent researchers to explore the underlying dimensions of reading motivation as assessed by the [scale]" (Watkins and Coffey 2004, p. 111). We chose to code these views of validity as unitary given that they refer to the modern view that all evidence bears upon construct validity (Sireci 2009). However, none of the recently published articles explicitly reported a unitary view of validity.

The second indicator examined whether articles referred to validity as a property of the test or as a property of the inferences of the test. The overwhelmingly majority of articles (95 %) referred to validity as a property of a test, and in some cases the property of a model (e.g., Janosz et al. 2000). For example, "...[these findings] present compelling evidence for the validity and utility of the [Academic Entitlement] scale" (Chowning and Campbell 2009, p. 994). Only one article (3 %) referred to validity as a characteristic of the inferences. This article stated, "construct validity of the interpretation of this difference as a diffusion effect was supported by comments by both the teachers themselves and by external observers" (Craven et al. 2001, p. 643).

The final two indicators for this category concerned references made to the Standards (e.g., AERA et al. 1999) and validity experts. Although the examined articles featured validation as a main part of their content, not one article made reference to the current Standards (AERA et al. 1999) or a previous version (e.g., AERA et al. 1985). More promising, however, was that six articles (32 %) made reference to one or more experts.

Sources of Validity Evidence

For the second category, articles were examined for whether they accurately presented any of the seven different sources of validity evidence: evidence based on response processes, consequences, internal structure, content, predictive

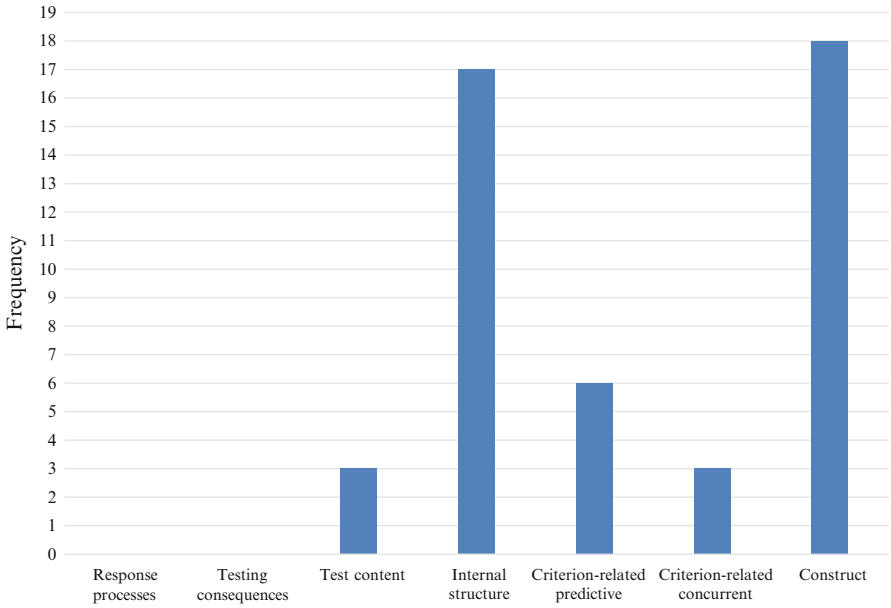

Fig. 7.1 Sources of validity evidence presented in articles published between 2000 and 2010

relations to other variables (i.e., criterion-related predictive), concurrent relations to other variables (criterion-related concurrent), and construct (Research Question 1b). The frequency with which the different sources of evidence were presented in articles published in 2000–2010 is shown in Fig. 7.1. As the figure shows, most articles reported construct evidence (95 %) and internal structure (89 %). Criterion-related predictive (32 %), criterion-related concurrent (16 %), and content evidence (16 %) were also reported in some articles. However, no articles reported evidence based on response processes or testing consequences.

Characterization of Internal Structure

In addition to coding whether articles reported evidence of internal structure, we also examined their characterization of internal structure as reliability evidence only, validity evidence only, or as reliability evidence that informs validity. Of the articles, most (89 %) reported evidence of internal structure (e.g., Cronbach's alphas). In all of these articles, internal structure was reported only as reliability. For example, "we examined the internal consistency reliability… of each of the factors constituting the given model" (Brockway et al. 2002, p. 215). In other words, none of the recent articles reported reliability as informing validity.

Components of Construct Evidence

As described above, four components of construct evidence were also documented: whether the article mentioned the term *construct validity*, undertook factor analysis (FA) or structural equation modelling (SEM) to explore constructs, provided convergent evidence, or provided discriminant evidence. All except one article reported at least one component of construct validity. The term, *construct validity*, was the most frequently reported component—84 % of articles mentioned construct validity. This was followed by reports of FA or SEM (68 %), convergent evidence (42 %) and discriminant evidence (37 %).

Number of Sources

For the third category, accurate reports of validity evidence were examined to ascertain the number of different sources of validity evidence that each article reported (Research Question 1c). Over half of the article reported one source of evidence (63 %); however, two sources (21 %, 4 articles), three sources (11 %, 2 articles), and four sources (5 %, one article) were also reported in several articles. For the article that reported four sources of evidence (i.e., Brockway et al. 2002), evidence based on test content, criterion-related predictive, criterion-related concurrent, and construct was provided.

Comparison Variables

Comparison variables were examined based on two indicators: whether articles provided justification for their choice of comparison variables and the types of comparison variables that articles used for providing evidence of validity (Research Question 1d). For the first indicator, we assessed articles on whether they provided *convincing justification*, *unconvincing justification*, or *no justification* for their choice of comparison variables. Convincing justification included a description of why the variable was chosen, whereas unconvincing arguments explained what the comparison variable was without justifying its choice or providing inaccurate reasoning for the choice of variable. Although not specifically mentioned in the current Standards (AERA et al. 1999), the provision of justification is not only good practice, but also necessary if criterion-related claims are to be upheld and clearly understood as evidence for validity.

Of the articles, 13 articles made reference to a comparison variable. Among these, eight provided no justification, two provided a convincing justification, and three provided an unconvincing justification. For one of the convincing justifications, Edwards and Schleicher (2004) explained how the variable of interest (tacit knowledge) had been correlated with the criterion variable (performance) in previous literature and what steps were needed to provide more criterion-related predictive validity evidence. For the unconvincing justifications, the articles did not

justify why the variables were chosen. For example, "Correlations between the newly developed... subscales and the published scales were examined for convergent and divergent [sic] validity" (Chowning and Campbell 2009, p. 984).

For the second indicator, we examined the types of comparison variables that articles used. Convergent (42 %), discriminant (37 %), and criterion-related predictive (32 %) variables were most frequently reported. In addition, criterion-related concurrent variables were mentioned by three articles (16 %).

Comparisons with Earlier Practice

The second aim of this study was to compare current practice with earlier practice in order to understand the degree to which validation practice has changed over time. Comparisons were made between recent articles and those published around half a century ago (from 1950 through 1960). Results are described below.

Validity Characteristics

As noted above, four indicators were used to assess validity characteristics. Table 7.1 shows the results for the four indicators by decade and in total. Before comparing the results across the two periods, it is important to present the findings from articles published in the earlier time period. As Table 7.1 shows, 69 % of articles published in 1950–1960 mentioned validity as a stand-alone concept; however, this pre-dates the contemporary unitary view of validity and refers to a singular entity—often a coefficient—that was considered proof of validity before 'types' of validity became prominent. For example, "As can be seen from a comparison of the validity coefficients for the two forms of the scale... differences in validity of the two types of response are negligible" (Neidt and Merrill 1951, p. 435). Of the other articles from that time period, 19 % mentioned multiple types of validity. These articles were published in the latter half of the 1950s and reflect the terminology change towards *types of validity* that occurred after the first Standards (APA 1954) were published. The remaining articles were unclear as to their characterization of validity. In comparing these results with those from the 2000–2010 articles, we see that the characterization of validity has changed. Contemporary articles were more likely to report 'all evidence as bearing on construct validity' (from no articles in 1950–1960 to 26 % of articles in 2000–2010). In addition, many more articles referred to multiple types of validity (from 19 % of articles in 1950–1960 to 63 % of articles in 2000–2010). However, no articles from either time period explicitly reported a unitary view of validity.

For the second indicator (i.e., validity as a property of the test versus a property of the inferences), almost all articles (94 %) published in 1950–1960 referred to validity as a property of a test. For example, "the validity of an English

Table 7.1 Characterization of validity

	1950–1960	2000–2010	Total
View of validity			
Unitary			
Explicit statement of a unitary view	0	0	0
All evidence bearing on construct validity	0	5	5
Not unitary			
Multiple types mentioned	3	12	15
Only validity mentioned	11	0	11
Unclear	2	2	4
Property of test or inferences			
Test	15	18	33
Inferences	0	1	1
Unclear	1	0	1
Reference to test standards			
Current version	0	0	0
Previous version	0	0	0
No reference made	16	19	35
Reference to experts			
Yes	2	6	8
No	14	13	27

Examination for Foreign Students was tested against the criterion of final grades. . .” (Lorge and Diamond 1954, p. 214). The only article that did not refer to validity as a property of a test was unclear in its characterization and so could not be coded. Comparing these results with current practice, we see that very little has changed. Almost all articles in both time periods refer to validity as a property of a test: 94 % of articles in 1950–1960 and 95 % of articles in 2000–2010.

The final two indicators for this category were concerned with references made to the Standards (e.g., AERA et al. 1999) and experts. Similar to the findings among recent articles, no articles published in 1950–1960 made reference to a previous version of the Standards (e.g., APA 1954). In terms of references made to experts, results suggest some change in validation practice. Only two articles (13 %) published in 1950–1960 made reference to one or more experts; however, this increased to six articles (32 %) published in 2000–2010. Experts cited included Messick (1995; cited three times), Campbell and Fiske (1959; cited twice), Cook and Campbell (1979; cited twice), Cronbach (1949; cited once), Cronbach and Meehl (1955; cited once), and Messick (1989; cited once).

Sources of Validity Evidence

The frequency with which the different sources of evidence were presented within each time period and in total is shown in Table 7.2. Among articles published in 1950–1960, there were several sources of evidence that were reported with similar

Table 7.2 Sources of validity evidence accurately reported

	1950–1960	2000–2010	Total
Response processes	0	0	0
Testing consequences	0	0	0
Test content	0	3	3
Internal structure	8	17	25
Criterion-related predictive	10	6	16
Criterion-related concurrent	7	3	10
Construct	2	18	20

Table 7.3 Characterization of internal structure

	1950–1960	2000–2010	Total
As reliability only	7	17	24
As validity only	0	0	0
As reliability and validity	1	0	1
No evidence	8	2	10

frequencies. The most frequently reported source of evidence was criterion-related predictive evidence (63 %). This was followed closely by internal structure (50 %) and criterion-related concurrent (44 %) sources of evidence. Construct evidence (13 %) was also reported, although less frequently. In comparing these results with current practice, we see that there has been an increase in reports of construct evidence (from 13 % of articles in 1950–1960 to 95 % of articles in 2000–2010) and internal structure (from 50 % of articles in 1950–1960 to 89 % of articles in 2000–2010). There was also a decrease in the two types of criterion-related evidence since the 1950s: from 63 % in 1950–1960 to 32 % in 2000–2010 for criterion-related predictive evidence and from 44 % in 1950–1960 to 16 % in 2000–2010 for criterion-related concurrent evidence.

Characterization of Internal Structure

Table 7.3 shows the characterization of internal structure across decades and in total. From 1950 to 1960, eight (50 %) of the articles provided evidence of internal structure. Of these, only one article published in 1960 reported reliability as a component of validity. In that singular case, the authors provided reliability coefficients and explained, "to further explore the question of validity of the need measures, coefficients of internal consistency of two of these measures were computed" (Uhlinger and Stephens, 1960, p. 263). All of the remaining articles that reported internal structure presented it as reliability only (i.e., not as part of validity). Comparisons across decades reveal that the vast majority of articles were and still are reporting internal structure as reliability only. However, a greater proportion of recent articles reported internal structure in any form suggesting greater uptake of this practice over time (from 50 % of articles in 1950–1960 to 89 % of articles in 2000–2010).

Table 7.4 Components of construct evidence

	1950–1960	2000–2010	Total
Construct validity mentioned	0	16	16
FA or SEM conducted	0	13	13
Convergent	2	8	10
Discriminant	0	7	7

Table 7.5 The number of different sources of validity evidence presented in articles

	1950–1960	2000–2010	Total
One source	13	12	25
Two sources	3	4	7
Three sources	0	2	2
Four sources	0	1	1

Components of Construct Evidence

Table 7.4 shows the four additional components of construct evidence that we examined and the frequency with which they were reported within the two time periods and in total. In 1950–1960, convergent evidence was the only reported component of construct evidence and it was reported in only two articles (13 %). A comparison of practice across the decades reveals that a great deal more articles reported the components of construct evidence in the recent articles and this includes reports of all four components (e.g., 84 % mentioned construct validity; see Table 7.4).

Number of Sources

Table 7.5 shows the number of sources reported by decade and in total. In 1950–1960, most articles reported only one source of evidence (81 %), with the remaining reporting two sources of evidence (19 %). Comparing these results with recent articles, we see that most articles still only reported one source of evidence. However, a greater proportion of recent articles reported multiple sources suggesting some change in practice over time (from 19 % of articles in 1950–1960 to 37 % of articles in 2000–2010).

Comparison Variables

We also assessed articles on whether they provided convincing justification, unconvincing justification, or no justification for their choice of comparison variables. All the articles published in 1950–1960 mentioned at least one comparison variable;

Table 7.6 Types of comparison variables reported

	1950–1960	2000–2010	Total
Criterion-related predictive	10	6	16
Criterion-related concurrent	7	3	10
Convergent	2	8	10
Discriminant	0	7	7

however, most provided no justification (81 %, 13 articles). Only two articles provided a convincing justification (13 %). For example,

> Despite the limitations of teachers' grades as statistical variables, we must recognize that grades are criteria in a very real sense—they are actually the principal evaluation in most school situations. It is therefore essential that tests intended as predictors be correlated with grades. (Doppelt and Wesman 1952, p. 210)

In addition, one article (6 %) provided an unconvincing argument by describing the comparison variable, but not justifying why it was chosen: "Freshman grades in college, or honor point ratio, were used as the criterion of scholastic achievement" (Holland 1959, p. 136). Comparing these results with recent articles, we see practice has only changed slightly with most articles still providing no justification (81 % of articles in 1950–1960 and 62 % of articles in 2000–2010).

Table 7.6 shows the types of comparison variables that articles reported. In 1950–1960, criterion-related predictive (63 %) and criterion-related concurrent (44 %) variables were most frequently reported types of variables. Convergent variables were also reported less frequently (13 %), whereas discriminant variables were not mentioned in any articles at this time. Comparisons between practice in the two time periods reveal an increase in reports of convergent (13 % in 1950–1960 to 42 % in 2000–2010) and discriminant evidence (from none in 1950–1960 to 37 % in 2000–2010) and a decrease in reports of the two types of criterion-related evidence over time (from 63 % in 1950–1960 to 32 % in 2000–2010 for predictive evidence and from 44 % in 1950–1960 to 16 % in 2000–2010 for concurrent evidence).

Discussion

The first aim of the current study was to document validation practice reported in validation articles published in JEP from 2000 to 2010 (Research Question 1). The results revealed that current practice reflects modern validity theory in several ways. In particular, most of the recent articles reported construct evidence, which reflects contemporary ideas about construct evidence being the whole of validity (Sireci 2009). In addition, most of the articles reported internal structure evidence. However, the results also revealed several ways in which validation practice did not reflect the current Standards (e.g., AERA et al. 1999) and modern validation theory (e.g., Zumbo 2007). For example, most articles referred to multiple types of validity and explained that validity was a property of a test. Furthermore, no articles reported evidence based on response processes or the consequences of testing.

The second aim of the current study was to compare practice across two time periods to see whether and how validation practice has changed over time (Research Question 2). Results revealed that the recent articles more regularly cited relevant experts, and used a wider variety of comparison variables. However, in several other categories practice was very similar in the two time periods. In particular, in both time periods the Standards (e.g., AERA et al. 1999) were not referenced, internal structure was reported as reliability evidence only (not as also informing validity), only one source of evidence was generally provided, and justification was rarely provided for comparison variables. From this, we can conclude that validation practice across the two time periods is similar in many ways despite the passing of 40–50 years and the publishing of four test standards (AERA et al. 1985, 1999; APA et al. 1966, 1974) during that time. Three major findings and their implications are discussed below.

Response Processes and Consequences of Testing

The results revealed that none of the articles published in 2000–2010 reported evidence of response processes or testing consequences. Given that validation occurs through a process of accumulation of evidence (AERA et al. 1999), it is not necessary for articles to include all sources of validity evidence. At the same time, however, the fact that no articles reported these two sources of evidence suggests that they may be being ignored. This has important implications.

As described above, evidence based on response processes refers to examinations of participants' responses and why they chose those responses (AERA et al. 1999). Response processes are helpful for understanding whether there are major individual differences in processes for answering questions, why this may have occurred, and how this may affect the responses (AERA et al. 1999). For example, evidence based on response processes can reveal differences in interpretations of test questions across different subgroups of participants. This is important for understanding whether the questions are accurately measuring the construct or whether some other type of variance is causing differences in scores (e.g., different meanings among different subcultures). It can also provide understanding of why this occurs, which can be used to create better instruments that more accurately capture the construct or knowledge under examination across different subgroups.

Examinations of response processes are also useful for developing definitions of a construct by revealing understanding about how it is interpreted by participants. This can also help ensure that participants are interpreting the questions as expected and, in turn, that their responses reflect the construct or knowledge that the researcher is attempting to examine. An example of this appeared in Gaderman et al.'s (2011) research. They examined the response processes of children as they answered the Satisfaction with Life Scale adapted for Children (Gaderman et al. 2011). The analyses revealed greater understanding of the construct by

showing that the children used strategies to answer the questions that were meaningful and theoretically consistent with the literature.

Also important is evidence based on test consequences. As described above, evidence based on test consequences aims to investigate the intended and unintended consequences of testing (AERA et al. 1999). As a broad statement, this type of evidence is important for considering construct irrelevant variance so that, as Hubley and Zumbo (2011) note, based on construct delineation and definition intended social and personal consequences and unintended social and personal side effects emerge. As summed up by Shepard (1997), "consequences are evaluated in terms of the intended construct meaning" (p. 8). With the domain of educational psychology in mind, as Hubley and Zumbo (2011) suggest, when the social consequences and side effects of using an educational psychology measure are not congruent with our societal values and goals regarding that particular psychological domain such insights in the validation process may be used to adjust constructs, theories, and aspects of the measurement process until the desired congruence between purposes, goals, values, and consequences is accomplished. For example, consider the construct of self-efficacy, which refers to individuals' beliefs about their capabilities in a prospective and specific context (Bandura 1982, 1993). When researchers measure this construct, a consequence may be the promotion of self-efficacy as a positive characteristic in educational and developmental contexts. This may be an intended consequence given that research has highlighted self-efficacy is important for positive outcomes among individuals (e.g., greater achievement among students; Caprara et al. 2006). However, it is important to note that self-efficacy is not necessarily positive at very high levels.

For example, Brenner et al. (2012) examined the development of self-efficacy for teaching among student teachers as they engaged in their practicum placement in schools during their teacher education program. Through their analyses, they described one student teacher who reported consistently high levels of self-efficacy for teaching despite receiving low ratings of effectiveness by her faculty advisor (i.e., the faculty member who assessed student teachers' progress). Moreover, Brenner et al. explained that the student teachers' high levels of self-efficacy may have prevented her from realistically assessing her abilities and putting in the necessary effort to improve her teaching skills. This example highlights that too much self-efficacy can, in fact, be a negative such that individuals may not feel the need to work on improving their own practice. When researchers assess constructs like self-efficacy, one consequence is that the construct becomes valued and participants may feel that they need to experience high levels of it. In most cases, this may be positive. Nevertheless, it is a consequence and should be considered.

When researchers do not consider evidence based on response processes and testing consequences, the implications potentially include a weaker understanding of the construct under examination and the promotion of certain outcomes among participants (that may or may not be positive). Clearly, greater efforts to include these types of evidence in validation practice are needed.

Reference to Standards

The second major finding to be discussed refers to references that articles made to the Standards (e.g., AERA et al. 1999). Considering that a proposal for test standards was first published in 1952 (APA, Committee on Test Standards) and that the first set of test standards were created in 1954 (APA), it is understandable that articles published before and shortly after this time (i.e., between 1950 and 1960) did not cite any test standards. However, all the articles published between 2000 and 2010—when five different versions of the test standards (AERA et al. 1985; APA et al. 1966, 1974; APA 1954) have been published—did not cite any version of the Standards. This result highlights a disconnect between validity theory and contemporary validation practice. If researchers do not cite the Standards in their publications, then it is questionable as to whether they are consulting them in their research, which leads to further concerns about the dissemination of validity theory.

There are two plausible reasons for why this finding occurred. First, it is possible that researchers are simply not familiar with the current Standards (AERA et al. 1999). If this is the case, there is a need for efforts that endeavor to increase the visibility of the current Standards and to educate researchers about how to use them in their practice. The second reason is that authors may be aware of the current Standards, but not feel it is necessary to make reference to them. For this possibility, journals like JEP and their editors may want to raise expectations about what constitutes accurate validation practice by requiring articles to accurately address modern validity theory in their validation practice. Clearly, more visibility is needed to raise general awareness of the Standards. However, expectations must also be raised to ensure that visibility is followed through with accurate practice.

Validity Characterization

The third key finding prevalent among the recent JEP validation articles was the presentation of several characteristics of validation practice that do not conform to modern validity theory. In particular, the view of validity as having multiple types and the conceptualization of validity as the property of a test highlight that validation practice has not kept pace with the changes in core aspects of validity theory. This is particularly concerning given that the unitary view of validity was first articulated in the 1950s by Loevinger (1957) and first presented in the 1985 Standards (AERA et al.). Furthermore, validity as adhering to inferences rather than a test was introduced in the 1985 Standards (Goodwin and Leech 2003). In other words, despite the fact that these aspects of validity theory had been published in the Standards (AERA et al. 1985, 1999) for 15–25 years when the recent articles were published, researchers are still using outdated practices.

As to why this occurred, it is possible that it relates to semantic differences (Cizek et al. 2008). Cizek and colleagues' (2008) suggest that despite the different nomenclature, *sources of evidence* and *types of validity* may actually have the same underlying meaning for researchers. However, it is also possible that this reflects confusion about the meaning and nature of validity. Given that both the current study and other studies (e.g., Cizek et al. 2008) have found the same misconceptions about types of validity and validity as a property of a test, this is an issue that warrants further consideration and investigation. Furthermore, validity theorists may want to carefully consider the language they use to explain validity theory in order to emphasize the importance of specific terminology, as well as the meaning underlying it. Clearer language and a better understanding of why certain language is used will likely help to increase the accessibility of the theory for researchers.

Another plausible reason for the prevalence of outdated practice is that researchers are unaware of how to provide evidence that conforms to modern validity theory. Instead of referring to the Standards (e.g., AERA et al. 1999), researchers may refer to existing examples of validation practice in the literature. When these examples utilize older theory and practices of validation, this creates a cycle of outdated practice informing more outdated practice. In order to remedy this, examples that are in line with contemporary validity theory are needed. As noted above, journals and journal editors may want to raise their expectations regarding the validation practice that is published in journals. By expecting researchers to report validation practice that is based in modern validity theory, more accurate examples will begin to appear in the literature. This will ideally disrupt the current cycle to create a shift towards better dissemination and practice.

Limitations

There are several limitations to the current study that must be discussed. First, the study examined 35 articles published in JEP between 1950 and 1960, and 2000 and 2010. Given the small sample, the generalizability of the results is limited to validation articles published in JEP around these two times. The second limitation is that we interpreted articles' provision of validity evidence as the position of the article, when it may in fact reflect an editorial position. Authors may have been discouraged from including too much information on validity to ensure the readability of the article, or they may have excluded it from their submissions for fear of being rejected from publication for being overly psychometric. The nature of our examination did not allow us to determine the reasons behind authors' reports of validation practice. For this reason, rather than referring to the authors we have referred to the articles, as they, not the author's perspectives were the data.

Conclusions

The current study has shown that despite great changes in validity theory and the Standards (e.g., AERA et al. 1985, 1999) over the past half century, current validation practice (presented in JEP articles) does not appear to wholly conform to modern validity theory. This is an issue that was raised by Messick (1988) two decades ago, and by Borsboom et al. (2004) more recently: "The concept that validity theorists are concerned with seems strangely divorced from the concept that working researchers have in mind when posing the question of validity" (p. 1061). The main conclusion from the current study, therefore, is that more must be done to help validation practice draw level with validity theory.

Although JEP is not a journal that focuses on psychometrics or psychological measurement, it is a journal published by the American Psychological Association (APA), a key author in all versions of the Standards (AERA et al. 1985; APA et al. 1966, 1974; APA 1954). Despite this, the Standards' recommendations do not appear to be required practice in JEP validation articles. Perhaps this reflects the different demographic of readers of JEP, or as suggested in the limitations above, the position of an editor. However, it certainly provides an interesting question regarding the practice of modern validity theory: If the APA, as a key organization in the creation of the Standards, does not require that authors follow the Standards in one of their own publications, then is it really surprising that there is such a disparity between theory and practice across the field? Considering the many experts who have questioned the gap between theory and practice (e.g., Borsboom et al. 2004; Hubley and Zumbo 1996; Messick 1988), this is certainly a question that needs to be addressed. As noted above, we recommend that journals raise their expectations, especially journals published by the AERA, APA, and NCME, so that validation articles accurately adhere to modern validity theory. This will not only provide examples for other researchers, but it will aid in the much-needed dissemination of the modern conceptualization of validity theory so that the gap between theory and practice is reduced.

Appendix

References of Articles Published Between 1950 and 1960

Ausubel, D. P., Schiff, H. M., & Zeleny, M. P. (1954). Validity of teachers' ratings of adolescents' adjustment and aspirations. *Journal of Educational Psychology, 45*, 394–406. doi:10.1037/h0056091.

Brown, W. F., & Holtzman, W. H. (1955). A study-attitudes questionnaire for predicting academic success. *Journal of Educational Psychology, 46*, 75–84. doi:10.1037/h0039970.

Chappell, T. L. (1955). Note on the validity of the army general classification test as a predictor of academic achievement. *Journal of Educational Psychology, 46*, 53–55. doi:10.1037/h0044316.

Doppelt, J. E., & Wesman, A. G. (1952). The differential aptitude tests as predictors of achievement test scores. *Journal of Educational Psychology, 43*, 210–217. doi:10.1037/h0060030.

French, J. W. (1958). Validation of new item types against four-year academic criteria. *Journal of Educational Psychology, 49*, 67–76. doi:10.1037/h0046064.

French, J. W. (1959). The relationship of home and school experiences to scores on achievement tests. *Journal of Educational Psychology, 50*, 75–82. doi:10.1037/h0047991.

Gage, N. L. (1957). Logical versus empirical scoring keys: The case of the MTAI. *Journal of Educational Psychology, 48*, 213–216. doi:10.1037/h0047795.

Holland, J. L. (1959). The prediction of college grades from the California psychological inventory and the scholastic aptitude test. *Journal of Educational Psychology, 50*, 135–142. doi:10.1037/h0041909.

Lodge, W. J. (1951). A validity study of personality questionnaires at the upper elementary grade level. *Journal of Educational Psychology, 42*, 21–30. doi:10.1037/h0061737.

Lorge, I., & Diamond, L. K. (1954). Validity of an objective examination for the placement of foreign students in English courses. *Journal of Educational Psychology, 45*, 208–214. doi:10.1037/h0053872.

Neidt, C. O., & Merrill, W. R. (1951). Relative effectiveness of two types of response to items of a scale on attitudes toward education. *Journal of Educational Psychology, 42*, 432–436. doi:10.1037/h0056066.

Papavassiliou, I. T. (1953). The validity of the goodenough draw-A-man test in Greece. *Journal of Educational Psychology, 44*, 244–248. doi:10.1037/h0057111.

Schultz, D. G. (1954). Item validity and response change under two different testing conditions. *Journal of Educational Psychology, 45*, 36–43. doi:10.1037/h0059845.

Scott, O., & Brinkley, S. G. (1960). Attitude changes of student teachers and the validity of the Minnesota Teacher Attitude Inventory. *Journal of Educational Psychology, 51*, 76–81. doi:10.1037/h0040593.

Uhlinger, C. A., & Stephens, M. A. (1960). Relation of achievement motivation to academic achievement in students of superior ability. *Journal of Educational Psychology, 51*, 259–266. doi:10.1037/h0041083.

Woodcock, R. W. (1958). An experimental prognostic test for remedial readers. *Journal of Educational Psychology, 49*, 23–27. doi:10.1037/h0042526.

References of Articles Published Between 2000 and 2010

Bong, M. (2009). Age-related differences in achievement goal differentiation. *Journal of Educational Psychology, 101*, 879–896. doi:10.1037/a0015945.

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*, 170–181. doi:10.1037/0022-0663.98.1.170.

Brockway, J. H., Carlson, K. A., Jones, S. K., & Bryant, F. B. (2002). Development and validation of a scale for measuring cynical attitudes toward college. *Journal of Educational Psychology, 94*, 210–224. doi:10.1037/0022-0663.94.1.210.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*, 891–901. doi:10.1037/0022-0663.98.4.891; 10.1037/0022-0663.98.4.891.supp (Supplemental).

Chowning, K., & Campbell, N. J. (2009). Development and validation of a measure of academic entitlement: Individual differences in students' externalized responsibility and entitled expectations. *Journal of Educational Psychology, 101*, 982–997. doi:10.1037/a0016351.

Craven, R. G., Marsh, H. W., Debus, R. L., & Jayasinghe, U. (2001). Diffusion effects: Control group contamination threats to the validity of teacher-administered interventions. *Journal of Educational Psychology, 93*, 639–645. doi:10.1037/0022-0663.93.3.639.

d'Ailly, H. (2003). Children's autonomy and perceived control in learning: A model of motivation and achievement in Taiwan. *Journal of Educational Psychology, 95*, 84–96. doi:10.1037/0022-0663.95.1.84.

Edwards, W. R., & Schleicher, D. J. (2004). On selecting psychology graduate students: Validity evidence for a test of tacit knowledge. *Journal of Educational Psychology, 96*, 592–602. doi:10.1037/0022-0663.96.3.592.

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*, 613–628. doi:10.1037/0022-0663.100.3.613.

Gathercole, S. E., & Pickering, S. J. (2000). Assessment of working memory in six- and seven-year-old children. *Journal of Educational Psychology, 92*, 377–390. doi:10.1037/0022-0663.92.2.377.

Greene, J. A., Torney-Purta, J., & Azevedo, R. (2010). Empirical evidence regarding relations among a model of epistemic and ontological cognition, academic performance, and educational level. *Journal of Educational Psychology, 102*, 234–255. doi:10.1037/a0017998.

Grigorenko, E. L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E. J., & Sternberg, R. J. (2009). Are SSATS and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101*, 964–981. doi:10.1037/a0015906.

Janosz, M., Le Blanc, M., Boulerice, B., & Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. *Journal of Educational Psychology, 92*, 171–190. doi:10.1037/0022-0663.92.1.171.

Kardash, C. M., & Wallace, M. L. (2001). The perceptions of science classes survey: What undergraduate science reform efforts really need to address. *Journal of Educational Psychology, 93*, 199–210. doi:10.1037/0022-0663.93.1.199.

Legault, L., Green-Demers, I., & Pelletier, L. (2006). Why do high school students lack motivation in the classroom? Toward an understanding of academic amotivation and the role of social support. *Journal of Educational Psychology, 98*, 567–582. doi:10.1037/0022-0663.98.3.567.

Mokhtari, K., & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology, 94*, 249–259. doi:10.1037/0022-0663.94.2.249.

Naglieri, J. A., & Rojahn, J. (2004). Construct validity of the PASS theory and CAS: Correlations with achievement. *Journal of Educational Psychology, 96*, 174–181. doi:10.1037/0022-0663.96.1.174.

Robbins, S. B., Allen, J., Casillas, A., Peterson, C. H., & Le, H. (2006). Unraveling the differential effects of motivational and skills, social, and self-management measures from traditional predictors of college outcomes. *Journal of Educational Psychology, 98*, 598–616. doi:10.1037/0022-0663.98.3.598.

Watkins, M. W., & Coffey, D. Y. (2004). Reading motivation: Multidimensional and indeterminate. *Journal of Educational Psychology, 96*, 110–118. doi:10.1037/0022-0663.96.1.110.

References

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

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

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, 201–238.

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist, 7*, 461–465.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

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

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37*, 122–147.

Bandura, A. (1993). Perceived self-efficacy in cognitive development and functioning. *Educational Psychologist, 28*, 117–148.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Brenner, C. A., Perry, N. E., & Collie, R. J. (2012, September). Student teachers' developing practices that promote self-regulated learning: Linking efficacy and utility beliefs to effectiveness. In N. Perry & B. Kramarski (Co-chairs), *Metacognition and self-regulation in developing professionals*. Symposium presented at the biennial meeting of the European Association for Research on Learning and Instruction, Milan, Italy.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology, 44*, 473–490. doi:10.1016/j.jsp.2006.09.001.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*, 397–412. doi:10.1177/0013164407310130.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1949). *Essentials of psychological testing*. New York: Harper.

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

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2011). Investigating the substantive aspect of construct validity for the satisfaction with life scale adapted for children: A focus on cognitive processes. *Social Indicators Research, 100*, 37–60. doi:10.1007/s11205-010-9603-x.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27*, 197–222.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and Evaluation in Counseling and Development, 36*(3), 181–191.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *Journal of General Psychology, 123*, 207.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research: An International Interdisciplinary Journal for Quality of Life Measurement, 103*, 219–230. doi: http://dx.doi.org/10.1007/s11205-011-9843-4.

Jonson, J. L., & Plake, B. S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement, 58*, 736–753. doi:10.1177/0013164498058005002.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037/0003-066X.50.9.741.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16*, 296.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (Vol. 19, pp. 405–450). Washington, DC: AERA.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*, 5–8, 13, 24.

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

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte: Information Age Publishing.

Spies, R. A., & Plake, B. S. (Eds.). (2005). *The sixteenth mental measurements yearbook*. Lincoln: Buros Institute of Mental Measurements.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Psychometrics, Vol. 26, pp. 45–79). Amsterdam: Elsevier Science.