

# Psychometric Validity: Establishing the Accuracy and Appropriateness of psychometric measures

# Psychometric Validity: Establishing the Accuracy and Appropriateness of Psychometric Measures

David J. Hughes

Alliance Manchester Business School

"The problem of validity is that of whether a test really measures what it purports to measure" (Kelley, 1927, p.14).

"The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another and low for a third" (Cureton, 1951, p. 621).

Validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13).

"A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes" (Borsboom, Mellenbergh, & van Heerden, 2004, p. 1061).

"At its essence, validity means that the information yielded by a test is appropriate, meaningful, and useful for decision making – the purpose of mental measurement" (Osterlind, 2010, p. 89).

*Validity* is widely acknowledged to be "the most fundamental consideration" in the development of psychometric measures (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014, p. 11). The psychometrician's principal goal is to develop valid measures and no self-respecting researcher or practitioner would knowingly use an invalid test - or would they? The answer to that question will vary, perhaps substantially, according to who is answering it and how they choose to define validity. As the quotes above demonstrate, there are a surprising number of conceptualisations of validity, some not even closely analogous. The variety of meanings associated with the word validity means that many see it as a complicated and daunting concept whilst reading that a measure (or an inference based upon a measure) is "valid" is somewhat meaningless unless accompanied by an explicit definition. Despite the conceptual melée surrounding validity, it remains the ultimate aim of psychometric test development and thus, validity remains a central tenet of this Handbook of Psychometric Testing.

Much of the current debate in the validity literature concerns discussions of epistemology or ontology, truth absolute or otherwise, the existence of constructs, the role of belief true or justified, and a number of other philosophical debates (Borsboom et al., 2004; Markus & Borsboom, 2013; Hood, 2009; Kane, 2006, 2016). Equally, validity theorists have been grouped as traditionalists, liberals, conservatives, ultra-conservatives; as expanders, unifiers, partitioners; as realists, anti-realists, constructivists, and pragmatists (e.g., Markus & Borsboom, 2013; Hood, 2009; Newton & Shaw, 2016). The focus of this chapter, however, is to explore the contribution of validity theory in answering two essential psychometric questions:

1. Am I measuring what I want to measure?
2. Is my measure useful?

The remainder of this chapter is broken down into five main sections. First, I will consider why we need the concept of validity and why it is so important for psychometricians. Second, I review the evolution of the concept of validity focusing on seminal works throughout the last century. Third, I consider the surprisingly tumultuous state of validity theory and practice within the 21st century. Fourth, I consider suggestions for advancing validity theory and provide my own model designed to improve the theoretical clarity and practical utility of validity theory. Finally, I provide practical guidance regarding what forms of 'validity evidence' there are, which of the questions each form of evidence can help us answer, and what tools are available to generate such evidence.

**Why do we need validity?**

Develop a theory, design a measure, collect data using the measure, test the theory, and if the theory is supported, then use your measure to make decisions based on the theory. In my view, this is science. The same broad approach applies whether examining gravity (or

gravitational waves) through to the design of a spaceship, or whether examining cognitive ability through to placing a student within a particular school class. There are three main elements here: theory, measurement, and decisions. Theory and measurement are the core aspects of any science, social or otherwise, and if we want to use our theories and measures, we must also consider our decision-making processes. As Cone and Foster (1991) note, "measurement provides the foundation for all other scientific pursuits" with "developments in all areas of science follow[ing] discoveries of appropriate measurement techniques" (p. 653). Psychometrics are measures, their development does indeed contribute to theory development (see Booth & Murray, Chapter 29), and they are often used to make important real-world decisions (e.g., who to hire; Hughes & Batey, in press). Validity is the word commonly assigned to describe evaluations of psychometric measurement and decisions made based on this measurement. Primary within this arena are two questions: are you measuring what you think you are measuring? If so, are your measures useful for decision-making?

Establishing what a psychometric measures, it turns out, is a rather difficult task. Most measurement is directed towards constructs that we cannot directly observe, such as attitudes, mental health, knowledge, executive functions, personality traits, political preferences, culture, cognitive biases, and motives. In circumstances when we cannot directly access the construct of interest, we must observe theoretically relevant behaviours and infer from these observations the existence and nature of the underlying construct (c.f., Borsboom, Mellenbergh, & van Heerden, 2003). Let us look at an example: an affluent person gives money to a homeless person they pass on the street; clearly, this is a marker of empathy driven generosity. Alternatively, our focal person might consider the homeless an irritant, but nevertheless give money to the homeless person in order to demonstrate empathetic generosity to their newly acquired romantic partner, who is walking beside them,. After all, the monetary cost is trivial but the reputational reward could be substantial. Both

explanations for this behaviour are plausible, but which is right? Is the giving driven by the construct of empathy or the construct of Machiavellianism? In the realm of psychometrics, this problem would manifest in the query: what does the questionnaire item "I give money to the homeless" really measure (if anything at all)? It is important that we can answer this question before we can make decisions based on our measure (questionnaire item) to guide theoretical developments or practical decisions (Cizek, 2016; Kane, 2016; Sireci, 2016).

A simple example such as this highlights the real difficulty associated with psychological measurement through behavioural observation, namely, it rests on assumptions concerning things and processes we cannot see. Our measures provide us with some information or data, but it is not the information or data that we really want (Zumbo, 2007). In the example above, the data we want pertains to the degree to which our focal person is empathetic or Machiavellian, the data we have is a count of how often that person gives to the homeless. One can see immediately a genuine gap between the knowledge that we want and the knowledge that we have. Validity refers to the evaluation of the relationship between the knowledge we want (the nature of the construct) and the knowledge we have (the measured behaviour) and the judgements regarding whether or not this relationship justifies the use of a measure for decision-making.

**The evolving notion of validity**

Papers as early as 1884 readily use the word validity within article titles and abstracts, suggesting that the word was in common use. However, none of these early papers defined validity. The context of the use suggested that validity referred either to the accuracy of a measure (i.e., does the measure, measure the construct of interest) or the appropriateness of a measure (i.e., is the measure useful) for some form of decision-making (e.g., Kiernan, 1884; Germann, 1895). The absence of a clear definition within these early articles perhaps explains why Buckingham's 1921 article, which used the word validity to describe the property that a test measures what it purports to measure, is the most the widely cited origin of the concept of validity.

Since 1921, the concept of validity has been a contentious subject of debate and has evolved a great deal. Numerous highly regarded scholars have proposed a number of revisions and expansions of the validity domain (for detailed historical developments see Kane, 2001; Newton & Shaw, 2013, 2014). From 1921-1950 validity was predominantly assessed by matching the content of psychometric measures with theory and by examining predictive capabilities (e.g., Buckingham, 1921; Cureton, 1950; Guilford, 1946; Kelley, 1927). From the mid-1950s to the mid-1980s, and notably following Cronbach and Meehl's (1955) seminal work outlining 'construct validity', validity became a much broader concept concerned with the examination of a constructs' nomological net. Construct validity dominated (eventually subsuming the notions of content and prediction as branches within the nomological net) until Messick's (1989) proposal of the unified model of validity, which retained all of construct validity but added the examination of the consequences of psychometric use. The extent of these revisions and expansions are such that arguably the most authoritative source on psychometric testing today (Hubley & Zumbo, 2011; Kane, 2013, 2016; Newton, 2012; Newton & Shaw, 2013, 2014; Sireci, 2016; Zumbo 2007, 2009),

the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999, 2014), defines validity very differently from Buckingham (1921). The *Standards* draws heavily on Messick's (1989) unified model and states that,

> "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself." (AERA, APA, NCME, 2014, p. 11).

There is one rather notable departure from the early definitions of validity. In this definition, validity is not a question of whether a psychometric measures what it purports to, but rather is concerned with the interpretations of psychometric scores and the decisions that one makes based on those interpretations. Validity is not a property of a measure but of an interpretation and thus one validates not a psychometric but an interpretation of a psychometric score (Kane, 2006). The "consensus position" (Newton & Shaw, 2013, p. 13) espoused within the *Standards* (2014) makes several additional points regarding the nature of validity, namely:

- Validity is not an all or nothing property, it is a matter of degree; the use of two tests can be somewhat valid in both cases but one can be more valid than the other.
- Validity does not exist as distinct types; rather, validity is the single overall judgement of the adequacy of a test interpretation or use. Thus, one should not speak of X validity or Y validity. Validity is a single evaluative judgement.
- Whilst validity is a single entity, there are five forms of validity evidence: content, response processes (e.g., cognitive processes during item responding),

relations with other variables (e.g., convergent, discriminant, concurrent, and predictive validity evidence), internal structure (e.g., factor structure), and evidence based on consequences (i.e., whether test use is fair and unbiased).

- Validation is an on-going process in which various sources of validity evidence are accumulated to build an argument in favour of the validity of the intended interpretation and use of the test.

The view of validity put forward by the *Standards* (AERA, APA, NCME, 1999, 2014) encompasses much of the previous 100 years of validity debate. Central to this notion of validity are the five different sources of evidence required to demonstrate validity.

The first element of evidence relates to the content of a psychometric measure and suggests that psychometric measures should contain exclusively content relevant to the construct at hand and that the content should be representative of the whole construct domain. This is very similar to original validity arguments made by Buckingham (1921) who, working within the domain of cognitive ability, argued that measures of intelligence should cover general learning ability across domains rather than learning of a narrow topic that is overly sensitive to the recency with which the learning had taken place (Buckingham, 1921). Simply, if a psychometric is designed to measure X, it should contain representative coverage of X.

Evidence relating to response processes is perhaps one of the most sophisticated elements of validity (Borsboom et al., 2004; Cronbach & Meehl, 1955; Embretson, 1984, 1993, 1998, 2016) and is often ignored in practice (Borsboom et al., 2004). Briefly, if a psychometric is designed to measure a construct (e.g., school knowledge) then responding to the items of the psychometric should require the use of the construct (e.g., retrieval of school knowledge). It follows that the examination of these item responses is vital for establishing what is measured.

Evidence regarding relations with other variables represents the vast majority of previous validity debate and encapsulates criterion associations and a large proportion of construct validity (Cronbach & Meehl, 1955). Relations between psychometric measures and criterion variables have always been seen as important with many early validity discussions focusing heavily on predictive capabilities (Buckingham, 1921; Cureton, 1950; 1951; Guilford, 1942, 1946). The simple suggestion is that if, for example, intelligence drives learning then a measure of the rate at which learning occurs or of scholastic achievement should constitute a measure of intelligence (Buckingham, 1921). Thus, a positive correlation between a measure of intelligence and educational attainment would provide validity evidence (Cureton, 1950; 1951; Guilford, 1942, 1946). In addition to criterion relations, Cronbach and Meehl's (1955) conception of construct validity implied that a psychometric measure can, in part, be defined by its relationship with other measures or observables. In other words, we should be able to predict, using theory, how a construct will relate to other things. If our measure of that construct does indeed relate as predicted, then we can be confident that we are measuring the target construct. These predicted relations were termed a system of laws or nomological net. In practice, the nomological net is often assessed via convergent validity evidence (i.e., correlations between multiple measures of the same construct are high) and discriminant validity evidence (i.e., that there are small correlations between measures designed to assess different constructs; Campbell & Fiske, 1959). However, Cronbach and Meehl (1955) presented four additional methods to examine construct validity: the previously discussed response processes, known group differences (e.g., a depression measure should differentiate between those who have been diagnosed with depression and those who have not), item correlations and internal structure (e.g., factor structure), and changes over time (e.g., development or stability).

Evidence relating to the internal structure of a measure concerns the relationship between items and represents a separate branch of validity evidence according to the *Standards*. If ten items are expected to measure the same construct then they should be correlated and load onto a single factor (e.g., Spearman, 1904). If the items are hypothesised to measure four sub-components of one higher-order factor then all items should be correlated, but there should be four distinguishable factors and modelling a higher-order factor should be possible.

Evidence concerning consequences regards the suggestion that we must consider the intended and unintended effects of test use when deciding whether or how a psychometric test should be used (e.g., Cronbach 1988; Hubley & Zumbo, 2011; Messick, 1989; Zumbo 2007, 2009). In essence, consequences consider fairness and represent the major contribution of Messick (1989) to validity theory. The idea that the social consequences of psychometric use constitutes a branch of validity is quite a step from the scientific (or statistical) forms of validity evidence previously endorsed; indeed, it is inherently political. Even those who introduced consequences into the validity domain noted that they do not sit neatly within a unified validity model (e.g., Cronbach, 1988, p. 6). Ultimately, however, validity theorists concluded that given the status of validity as the key concern in psychometric development and the importance of consequences that the two should remain unified (Messick, 1989; Newton & Shaw, 2016; Shepard, 2016; Zumbo & Hubley, 2016).

Nevertheless, shoehorning consequences under the label of validity remains controversial today. As the editors of a recent special issue focussed on validity noted, “the controversy over consequences still looms largest” (Newton & Baird, 2016, p. 174). The most common line of argument put forward by critics of the Messick-inspired validity model is to depart from unification and split validity into: construct validity and consequences (Lees-Haley, 1996), or construct representation and nomothetic span (Embretson, 1983), or

internal and external validity (Lissitz & Samuelson, 2007), or validity and intended use (Cizek, 2016), or validity and quality (Borsboom et al., 2004). However, these suggestions are often rebuffed with authors arguing: that little is gained by a theoretical or conceptual split, that the validity of a test (or test score interpretation) is inextricably linked to its use and thus consequences, that a split would diminish the focus on consequences, or that a split would confuse researchers and practitioners (e.g., Kane, 2016; Moss, 1998, 2016; Shepard, 2016; Sireci, 2007, 2016; Zumbo & Hubley, 2016).

**Evolving notion of validity: Consensus, what consensus?**

So, the current "consensus position" (Newton & Shaw, 2013, p. 13) regarding validity, endorsed by major educational and psychological bodies (AERA, APA, & NCME, 1999, 2014) and numerous validity scholars (Hubley & Zumbo, 2011; Kane, 2016; Newton, 2012; Newton & Shaw, 2013; Shepard, 2016; Sireci, 2016; Zumbo & Hubley, 2016) states that validity pertains to interpretations not tests, is a single evaluative judgement based on five sources of evidence, is a matter of degree, and is a continuous process. If this view represents a genuine consensus then one would expect it to be reflected in psychometric practice, especially given that the 1999 and 2014 *Standards* are largely consistent regarding their validity guidelines, giving psychometricians more than fifteen years to familiarise themselves with and adopt the recommendations. Thus, it is surprising to note that two reviews conducted by Cizek and colleagues (2008, 2010) revealed that surprisingly few psychometric investigations follow these recommendations. Cizek, Rosenberg, and Koons (2008) found that just 2.5% of articles reviewed take a unitary perspective (i.e., that validity is a single evaluative judgement), that only 9.5% cite the *Standards* or Messick (1989), and that numerous authors refer to validity as a characteristic of a psychometric tool, with only

24.7% referring to validity as a property of score interpretations. Further, Cizek et al. (2008, 2010) noted that hardly any papers examined response processes and consequences.

These observations suggest (at least) two possibilities; either researchers are ignorant of the *Standards'* recommendations, or there may not be a consensus after all. A closer inspection of the validity literature suggests the latter (e.g., Borsboom et al., 2004, 2007, 2009; Cizek, 2012, 2016; Lees-Haley, 1996; Lissitz & Samuelson, 2007; Mehrens, 1997; Popham, 1997; Scriven, 2002; Shadish, Cook, & Campbell, 2002; Wiley, 1991). Discontent with the consensus position covers two fundamental issues. First, to what does validity refer? Researchers do not agree on whether psychometrics, interpretations of scores, or uses of psychometrics should be validated. Second, which pieces of evidence should be considered validity evidence? Over time, the *Standards* definition of validity has come to encapsulate every important test-related issue within a unified validity model, and as a result, validity has become a complex and cumbersome concept. Some would argue that as validity now covers everything it means nothing, and no longer provides a useful framework to highlight the critical features of high-quality measurement (Borsboom et al., 2004; Mehrens, 1997; Wiley, 1991). There is no dispute regarding the importance of the different types of validity evidence, but many have questioned whether all these types of evidence should be considered under the 'validity' banner (c.f., Borsboom et al., 2004; Cizek, 2012, 2016; Lees-Haley, 1996; Lissitz & Samuelson, 2007; Popham, 1997; Scriven, 2002; Shadish et al., 2002).

The turn of the 21st century saw Borsboom and colleagues (2004) make a compelling critique of the 'consensus position' on both counts and in doing so, argue for a return to the 1920's version of validity:

> Validity is not complex, faceted, or dependent on nomological networks and social consequences of testing. It is a very basic concept and was correctly formulated, for instance, by Kelley (1927, p. 14) when he stated that a test is valid if it measures what

> it purports to measure … a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure (Borsboom et al., 2004, p. 1061).

Borsboom et al. (2004) put forward a very simple view of validity. Validity is not about *interpretations* or *intended uses* but is a *property of tests*, and a test is valid if the construct you are trying to measure exists (in a realist sense) and causes the measured behaviour/response. In other words, completing an intelligence test should require the use of intelligence, and thus differences in test responses between persons would be the result of differences in intelligence. The simplicity of Borsboom et al.'s view is a virtue and provides clear guidance for establishing validity through a single question: does the construct of interest cause the observed/measured behaviour/response? If so, the test is valid.

From a validity evidence perspective, Borsboom et al.'s (2004) view sees validity limited to evidencing the response process and in making this point they deliver hefty critiques regarding the role of the other elements of the nomological network. Put simply, they argue that whilst a nomological network is useful for theory building, knowing that your measure correlates with other measures does not help identify what your measure actually measures: "it is farfetched to presume that such a network implicitly defines the attributes in question…. It is even more contrived to presume that the validity of a measurement procedure derives, in any sense, from the relation between the measured attribute and other attributes" (Borsboom et al., 2004, p. 1064). This critique is undoubtedly strong[1]. Observing a positive correlation between a measure labelled intelligence and educational achievement

[1] Whilst I agree with the notion that convergent and discriminant validity cannot tell us whether our measure accurately captures the target construct, it is worth noting that Borsboom et al. (2004) critique a caricature of the nomological net that is probably the predominant interpretation in current practice. However, convergent and discriminant evidence are only a small proportion of the nomological net as proposed by Cronbach & Meehl (1955). If one were to establish evidence based on all of Cronbach and Meehl's (1955) sources of evidence (which include response processes), one would have a strong evidence base to assess whether the target construct drives variation in the measured behaviour.

actually tells you nothing about what is actually measured. What makes a measure of intelligence so, is that responding to the test requires the use of intelligence, what makes a test of high school history knowledge so, is that answering the questions requires the use of knowledge gained during high school history class. This point seems indisputable.

In keeping with the impassioned and often antagonistic nature of the validity debate (Newton & Shaw, 2016), numerous authors have, in turn, critiqued Borsboom's view and similar views espoused by others (e.g., Cizek, 2012, 2016; Lissitz & Samuelson, 2007; Scriven, 2002; Shadish et al., 2002). One counter argument states that validity cannot be a measurement-based property of tests as measurement is conditional (i.e., a measure performs differently with different groups; Newton, 2012). This is true; any given measure is unlikely to operate equivalently across all populations or contexts. However, this is an empirical question ripe for empirical examination and not a point that warrants the abandonment of the notion that a psychometric test can be validated. All that is needed to negate this concern is a simple specification of the population or context of interest (as is the norm, e.g., child, or adult intelligence tests).

A second counter argument asserts that a test cannot be validated independently of an interpretation of a test score (e.g., Hubley & Zumbo, 2011; Kane, 2013, 2016; Newton, 2012; Sireci, 2007; 2016; Zumbo, 2007, 2009). For example, Kane states that it would be difficult to validate a test without a label, and a label requires an interpretation. This seems an odd concern. If a psychometric measure is designed to measure construct X, then the label would be 'measure of X'. This seems uncontroversial and if the interpretation is limited to "the psychometric measures the target construct" then a critique based on interpretations is weak. It is true that any scientific endeavour cannot be separated from values and the labelling process can be particularly value-laden (divergent thinking vs. scatter-brained, conscientious vs. boring, emotionally stable vs. cold). Nevertheless, one can still evaluate whether the

psychometric measures what it purports to measure. Simply, if you can show that construct X causes variation in responses to 'measure of X' then this is concrete evidence that the measure of X, measures X within the given population (Borsboom et al., 2004).

Third, validity theorists have argued that a psychometric cannot be validated unless one considers the purpose for which it will be used (Hubley & Zumbo, 2011; Kane, 2016; Sireci, 2007, 2016). For example, Sireci (2007) cannot imagine convening a committee of experts to evaluate a test without telling them the purpose for which the test is to be used. This argument is premised on the mistaken assumption that measurement informed decisions but not measurement per se represent a purpose. However, if the purpose of the test is to measure a construct (as Borsboom et al. suggest), then all the experts need to know is the theory surrounding the nature of that construct. If the construct to be assessed is sixth-grade mathematical knowledge (Sireci's example), then it is quite clear that the test should contain questions representative of knowledge delivered during sixth-grade mathematics education. What the test scores will subsequently be used for (e.g., training, awarding high school diplomas, college entry) is a separate point to that of whether or not the test accurately captures mastery of sixth-grade mathematics knowledge.

Fourth, Kane (2013, 2016) has argued that Borsboom and colleague's view is reductionist and would see many important elements of psychometric evaluation currently considered as indicators of validity evidence (e.g., predictive validity evidence) omitted from the definition of validity. Borsboom et al. (2004) counter this point within their original thesis by stating explicitly that they believe validity is already too large, and that as the unified model encompasses every test-related concern, it is essentially meaningless. Those who oppose segmenting the different forms of validity evidence are often most concerned that such a split would see psychometric developers ignore social consequences (e.g., Kane, 2016; Messick, 1989; Sireci, 2007, 2016). This is a noble concern, but one that is neither

scientifically relevant nor empirically supported. First, we should not define the concept of validity based on political concerns. If including evidence regarding psychometric measurement and evidence regarding the use of a psychometric for a given purpose represent separate concerns then we should not artificially combine them. Second, the *Standards* state explicitly that the responsibility for evaluating the social consequences of the primary intended use of a psychometric is to be evidenced by the developer, but that subsequent uses require the user to evaluate the consequences (AERA, APA, & NCME, 1999, 2014). Third, as evidenced by the two reviews conducted by Cizek and colleagues (2008, 2010), social consequences are already largely ignored. Current evidence would suggest that sacrificing a clear and logically coherent definition of validity in order to protect consideration of consequences does not work.

Sadly, as an observer of this debate, I concur with Newton and Shaw (2016), who have suggested that the style of academic discourse within the validity field promotes the taking of sides that generates a false dichotomy that is not conducive to progress (Newton & Shaw, 2016). Given that validity is "the most fundamental consideration" in the development and use of psychometric measures (AERA, APA, NCME, 2014, p. 11) that gives rise to numerous practical implications (i.e., guidance for test developers), it would be beneficial if we could generate a definition that was useful. How we define the term 'validity' is not simply an academic or semantic debate but one that sets the standards for what evidence is required before proclaiming our measure, our measure interpretation, or measure use is *valid*.

In summary, there are two notable areas of disagreement. First, to what should validity apply? Some argue that validity is a property of a test (Borsboom et al., 2004), some that validity refers to interpretations of test meaning (e.g., Cizek, 2012, 2016), some that validity refers to test interpretation and test use arguments (Kane, 2006, 2016), and others that validity refers to test use as without a use a test is meaningless (e.g., Sireci, 2016).

Second, which pieces of evidence constitute validity evidence? Borsboom et al. (2004) say validity lives and dies by evidence that test completion elicits use of the target variable whilst others see validity as established by demonstrating good psychometric properties (e.g., Lissitz & Samuelson, 2007). Cizek (2016) believes all elements of test evidence except consequences should be considered validity whilst others believe all test evidence including consequences should be considered validity evidence (e.g., Zumbo, 2007). Yet still, others suggest that validity is situation specific and that demonstrating validity in one instance to one audience is not synonymous with demonstrating validity in another instance to another audience (Kane, 2013, 2016). Thus, the current "consensus position" is actually anything but a consensus (c.f., Hood, 2009; Markus, 2016; Newton & Shaw, 2013, 2016). The differences in definition and content are not trivial; they represent radically different approaches to validity. The result is a validity concept (AERA, APA, NCME, 2014) that is amongst the most contested concepts I have come across (c.f. Newton & Shaw, 2016). Not only this, but the *Standards'* definition is so complex and imprecise that it promotes numerous misunderstandings (Newton, 2012; Newton, & Shaw, 2013) that lead to diverse and often sub-standard validation practices (Cizek, 2008, 2010), and acts as a barrier to successful communication in science and practice (Cizek, 2016; Koretz, 2016). It is possible that these problems stem from the inability of psychometricians and those who use our tools to fully appreciate validity theory, it is also possible that the unified model is simply incoherent and untenable (Borsboom, et al., 2004; Cizek, 2016; Popham, 1997)

**Evolving notion of validity: a way forward?**

So, what are we to do in defining the construct of validity, and importantly for this chapter, making useful recommendations for evaluating psychometric validity? Newton and

Shaw (2016), having grappled with these issues for longer than most, suggest three possible options for progress.

First, continue the academic debate until we arrive at a logical conclusion regarding the validity construct. Newton and Shaw suggest that whilst this is possible, it seems highly unlikely. I agree. As we have discussed, theorists are at loggerheads and base their arguments on differing assumptions. A review of twenty papers published – during the writing of this chapter – in a recent special issue concerning validity in, *Assessment in Education: Principles, Policy & Practice,* further demonstrates the distance between theorists, and the vigour with which they defend their own corner (e.g., Borsboom & Wijsen, 2016; Cizek, 2016; Kane, 2016; Shepard, 2016; Sireci, 2016; Zumbo & Hubley, 2016).

Second, Newton and Shaw suggest we could mould validity into a family resemblance concept so that it has a non-technical meaning that approximates 'good'. Using this approach would require only a limited change to the research lexicon and allow the term validity (read good) to be applied to any element of testing: a valid item, a valid selection programme, convergent validity, predictive validity, which is how it is currently used within academia and industry (Cizek et al., 2008, 2010; Newton & Shaw, 2013). This option might change the debate regarding 'validity', but it does not address the two core controversies: whether a test can be valid and which bits of testing belong in a single model.

Third, Newton and Shaw suggest that we can retire the word validity without any undue consequences: "given the omnipresent reality of imprecise and ambiguous usage, and the fact that even testing specialists use the word in quite different ways, it is hard to see how anything conceptually fundamental could be lost if the word were to be retired" (p. 190). On this particular claim, I agree. Validity is used in so many different ways that removing it would not pose a serious problem. We will explore this controversial claim below. However, where I disagree with Newton and Shaw is the notion that removing the term validity will

move the current debate forward. A change in lexicon is needed, but so too is a theoretically driven change in conceptualisation.

Here, I would like to propose another solution drawing upon the second and third of Newton and Shaw's suggestions and the work of Cizek (2012, 2016) and Borsboom et al. (2004). I suggest that we can drop the term validity altogether. However, if this seems too radical a proposition, then we can continue to use the word 'validity' as a family resemblance construct. In doing so, we strip the term validity from having any precise technical definition and thus need other words that can fill this void. We will come to this shortly. However, before we can address the semantic debate, we must focus on the core scientific value of the 'validity' debate: what types of evidence should count towards 'validity'. One extreme is that 'validity' should be evidenced by demonstrating that the measure accurately captures its intended target (e.g., Borsboom et al., 2004, 2009). The other extreme is that validity should be comprised of all information regarding test development and test use (e.g., Messick, 1989; Zumbo & Hubley, 2016). Thus, the major problem we must solve is whether it is logical to use a single model and label to refer to all elements of psychometric measure evaluation. My view, simply, is *no*, it is not logical. To make this argument, and provide a practically useful and theoretically meaningful split in the 'validity' domain, I return to the two questions posed at the outset of this chapter:

1. Am I measuring what I want to measure?
2. Is my measure useful?

These two questions guide measure development, refinement, and use. Establishing whether a measure accurately captures its intended target is a very different beast to establishing whether it is appropriate to use a measure in any given scenario. Let us look at an example using IQ tests:

1. Does my IQ test measure intelligence?

2. Is my IQ test useful for employee selection/guiding the design of educational programmes?

Hopefully, it is clear to all readers that the information and evidence necessary to address question one is different from that needed to answer question two. In Borsboom et al.'s (2004) terms, question one requires evidence that completing the test requires the use of one's intelligence and that the test captures that variation accurately. The second question(s) requires a great deal more evidence and we must consider whether the test predicts job or school performance, whether the prediction offered holds across groups, what cut-off scores should represent 'success', whether it is ethical to segregate children based on intelligence, and so on. As Cizek (2016, p. 215) puts it, "a diverse array of empirical and logical rationales confirms that [test score interpretation and test score use] require distinct sources of evidence bearing on differing purposes and that a single synthesis of evidence on both inference and use is not possible." I could not agree more. Establishing that a psychometric accurately measures intelligence has only a minor bearing on whether it is deemed appropriate as a selection tool. Equally, evidence that a test predicts job performance might indicate its appropriateness for use in selection but it does not give any indication of whether that test measures intelligence. The evidence needed to answer question one differs markedly from the evidence required to address question two and no matter how strong the evidence in relation to either issue, it is not enough to answer the other.

I am not the first author to point out that validity evidence would be better suited to a two-component model. However, my approach has two major differences. First, the divide is driven by the pragmatic need to address the two fundamental questions of test evaluation (as opposed to an attempt to remove consequences from validity) and the theoretical argument that these two endeavours cannot meaningfully be combined. Second, I am not precious about the word validity. In contrast, previous discussions along similar lines have made

impassioned pleas that validity pertains to question one calling evidence relating to the other question utility, consequences, quality, or some other label for 'good' or 'useful' (Borsboom et al., 2004; Cizek, 2016; Lees-Haley, 1996; Lissitz & Samuelson, 2007). Rather than take this antagonistic approach – perhaps inspired by the lexical tradition held dearly by my native world of personality research – I turned to the dictionary to identify two words that match the two fundamental questions, which I believe can serve us well and move us forward. Here are some definitions taken directly from the Oxford English Dictionary Online:

- *Accuracy* (n.d.): The closeness of a measurement, calculation, or specification to the correct value. Contrasted with *precision* (the degree of refinement of the measurement, etc.)
- *Appropriate* (n.d.): Suitable or proper in the circumstances
- *Validity* (n.d.): Well-founded on fact, or established on sound principles, and thoroughly applicable to the case or circumstances; soundness and strength (of argument, proof, authority, etc.)

Question one, refers to whether your psychometric measures what it purports to - or in other words, whether your measure is *accurate*[2]. It is important to note that in the dictionary (and scientific discourse more broadly) accuracy is contrasted with precision which is essentially reliability (see Revelle & Condon, Chapter 23). Note also that the concept of accuracy as defined here aligns to the use of the term validity when it was first embraced in psychological science. Perhaps Buckingham (1921) simply chose the wrong word. Question two, refers to how useful a measure is - or in other words, how *appropriate* it is to use your psychometric for a given purpose in a given situation.

[2] The terminology of measurement within psychology is hotly contested and the phrase 'correct value' is likely to annoy some (Markus & Borsboom, 2013). However, taking a realist approach (Borsboom et al., 2004) that abilities, traits, attitudes, and knowledge exist and drive item responses then standard psychometric techniques could justifiably be labeled as measures seeking to arrive at the 'correct value' (Markus & Borsboom, 2012). Afterall, this is the fundamental principle of psychometrics (psycho = mental, metric = measurement).

Using the terminology suggested here we can say that the goal of psychometric development is to generate a psychometric that *accurately* measures the intended construct, as *precisely* as possible, and that uses of the psychometric are *appropriate* for the given purpose, population, and context. A psychometric should be accurate, precise, and appropriate or in current terminology, a psychometric (or psychometric score interpretation) should be valid, reliable, and valid. The ambiguity of 'validity' is inherent in its inability to differentiate between these two markedly different aims. The 'validity' label does not serve us well and we should change it.

**Two questions, two steps**

Establishing the accuracy and appropriateness of psychometric measures are both equally noble pursuits; neither one is necessarily more interesting or worthwhile. They do however, concern different elements within the life span of a psychometric. Accuracy is likely to be most relevant during initial conceptualisation and development whereas appropriateness concerns the later use of a psychometric for theory building or decision-making. As implied by the previous sentence, I see accuracy and appropriateness forming a two-step approach to psychometric evaluation.

The first-step is to ascertain whether your psychometric accurately measures what it purports to. If we cannot say with confidence what our psychometric is measuring then I would suggest that use of the psychometric for theory building or decision-making is inappropriate. Thus, establishing a measure's accuracy is a fundamental base for all other psychometric activities, and is a necessary but not sufficient condition for establishing the appropriateness of a particular psychometric use (e.g., Cizek, 2016, Kane, 2013, 2016; Sireci, 2016). Some might disagree and suggest that if a psychometric predicts an outcome then its use is appropriate. I disagree. The atheoretical use of a psychometric which measures an

unknown source is unwise, unethical, and possibly illegal (i.e., in the case of employee selection). Once we have some degree of confidence that our psychometric accurately measures what it purports to, then we can assess the appropriateness of its use for a particular purpose, within a particular population, within a particular context. So how do we establish psychometric accuracy and appropriateness? The answers to these questions draw heavily on the hundred years of discussion of validity summarised earlier in the chapter. There has been a great deal of work regarding the assessment of different elements of psychometric measures and much of it is useful here.

**Establishing the accuracy and appropriateness of psychometric measures**

'Validity evidence' comes in many forms. Indeed, Newton and Shaw (2013) listed 151 types of 'validity evidence' (or 'types of validity' as they are commonly referred to) that they identified within the literature. Given the definitional maze that is 'validity', it is not surprising that researchers have invented ever more types of 'validity' evidence. After all, if 'validity' encapsulates all things that can be 'good' about a psychometric (or interpretations and uses) then why not demonstrate its goodness in a myriad ways, especially as judgements of psychometric quality tend to be related to the number of pieces of 'validity evidence' present (Cizek et al., 2008). As a result, navigating the various forms of 'validity evidence' can be overwhelming. Having reviewed a host of sources that discuss 'validity', I found a surprising dearth of succinct and useful 'validity evidence' checklists. The one exception is perhaps the *Standards* (1999, 2014) but even this tends to shy away from providing precise definitions and practical guidance. This, I believe, is unhelpful for those wishing to design and evaluate psychometrics. Thus, based on two separate reviews, I have constructed a succinct and user-friendly checklist of types of evidence that can be used to establish the accuracy and appropriateness of psychometric measures.

The first review conducted for this chapter examined introductory textbooks, 'validity' chapters, and whole 'validity' books and was focussed on understanding how types of 'validity evidence' are generally defined or mistakenly defined (AERA, APA, NCME, 1974, 1985, 1999, 2014; Carmines & Zeller, 1979; Carver & Scheier, 2003; Cooper, 2002; Hubley & Zumbo, 2013; Kline, 2000; Larsen, Buss, & Wismeijer, 2013; Markus & Borsboom, 2013; Osterlind, 2010; Smith, 2011; Woods & West, 2010). The second review examined all newly developed scales published in two highly regarded psychometrics journals, *Psychological Assessment* and *Assessment*, between April 2015 and June 2016. I use the data gathered from these two reviews to provide a list of the major forms of 'validity

evidence' with clear and distinct definitions. The second review is useful in presenting a snapshot of current 'validation' practices. The user-friendly checklists are displayed in Tables 24.1 and 24.2, and contain definitions of evidence types, methods that are particularly useful for generating each type of evidence, and also the percentage of papers introducing new psychometrics that report each type of evidence.

**Accuracy**

Accuracy is defined as the closeness of a measurement to the correct value and is contrasted with precision or reliability (the degree of refinement of the measurement). In psychometric terms, a measure is accurate when it measures what it purports to measure or perhaps more specifically, when variations in the target construct are clearly represented by variations in item responses (c.f., Borsboom et al., 2004). Previous models that have separated 'validity' into two components have tended to include the majority of types of 'validity evidence' under question one (e.g., Cizek, 2016), including criterion associations and correlations with other variables that are central to the classic construct validity model (Cronbach & Meehl, 1955). However, I agree with Borsboom et al. (2004, 2009), who say that whilst correlations between psychometric measures are interesting and can be informative, they do not tell us anything about what is actually measured. Thus, in establishing psychometric accuracy, I will focus only on evidence that relates directly to determining the nature of the measurement, namely, response processes, content representativeness, and content structure. Table 24.1 provides an overview of the most useful forms of evidence for establishing psychometric accuracy.

- - INSERT TABLE 24.1 - -

Table 24.1

Evidence for establishing psychometric accuracy: labels, definitions, methods, and prevalence of current use

| Type of evidence | Definition | Example methods | % used |
|---|---|---|---|
| Content | The degree to which the content (i.e. items, tasks) of a psychometric measure comprehensively captures the target construct | Matching content to standardised descriptions (i.e., curriculum lists or diagnostic criteria) | 55 |
| | | Matching content to theory using expert ratings | |
| Response processes | The mechanism by which the target construct causes item responses | Cognitive mapping | 0 |
| | | Think aloud protocols | |
| Structural | The degree to which the relationships among psychometric content (items, tasks) reflect the theoretical framework | Exploratory factor analysis | 90 |
| | | Confirmatory factor analysis | |
| Stability across groups | The degree to which the content, structure, and response processes remains stable across groups | Differential Item Functioning | 35 |
| | | Invariance measurement | |

**Response Processes**

The gold standard for estimating accuracy lies in the investigation of response processes. Psychometrics are designed to measure unobservable constructs (e.g., abilities, attitudes, traits, knowledge), and the core underlying assumption of psychometrics is that variation in the construct is reflected in item responses (Cronbach & Meehl, 1955). Thus, whenever a psychometric measure is to be interpreted as a measure of a construct, we should be most interested in understanding the mechanisms through which the item responses come about (Borsboom et al., 2004, 2009; Embretson, 1983, 1994). Accuracy here then, concerns the match between the psychological processes hypothesised to be under investigation and the processes that respondents actually engage in or draw upon when responding to items.

To examine response processes, one must first identify the processes, decisions, strategies, and knowledge stores involved in generating item responses. Recall the 'item' proposed at the beginning of this chapter, "I give money to the homeless". If this item is designed to measure empathy-driven generosity, then one might hypothesise that the respondent would engage in a number of processes (see Karabenick et al., 2007, figure 1 for a formalised and generic information-processing model of item responses):

1. Read and interpret the item.
2. Recall from memory seeing a homeless person (perhaps how they felt when considering how the homeless person must feel living outside, vulnerable to the weather and financially unsupported).
3. Have an emotional reaction to this sequence of thoughts.
4. If the emotional reaction is negative (i.e. sadness), decide to give money because it will alleviate some of the homeless persons' pain and perhaps reduce personal feelings of guilt. If the emotional reaction is neutral or positive then it is unlikely that the person will respond to give money.

5. Read and interpret the response options.
6. Select a response option that is congruent with their thoughts and feelings. In this instance, the stronger the emotional reaction, the more extreme the response on the rating scale.

One could examine the processes the respondent actually engages to clearly examine whether the item responses measure empathy-driven generosity. If however, a respondent has a neutral emotional reaction, but notes the social esteem one gets from giving to the homeless and then endorses the item strongly, we can conclude that this item might be accurately measuring empathy for some respondents but not for others (instead measuring Machiavellianism). One could then change the item to "would you give to a homeless person even if no one ever knew about it?" Hopefully, responses to this item would give a more accurate reflection of empathy-driven generosity.

I can imagine some test developers thinking 'this seems like a lot of work' - and they would be correct. Identifying response processes, especially for attitudinal measures is not easy but it is worthwhile work. Many psychometrics are designed to assess complex behaviours and in such cases, the response processes are unlikely to be unidimensional in the purest sense. However, many items have been shown to conform to criteria supporting unidimensionality and provide accurate and precise measurement of only one construct. If however, the construct you wish to measure cannot be measured with unidimensional items then an examination of response processes will uncover this and help in the accurate modelling of item-level data (e.g., Booth & Hughes, 2014). Nevertheless, a focus on response processes allows us to derive firm conclusions regarding the accuracy of our psychometric measures and accurate psychometric measures are essential. Using inaccurate psychometric measures can stifle theory and lead to precarious real-world consequences. The benefits of understanding response processes has the potential to improve measurement accuracy but

also theory development and decision-making. If we understand how a child solves a mathematics problem, we can improve teaching; if we can understand how Warren Buffet makes investment decisions, we can test this and select better investors. Whatever your intended use for a psychometric, establishing that it accurately captures the response process you are interested in is paramount. It is therefore disappointing to see that not one of the twenty papers reviewed for this chapter assessed the processes elicited during item response (see Table 24.1).

There are a number of different methods for evaluating response processes. For example, cognitive design systems are very useful for evaluating the accuracy of decision-making or problem solving items (e.g., Embretson, 1994, 1998, 2016). Cognitive design systems posit a theory of item response processes, build a series of hypotheses regarding these processes, and then test them. In many respects, cognitive models reflect good quality theory testing. In very simple terms, if we posit that compared to item X, item Y requires a greater use of construct A, then there should be a greater number of correct responses to item X than item Y. This kind of modelling allows one to develop expected item response patterns and then examine whether or not these response patterns are met. Embretson (1998, 1999, 2010, 2016), Embretson and Gorin (2001), Jansen and van der Maas (1997), and Mislevy and Verhelst (1990) provide useful empirical examples of how cognitive models of item responses can be evaluated.

Another method of investigating item response processes is to use think-aloud protocols (Ericsson & Simon, 1980, 1993, 1998). Think-aloud protocols require participants to say whatever comes into their mind as they respond to an item (what they are thinking, recalling, feeling etc.). This technique can give researchers an understanding of the cognitive processes undertaken and would be useful for the hypothetical example concerning empathy-driven generosity/Machiavellianism described above. Researchers should train respondents in

the think-aloud technique and provide clear guidance regarding the type of information they are looking for. Think-aloud protocols can either be conducted concurrently (as the participant responds) or retrospectively (upon completion of the response). Both have merits and weaknesses as concurrent protocols allow for more spontaneous and potentially accurate insights but increase cognitive demand on participants, whereas retrospective protocols (sometimes termed cognitive interviews) do not interfere with item responses but, like all retrospective analyses, are subject to forgetting and bias, and cannot detect dynamic real-time changes in response processes. Some practical examples of the think-aloud process can be found in Ericsson and Simon (1998), Darker and French (2009), Durning et al. (2013), Vandevelde, van Keer, Schellings and van Hout-Wolters (2015), with van Someren, Barnard, and Sandberg (1994) providing a comprehensive practical guide to think-aloud protocols.

Although think-aloud protocols and similar cognitive interview techniques are valuable, there remain numerous questions regarding the optimal approach to their implementation (e.g., Presser et al., 2004). For example, there are questions concerning which constructs are suitable for cognitive probing, the optimal number of probes, the merits of concurrent or retrospective probing, whether the think-aloud distorts the true respondent process, and how to amalgamate, analyse, and interpret the results. Nevertheless, despite slow methodological progress in this area (driven largely by test developers reluctance to carry out appropriate studies), understanding response processes is vital and is to be encouraged whenever applicable.

**Content**

Whilst evidence that the target construct drives item responses is undoubtedly the most important element of evidence necessary to establish psychometric accuracy, other forms of evidence are also important. Content evidence (‘content validity’) is well known and widely discussed within the psychometric literature (Nunnally & Bernstein, 1994; Haynes,

Richard, & Kubany, 1995). Evidencing accurate content involves demonstrating a match between the content theorised to be related to the construct and the actual content of the psychometric. So, if psychopathy is theorised to consist of four main content dimensions then items should represent those four dimensions. Within our review, content evidence was one of the most widely reported with 55% of articles explicitly examining the accuracy of the psychometric content of items.

All examinations of content accuracy are inherently theoretical. The first and most important step in assessing content accuracy is to define the domain and the facets of the target construct (Nunnally & Bernstein, 1994). A poorly defined construct, based on a weak theoretical footing will, almost certainly, lack content accuracy. In many fields, there are authoritative sources that provide a 'gold standard' construct definition that can be used to guide content development and evaluation (e.g., the APA's Diagnostic and Statistical Manual of Mental Disorders). However, relying solely on a single widely accepted model is not always suitable and certainly should not be done blindly. Even well-established models deserve thorough examination and it is not uncommon to find dominant construct definitions wanting (see Irwing & Hughes, Chapter 1) . Once one has carefully defined the construct domain, content must be generated that is likely to elicit appropriate response processes. In order to generate such content, one can conduct expert and general population sampling to identify relevant content (Haynes et al., 1995; Irwing & Hughes, Chapter 1). The generated content should then be assessed using formalised rating procedures whereby multiple experts rate the quality of the content based on its relevance, representativeness, specificity, and clarity (see, Nunnally & Bernstein, 1994; Lynn, 1986; Stewart, Lynn, & Mishel, 2005). Content deemed to be content accurate based on theoretical examination can then be examined empirically for evidence that the content does indeed elicit the expected response processes across all facets of the construct.

There are two major threats to content accuracy, namely, construct-irrelevance and construct under-representation (Messick, 1989). If a psychometric is designed to measure psychopathy but contains items pertaining to intelligence, we say the psychometric has construct-irrelevant content, whereas if the psychometric only measures three out of four psychopathy factors then we say that we have construct under-representation.

Content-based evidence provides one of the most compelling demonstrations of the failings of the unified model of 'validity' (AERA, APA, NCME, 2014). There are two common goals when reviewing psychometric content, one is to optimise accuracy (i.e., full content representation) and the other is to optimise predictive properties whilst reducing adverse impact (or negative social consequences). These two aims of psychometric test construction are not convergent. For example, a truly accurate measure of personality would require the measurement of many facets, taking respondents a long time. In contrast, the most appropriate way to use personality for prediction (e.g., employee selection) involves measuring only relevant facets (e.g., Hughes & Batey, in press). If, however, we accept that these two lines of enquiry are distinct we can say clearly that one should first build a fully accurate measure of personality (with many facets, regardless of their predictive capabilities) and then, when using personality for prediction, compile a measure with only the most appropriate facets (Hughes & Batey, in press).

**Structure**

Closely related to content evidence is evidence relating to how that content is structured. Evidence of an accurate structure involves the demonstration of a match between the theorised content structure and the actual content structure. Using the psychopathy example from above, if we hypothesize four factors of psychopathy then we should be able to identify those factors within our item responses, but we should also be able to model the

general, higher-order psychopathy factor. Structural evidence is most commonly amassed using forms of factor analysis, including exploratory and confirmatory factor analysis (see Mulaik, Chapter 8; Cai & Moustaki, Chapter 9; Jennrich, Chapter 10; Timmerman, Lorenzo-Seva, & Cuelemans, Chapter 11) and more recently, exploratory structural equation modelling (Asparouhov & Muthén, 2009; Booth & Hughes, 2014). Each of these techniques is popular within current psychometric practice (Table 24.1) and is discussed in excellent fashion within this Handbook of Psychometric Testing and elsewhere, so I will not discuss them further.

It is also important to establish whether or not the structure you have identified holds across different groups. In general, response process, content, and structure should be examined across all populations of relevance (e.g., across ability levels, ages, national groups). Two tools are of particular value when examining structural stability across groups. First, invariance analysis (Millsap & Kim, Chapter 26) can examine whether the number of factors is stable across groups (configural invariance), whether the factor loadings are of the same magnitude (metric invariance), whether the intercepts are stable (scalar invariance), and whether the unique factor variances are stable (strict invariance, though Little [2013] provides compelling arguments as to why this is not an appropriate criterion for measurement equivalence). Thirty-five percent of articles in our review reported some form of invariance analyses. Second, differential item functioning (Drasgow, Nye, Stark, & Chernyshenko, Chapter 27) identifies items for which members of different subgroups with identical total test scores show differing response patterns and thus can identify potentially biased items.

**Appropriateness**

Whereas there are currently a limited number of methods for establishing the accuracy of a psychometric measure, there are many potential methods for establishing

appropriateness. A psychometric measure can be used for many different purposes across many different contexts and for each a unique set of evidence could be required to demonstrate appropriateness (Kane, 2006, 2013, 2016). Broadly speaking, there are two main classifications of use: theory testing and decision-making. Theory testing captures the majority of research activities (does X correlate with Y, how does X interact with Y to produce Z, etc.), whilst decision-making refers to applied use (e.g., selecting or training employees, diagnosing mental health problems, placing students within ability-based classes). In keeping with currently accepted nomenclature, I will discuss evidence concerning three main categories, namely, evidence based on relationships with other variables, evidence relating to consequences and fairness (e.g., AERA, APA, NCME, 2009, 2014; Cronbach & Meehl, 1955; Messick, 1989), and evidence relating to feasibility concerns (e.g., Cizek, 2016). Table 24.2 provides a summary of the most common forms of evidence that can be used to establish psychometric appropriateness.

-- TABLE 24.2 –

Table 24.2

Evidence for establishing psychometric appropriateness: labels, definitions, methods, and prevalence of current use

| Type of evidence | Definition | Example methods | % used |
|---|---|---|---|
| Convergent | The relationship between psychometric measures of a construct and other measures of the same construct. | Correlations | 75 |
| Discriminant | The relationship between test scores and scores on measures assessing different constructs. | Correlations | 85 |
| | | Confirmatory factor analysis (CFA) | |
| Predictive | The ability to longitudinally predict criterion scores based on test scores. | Time-lagged or longitudinal regression models | 10 |
| | | Time-lagged or longitudinal structural equation models (SEM) | |
| Concurrent | Cross-sectional prediction with both predictor and criterion data collected at the same time. | Regression models | 60 |
| | | SEM | |
| Incremental | Improvements in prediction of a criterion variable added by a particular test over and above other measures. | Regression models | 20 |
| | | SEM | |

| | | | |
|---|---|---|---|
| Known groups | The extent to which a psychometric measure correctly discriminates between those known to be low and those know to be high in a construct. | T-tests<br>ANOVA<br>Latent mean differences (mean structures analysis) | 10 |
| Consequences | The intended and unintended consequences of test use. | Differential item functioning<br>Invariance analysis<br>Estimates of adverse impact<br>False-positive and false-negative rates | 0 |
| Feasibility | The practical concerns related to psychometric use. | Cost<br>Time<br>Respondent reactions | 0 |

**Relationships with other variables**

Perhaps the most common use of psychometrics is as a predictor of a criterion. Criterion association evidence comes in two major forms: concurrent and predictive. Concurrent criterion associations are essentially cross-sectional with both the predictor and criterion measured at roughly the same time. Sixty per cent of articles reviewed reported concurrent criterion associations. In contrast, predictive criterion associations are time-lagged. The predictor is measured at time one and the criterion at some later point in time. Predictive criterion associations are more powerful for all the reasons longitudinal research is superior to cross-sectional (Menard, 2002). Of articles reviewed, only 10% reported predictive criterion relations though many who presented concurrent relations claimed that they were predictive. In addition to simple models with a single predictor and a single criterion, it is also advisable to demonstrate that a measure offers incremental prediction beyond other established predictors (see Smith, Fischer, & Fister, 2003, for an accessible treatment of how to consider incremental prediction during measure development). Such models use multiple predictors and focus on the additional variance explained by the focal measure, thus allowing for more concrete claims regarding the appropriateness of a measure. For example, if we wanted to predict job performance it would be of great value to know that our newly developed measure predicted even when modelled alongside well-known predictors such as intelligence and conscientiousness. In such a situation, we could then say confidently that using our measure during selection is appropriate. As displayed in Table 24.2, criterion relationships are commonly reported within new scale development papers but overwhelmingly these associations are concurrent. Common methods include various regression models, structural equation models, and tests of known group differences (e.g., those diagnosed with a mental disorder and those who are healthy).

Convergent and discriminant evidence were first introduced by Campbell and Fiske (1959). Convergent evidence is "represented in the agreement between two attempts to measure the same trait through maximally different methods" (Campbell & Fiske, 1959, p. 83) though the same traits measured using the same method can be considered a weak form of convergent evidence. As we can see in Table 24.2, convergent evidence is commonly reported within new scale development papers but in the overwhelming majority of cases the construct is measured using the same method (i.e., two self-report measures of Narcissism). The review of new scales also revealed that convergent evidence is often claimed when the methods used are the same (i.e. two self-report questionnaires) or simply when a psychometric correlates with other measures that theoretically it should (i.e., intelligence and school performance), and often Campbell and Fiske (1959) were cited as the justification for these analyses. However, this is not 'convergent validity' as proposed by Campbell and Fiske (1959). In fact, it is not entirely clear what 'type' of evidence this is, if any. Although often described as 'nomological validity' or 'construct validity', evidencing positive correlations between two measures is perhaps best viewed as a test of theory, not a specific form of validity evidence, especially given the breadth of both the nomological net and construct validity (Cronbach & Meehl, 1955).

Discriminant evidence (frequently mislabelled as divergent evidence in the literature) is the degree to which measures of theoretically distinct constructs are empirically unrelated to one another. Discriminant evidence is particularly important for showing that a new measure is actually new and preventing construct proliferation (e.g., Le, Schmidt, Harter, & Lauver, 2010; Reeve & Basalik, 2014; Shaffer, DeGeest, & Li, 2016). Construct proliferation occurs either when researchers propose multiple scales that claim to measure the same underlying construct but actually do not (Jingle Fallacy) or when ostensibly new constructs are proposed that are theoretically and/or empirically indistinguishable from existing

constructs (Jangle Fallacy, Kelley, 1927; Ziegler, Booth, & Bensch, 2013). Construct proliferation impedes the creation of cumulative knowledge and hinders the development of parsimonious theories (Le et al., 2010). Discriminant evidence is designed for this purpose (Bagozzi, Yi, & Phillips, 1991; Campbell & Fiske, 1959) and is critical in establishing whether a new measure is appropriate for theory building and testing (Harter & Schmidt, 2008). Discriminant evidence can be garnered through the use of the multitrait-multimethod approach (Campbell & Fiske; 1959; see Koch, Eid, & Lochner, Chapter 25) and via confirmatory factor analysis (Bagozzi et al., 1991; Fornell & Larcker, 1981; Voorhees, Brady, Calantone, & Ramirez, 2016). Shaffer et al. (2016) recently presented a practical guide to conducting more rigorous investigations of discriminant properties of measures by taking into account measure reliability. They also make the very important point that researchers should select a reasonably broad number of measures of theoretically similar constructs (e.g., happiness and joy rather than happiness and fear) when investigating the discriminant properties of a measure. Failing to do so provides a weak examination of discriminant properties (Shaffer et al., 2016).

**Consequences and fairness**

Consistent controversy arises when the consequences of psychometric use are argued to be a component of 'validity' (c.f., Cizek, 2016; Newton & Shaw, 2016; Sireci, 2016; Zumbo & Hubley, 2016). I cannot imagine the same degree of controversy if we say that examining the consequences of psychometric use is a core component of establishing whether that use is appropriate. By explicating the difference between accuracy and appropriateness, consequences become more obviously relevant within appropriateness and clearly irrelevant to questions or accuracy. When producing a new psychometric or trying to sell one, researchers or test publishers can currently make a convoluted, lengthy, and

convincing ‘validity argument’ whilst ignoring consequences completely. Indeed, the review conducted specifically for this chapter showed that not a single publication considered potential social consequences (see Table 24.2). However, if psychometric publications had to make two specific evidence-based statements regarding accuracy and appropriateness, it would be much more difficult to hide the lack of consideration of consequences. Any appropriateness statement would be clearly incomplete without an assessment of possible biases and their consequences. Ironically, this is contrary to the commonly raised concern that changes to the current ‘validity’ consensus (AERA, APA, NCME, 2014) would see consequences ignored (Kane, 2013, 2016; Sireci, 2007, 2016).

Consequences are undoubtedly important and psychometric users must always examine whether or not their intended use will have adverse effects on any particular group (Cronbach, 1988; Messick, 1989; see Reckase, 1998 for practical recommendations). Group differences can be assessed using techniques such as multi-group confirmatory factor analysis (Koch et al., Chapter 25), invariance analysis (Millsap & Kim, Chapter 26), and differential item functioning ( Drasgow, et al., Chapter 27). Equally, if psychometrics are used to choose people for jobs or educational opportunities, careful examination of success rates are also important. It is important here to restate that not all group differences are due to bias, they might be due to naturally occurring group differences (e.g., height or conscientiousness between males and females). Whether the use of a psychometric that has known group differences, and thus will adversely impact one group, is appropriate is a difficult decision to make and one for which generic guidance is difficult to give.

**Feasibility**

Finally, when deciding whether the use of a psychometric is appropriate for a given purpose within a given context we often have to consider more than predictive capabilities

and biases. We also have to address practical considerations including monetary cost, time cost, respondent reactions, and administration procedures. Organisational researchers and employee selection practitioners have considered these elements for a long time (e.g., Hough, Oswald, & Ployhart, 2001; Hughes & Batey, in press) and in his recent paper, Cizek (2016) noted that educational programme evaluation models could also provide a framework to guide considerations of feasibility (e.g., Patton, 2008; Shadish, Cook, & Leviton, 1991).

## Conclusion

Adopted by psychometricians in the 1920s 'validity' has been the subject of continuous debate and revision but always retained its status as the most fundamental consideration in test development. Unfortunately, the current unified 'validity' model (AERA, APA, NCME, 2014) is so complicated and convoluted that it baffles researchers, practitioners, and even 'validity' theorists (Cizek et al., 2008, 2010; Markus & Borsboom, 2013; Newton & Shaw, 2013, 2016). Much of the theoretical confusion and practical difficulty stems from tensions that arise when trying to shoehorn considerations of a measure's accuracy, reliability, factor structure, predictive capability, test scoring, test administration, social consequences, and more into a single unified model (Borsboom et al., 2004; Cizek, 2012, 2016; Koretz, 2016; Newton & Shaw, 2016). Many of these sources of 'validity evidence' are not just difficult to combine, they are sometimes diametrically opposed and cannot be meaningfully represented within a unified model. This fact has led numerous authors to call for substantial revisions to validity theory (e.g., Borsboom, et al., 2004, 2009; Cizek, 2012, 2016; Koretz, 2016; Newton, 2012; Newton & Shaw, 2013, 2016; Popham, 1997; Lissitz & Samuelson, 2007). I agree with the major premise of these critiques and hope that the introduction of the accuracy and appropriateness model of psychometric evaluation goes some way to providing the theoretical advance required.

The accuracy and appropriateness model has a number of notable advantages over the unified model (AERA, APA, NCME, 2014; Messick, 1989). First, the model answers calls for a coherent partition of the different types of 'validity evidence' according to the two major questions in psychometric evaluation (Cizek, 2016), namely, what do psychometrics measure and are they useful? This reconfiguration addresses the logical inconsistencies associated with the unified model (Borsboom et al., 2004; Cizek, 2016; Popham, 1997) and means that each of the major forms of 'validity evidence' espoused in validity models

throughout history now sits in a sensible place alongside other forms of ‘validity evidence’ that help address the same question.

Second, structuring the two main questions and associated sources of ‘validity evidence’ into a two-step process provides psychometric developers and users with a clear and coherent model to follow in practice. This theoretically coherent and conceptually simple model leaves little question regarding which types of evidence are needed to establish accuracy and appropriateness and thus has the potential to improve psychometric development and use (Cizek, 2016; Koretz, 2016).

Third, the model allows us to drop the emotive word ‘validity’ (Newton & Shaw, 2016) and in doing so move past the adversarial and antagonistic semantic debate that has distracted theorists for decades (Hood, 2009; Markus, 2016; Newton & Shaw, 2016). In turn, we can refocus our attention on the scientific and ethical questions that really form the core of psychometric evaluations. Those who might be sceptical of this, are encouraged to revisit the sections of this chapter which outline methods for establishing accuracy and appropriateness. These sections were written without the use of the word ‘validity’ except when in reference to previous work, demonstrating that test evaluation can be discussed without our prized and poorly defined label.

Fourth, the labels accuracy and appropriateness address calls to provide simple yet precise terminologies that allow us to communicate whether or not a psychometric measures what it claims to (Borsboom, 2012) and whether or not a specific test use is justified (Cizek, 2016; Koretz, 2016). The precision and simplicity of these terms make it easy to communicate within and across research domains, and across the research-practice divide.

Fifth, the importance of consequences can be stated without invoking the theoretical and logical critiques promoted by the inadequacies of the unified model. Numerous authors have correctly emphasised the importance of test consequences (Kane, 2016; Sireci, 2016;

Zumbo & Hubley, 2016) and simultaneously argued that changes to the unified model would see them ignored. However, as we have discussed, consequences are already ignored (see Cizek, et al., 2008, 2010 and this chapter's review). Currently, psychometric developers are able to hide a lack of consideration for consequences by presenting a myriad of other forms of 'validity evidence' (Cizek et al., 2008, 2010). However, if psychometricians were to adhere to the model espoused here, they would need to present two explicit bodies of evidence, one concerning accuracy and a second concerning appropriateness. It is inconceivable that consideration of consequences could be omitted from a discussion of whether or not it is appropriate to use a psychometric for a specific real-world purpose.

This chapter set out to provide psychometric researchers and practitioners with a guide for 'validating' measures. I believe the current chapter has achieved its aims by stating clearly that psychometric 'validity' is concerned with establishing the *accuracy* and *appropriateness* of measures. I hope that the theoretical discussion and practical guidelines provided will serve as a stimulus for further theoretical clarification and prove useful for those developing and using psychometric measures. Perhaps the key practical take-home messages for those who develop, evaluate, and use psychometrics are as follows. The accuracy of a psychometric can be established through examination of: participant response processes, psychometric content, and the structure of psychometric content. Whether it is appropriate to use a psychometric for a given purpose can be established through examination of: the relationship between psychometric scores and other variables, the potential or actual consequences of psychometric use, and the practical feasibility of psychometric use.

References

Accuracy. (n.d.): In Oxford English Dictionary Online. Retrieved from http://www.oed.com

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: American Educational Research Assocation.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (4th ed.). Washington, DC: American Educational Research Assocation.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (5th ed.). Washington, DC: American Educational Research Association.

Appropriate. (n.d.): In Oxford English Dictionary Online. Retrieved from http://www.oed.com

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438.

Bagozzi, R. P., Yi, Y., & Phillips, L. W. (1991). Assessing construct validity in organizational research. *Administrative Science Quarterly, 36*(3), 421-458.

Booth, T., & Hughes, D. J. (2014). Exploratory structural equation modeling of personality data. *Assessment, 21*(3), 260–271.

Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptions of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 38-41.

Borsboom, D., & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton, & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education* (pp. 85-115). Cambridge: Cambridge University Press.

Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: The value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 281-283.

Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203-219.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium--XIV. *Journal of Educational Psychology*, *12*(5), 271-275.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105.

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage.

Carver, C. S., & Scheier, M. F. (2003). *Perspectives on personality* (5th ed.). Boston: Allyn & Bacon.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212-225.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*(5), 732-743.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Cone, J. D., & Foster, S. L. (1991). Training in measurement: Always the bridesmaid. *American Psychologist, 46*(6), 653-654.

Cooper, C. (2002). *Individual differences* (2nd ed.). London: Arnold.

Cronbach, L. J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika, 53*(1), 63-70.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Cureton, E. E. (1950). Validity, reliability and baloney. *Educational and Psychological Measurement, 10,* 94-96.

Darker, C. D., & French, D. P. (2009). What sense do people make of a theory of planned behaviour questionnaire? A think-aloud study. *Journal of Health Psychology, 14*(7), 861-871.

Durning, S. J., Artino Jr, A. R., Beckman, T. J., van der Vleuten, C., Holmboe, E., & Schuwirth, L. (2013). Does the think-aloud protocol reflect thinking? Exploring functional neuroimaging differences with thinking (answering multiple choice questions) versus thinking aloud. *Medical Teacher, 35,* 720–726.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175–186.

Embretson, S. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive Assessment* (pp. 107-135). New York: Springer.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396.

Embretson, S. E. (2010). Cognitive design systems: A structural modeling approach applied to developing a spatial ability test. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 247-273). Washington, DC: American Psychological Association.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64*(4), 407–433.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.

Embretson, S. E. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice, 35*(3), 6–22.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178-186.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis (revised edition).* Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Fincham, R., & Rhodes, P. S. (2005). *Principles of organizational behaviour.* Oxford: Oxford University Press.

Fornell, C., & Larcker, D.R. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error, *Journal of Marketing Research,* 18, 39-50.

Guilford, J. P. (1942). *Fundamental statistics in psychology and education.* New York, NY: McGraw-Hill.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(4), 427-438.

Harter, J. K., & Schmidt, F. L. (2008). Conceptual versus empirical distinctions among constructs: Implications for discriminant validity. *Industrial and Organizational Psychology, 1*(1), 36-39.

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247.

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19*(4), 451-473.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*(1-2), 152–194.

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, et al. (Eds.), *APA handbook of testing and assessment in psychology* (Vol. 1, pp. 3-19). Washington, DC: American Psychological Association Press.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*, 219-230.

Hughes, D. J., & Batey, M. (in press). Using personality questionnaires for selection. In H. Goldstein, E. Pulakos, J. Passmore, & C. Semedo (Eds.), *The Wiley Blackwell Handbook of the Psychology of Recruitment, Selection & Retention.* Chichester: Wiley-Blackwell.

Jansen, B. R., & van der Maas, H. L. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17*(3), 321-357.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198-211.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C. R., et al. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist, 42*, 139-151.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York: World Book.

Kiernan, J. G. (1884). Feigned insanity: An enquiry into the validity of the reasons for recent diagnoses of this kind. *Journal of Nervous and Mental Disease, 11*(2), 177-184.

Kline, P. (2000). *The handbook of psychological testing* (2nd ed.). London, United Kingdom: Routledge.

Koretz, D. (2016). Making the term 'validity' useful. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 290-292.

Larsen, R. J., Buss, D. M., Wismeijer, A., & Song, J. (2013). *Personality psychology.* Maidenhead: McGraw-Hill Higher Education.

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes, 112*(2), 112-125.

Lees-Haley, P. R. (1996). Alice in Validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51*(9), 981-983.

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-386.

Markus, K. A. (2016). Validity bites: comments and rejoinders. *Assessment in Education: Principles, Policy & Practice, 23*(2), 312-315.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning.* New York, NY: Routledge.

Markus, K. A., & Borsboom, D. (2012). The cat came back: evaluating arguments against psychological measurement. *Theory and Psychology 22*, 452–466.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.

Menard, S. W. (2002). *Longitudinal research (2nd ed.).* Thousand Oaks, CA: Sage Publications.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: The American Council on Education & The National Council on Measurement in Education.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*(2), 236-251.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 1-29.

Newton, P. E., & Baird, J.-A. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice, 23*(2), 173-177.

Newton, P. E., & Shaw, S. D. (2016). Disagreement over the best way to use the word 'validity' and options for reaching consensus. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 178-197.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-319.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment.* London: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Osterlind, S. J. (2010). *Modern measurement: Theory, principles, and applications of mental appraisal (2nd ed.).* Boston, MA: Pearson Education.

Patton, M. Q. (2008). *Utilization-focused evaluation.* Thousand Oaks: SAGE.

Popham, W. J. (1997). Consequential validity: right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Presser, S., Rothgep, J.,Couper, M., Lessler, J., Martin, E., Martin, J., & Singer, E. (Eds.) (2004). *Methods for testing and evaluating survey questionnaires*. New York, NJ: Wiley.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16.

Reeve, C. L., & Basalik, D. (2014). Is health literacy an example of construct proliferation? A conceptual and empirical evaluation of its redundancy with general cognitive ability. *Intelligence, 44*, 93–102.

Scriven, M. (2002). Assessing six assumptions in assessment. In H.I. Braun, D.N. Jackson, & D.E. Wiley (Eds.) *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice.* Newbury Park: Sage.

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the problem of construct proliferation a guide to assessing the discriminant validity of conceptually related constructs. *Organizational Research Methods, 19*(1), 80-110.

Shepard, L. A. (2016). Evaluating test validity: reprise and progress. *Assessment in Education: Principles, Policy & Practice, 23*(2), 268-280.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226-235.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Smith, G. T., Fischer, S., & Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15*(4), 467-477.

Smith, M. (2011). *Fundamentals of management* (2nd ed.). Maidenhead: McGraw-Hill Higher Education.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72-101.

Stewart, J. L., Lynn, M. R., & Mishel, M. H. (2005). Evaluating content validity for children's self-report instruments using children as content experts. *Nursing Research, 54*(6), 414-418.

V. (n.d.): In Oxford English Dictionary Online. Retrieved from http://www.oed.com

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. (1994). *The think aloud method: a practical approach to modelling cognitive processes.* Londen: Academic Press.

Vandevelde, S., van Keer, H., Schellings, G., & van Hout-Wolters, B. (2015). Using think-aloud protocol analysis to gain in-depth insights into upper primary school children's self-regulated learning. *Learning and Individual Differences, 43,* 11–30.

Voorhees, C. M., Brady, M. K., Calantone, R., & Ramirez, E. (2016). Discriminant validity testing in marketing: an analysis, causes for concern, and proposed remedies. *Journal of the Academy of Marketing Science, 44*(1), 119–134.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow, & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach.* Hillsdale, NJ: Erlbaum.

Woods, S. A., & West, M. A. (2010). *The psychology of work and organizations.* Andover: South-Western Cengage Learning.

Ziegler, M., Booth, T., & Bensch, D. (2013). Getting entangled in the nomological net. *European Journal of Psychological Assessment, 29,* 157-161.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity:*

*Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: Information Age Publishing.

Zumbo, B. D., & Hubley, A. M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 299-303.