Dismantling Face Validity - Why the Concept Must Live On, But the Term Must Finally Die

Tzur M. Karelitz

National Institute for Testing and Evaluation, Israel

Charles Secolsky

Mississippi Department of Education

Paper submitted to the IAEA 2015 conference in Lawrence, KS

Introduction

In validity studies, researchers collect evidence to support or refute claims about the interpretation and use (IU) of test scores. In this article we distinguish between two types of evidence researchers could use in achieving this goal. First, evidence can be score based (SBE), resulting from analyzing test scores. A typical example might be correlating scores with other variables to support or refute claims of criterion or construct validity. Consider a college admissions test that is validated by correlating scores with the grade-point averages of enrollees. Another type of evidence is perception based (PBE), resulting from analyzing people's reported perceptions about the test. Perception is an interpretive process influenced for each individual by a variety of factors, such as past experiences, knowledge, beliefs, and attitudes. One form of PBE used to establish content validity might involve measuring the extent of consensus among experts about whether the test content captures the test domain as defined in statements concerning the purpose of the test.

Face validity (FV) is a controversial form of PBE. Face validity is typically defined as whether, on the face of it, a test looks as if it measures what it purports to measure. The consensus among measurement experts is that asking laypeople if they think the test looks valid is not sufficient evidence to support the IU of test scores (Messick, 1989; Secolsky, 1987). In this article we suggest that, in warning against FV, measurement experts obscured its usefulness. Simply put, while the methodological literature on analyzing SBE flourished, investigations of PBE—specifically FV—did not evolve to nearly the same extent.

In our opinion, the definition of FV provided previously is simplistic and outdated. In this article we review how this definition came about and discuss why we believe it is flawed. We believe that if we revisit the ideas that influenced the inception of FV, based on modern

understanding of validity, a broader, more useful concept emerges. PBE represents perceptions about various aspects of the test (e.g., item content, content relevance, score usage), as seen through the eyes of constituencies other than test developers and including the examinees themselves. This type of data does not have the same psychometric value as SBE and should not be interpreted or used as such. Nevertheless, PBE can support the interpretation of SBE by providing insights about why items perform a certain way or why examinees are confused about their scores. Such evidence can be used to improve the overall quality of the test and minimize its negative impacts. Therefore, we contend that the value of PBE first lies in its usefulness when embedded in the process of test development.

PBE is also useful for the validity argument framework (Kane, 2013). Perceptions of various stakeholders represent alternative IU of test scores, which are essential for the evaluation of validity arguments. Moreover, PBE is specifically relevant for evaluating the clarity and plausibility of the IU argument. If the argument is clear and plausible, then the relevant constituents should perceive it as such. Using PBE in this way, along with the relevant SBE, can help support or refute validity claims. To be clear, we do not argue that researchers should prefer the perceptions of laypeople over experts or vice versa. We simply believe that the researchers should consider both views when evaluating their validity argument.

Nearly 70 years ago, Guilford (1946) and others noted that FV was important primarily in making test results palatable to the public. At that time, this did not seem like a particularly important aspect of testing. Today, as social media facilitates the use and misuse of information, public opinion about tests can have a considerable negative influence, not only adversely affecting the test takers but even inducing decision makers to abolish tests. These days, the importance of public opinion is peaking, and we believe that the measurement community has a

scientific need as well as a moral obligation to seriously consider what the public thinks about the tests that we develop. For these reasons there is a pressing need to reconsider the notion of FV and PBE more generally and how they relate to current theory and practice of validity.

In this article we argue for the importance of the ideas that underlie FV in test development and validation and, more pragmatically, their reinstatement to the field within the more general term of PBE. The notion conveyed by FV is that negative perceptions about the test can be harmful. Therefore, studying these perceptions can be useful for improving the test by making it less susceptible to such threats. To be clear, we object to using the term *FV* in any way to describe a test. We believe, however, that PBE is crucial for the test development and the evaluation of validity arguments.

The article has two main sections. In the first section, we review the older literature on measurement. We believe that understanding the early development of validity provides an explanation of why FV was dropped from subsequent developments in measurement theory. The review also highlights that many authors believe there is value in collecting PBE for various validation purposes (although they never refer to this type of evidence collectively in this way). In the second section, we discuss the integration of PBE into the validity–argument framework and provide concluding remarks and recommendations.

Historical Review

In this section we present an overview of the development of validity research, focusing on how FV was treated at various times. The section is divided into five parts: (a) early formulation of FV, (b) classical test theory, (c) distinct types of validity, (d) criterion-referenced measurement, and (e) validity as argument. In this review we emphasize how the critical view of

FV mainly objected to using it as the only type of validity but essentially supported the importance of various forms of PBE for validation purposes.

Early Formulation of FV

From the earliest studies on validity (Vincent, 1924) until the present, determining the extent to which an item or test measures what it purports to measure continues to challenge measurement practitioners and theorists. As Kelly (1927) stated, "The question is thoroughly roused from the slumber of centuries, probably never to sleep again" (pp. 13-14). Consistent with Kelly's observation, the concept of validity evolved during the 1930s until the mid-1940s, and significant contributions were made by Foran (1930), Thurstone (1931), Lindquist and Cook (1933), Turney (1934), and Lindquist (1935, 1936), among others. Foran (1930) differentiated validity, defined as the degree to which a test measures what it was intended to measure, from discriminative capacity, which is the degree to which a test or item discriminates between high- and low-ability examinees. He also made a distinction between validity of content and validity of form. The content of an item, Foran claimed, may not be valid regardless of the form in which it was presented. Conversely, an item may be phrased so that a correct answer would not measure the intended ability. His work formed the basis of what was later referred to as content validity.

Thurstone (1931) defined discriminative capacity as the correlation of a test with a criterion, which laid the groundwork for what later became known as predictive validity. Seemingly in opposition to Thurstone, Lindquist and Cook (1933) discussed the shortcomings of the discrimination index as a measure of item validity and argued for the need for a subjective ingredient in order to improve on item validity. Their work marked the beginning of FV at the item level for the purpose of validating a test. The roots of FV can also be inferred from Lindquist's (1936) assertion that items may often be missed by superior students rather than by

inferior students, because test items are sometimes open to more than one interpretation. That is, test developers or subject matter experts may have one interpretation of what an item measures, and examinees or others may have another. Recognizing this possible dilemma, Turney (1934) stated that statistical analysis should be reserved largely for understanding the apparent consensus among expert judges regarding what they believed test items were measuring. These ideas laid the foundation for Rulon's (1946) notion of "obviously" valid tests. Rulon was unsatisfied with the then-current definition of validity as a test measuring what it purports to measure because a test's validity could be completely altered by changing its purpose arbitrarily.

Even this brief historical review of the literature provides a context of justification for recognizing that differences in the interpretation of what test items measure are not unlikely events. Given that fact, it seems only prudent to collect PBE along with SBE to account for the possibly differing views of what tests or items measure, whether obvious or not so obvious, to borrow from Rulon (1946), and ideally determine the degree to which differing views converge.

While no sharp chronological demarcation can be identified in separating schools of thought with respect to conceptions of validity, differences did emerge among measurement theorists regarding validity and, in particular, FV. Mosier (1947) argued that FV is used by the measurement community to represent various types of validity evidence. To Mosier, some uses of FV are legitimate and some are not; the use of a single term obscures the difference between these two uses and thus can be harmful. For example, tests that are valid by assumption "appear on their face" to have a commonsense relationship to the purpose of the test. Supposedly, tests do not require statistical evidence for this kind of face validity because a lack of validity may be disregarded based on the strength of the assumption alone. Mosier objected to validity by assumption, saying that using it is "totally unscientific and indefensible" (p. 198).

We contend that validity by assumption does, in fact, require statistical evidence because there is the strong possibility that commonsense notions about what an item or a test measures can vary among test takers and other constituent groups of non-experts. These different views are informative because they highlight alternative interpretations of the test's purpose and content. Nevertheless, they are not, by themselves, evidence of validity and do not possess such psychometric quality.

A test has validity by definition if, in the opinion of subject matter experts, the sample of items selected for inclusion on the test represents "adequately the total universe of appropriate questions" (Mosier, 1947, p. 192). In this case, when the criterion is linked directly and intimately to the test items, the use of FV is justifiable. A test has FV via the appearance of validity if, in addition to being valid for pragmatic or statistical reasons, it appears valid in the situation and for the particular purpose it is being used. In this respect, Mosier points out that it is highly desirable that tests be acceptable to users and examinees. A fourth concept, validity by hypothesis, refers to the level of confidence developers have in the appropriateness of the items. In selecting the particular items and tasks, developers form a hypothesis about their appropriateness, which needs to be tested. Mosier (1947) concluded that "Since the term 'face validity' has become overlaid with a high degree of emotional content and since its referents are not only highly ambiguous but lead to widely divergent conclusions, it is recommended that the term be abandoned. Anyone intending to use the term should, instead, describe fully the concept which he originally intended to denote by 'face validity'" (p. 205).

To summarize, Mosier warned against using perceptions as the only source of validity evidence (i.e., validity by assumption). He also argued that some types of PBE could be useful and informative (appearance of validity, validity by hypothesis) and in some situations are

sufficient for supporting the validity of the test (validity by definition). For almost three decades after the publication of Mosier's paper there was hardly any methodological discussion of FV.

Classical Test Theory

The second meaningful period in the development of the concept of validity stems from the enormous influence of statistical methods in psychology. The emphasis given to factor analytic methods by Guilford (1940) is indicative of this movement toward the use of statistical, rather than judgmental, approaches. Guilford recommended only two types of validity for test evaluation: factorial validity and practical validity. Factorial validity is expressed as factor loadings using meaningful common reference factors. Factorial validity not only determined whether a test measures what it is supposed to measure but more precisely answers the question of what the test actually measures. Practical validity is expressed as correlations of the test with meaningful criteria (i.e., criterion-related validity). For Guilford, a "test is valid for anything with which it correlates" (1946, p. 429), and FV was mainly important for making tests more palatable to the public, not for constructing more valid tests.

Guilford believed that only factors derived from factor analysis are dependable enough to be used for validation. On the contrary, we believe that the implicit use of the operational definitions of variables and their statistical analysis might hide what is, in fact, observable. Simply stated, factor analysis is used in interpreting test scores based on the variation in examinee performance. Similarly, we believe that one should analyze the variation in people's perception about tests, items, and scores. The notion of FV as a legitimate area of empirical inquiry stems from the juxtaposition of SB and PB evidence with the goal of gaining insights about the validity of IU of test scores.

The work reviewed thus far traces the ongoing debate between statistical and judgmental conceptions of validity and its implications for FV. The beginning of the postwar developments in measurement theory continued this trend. For example, Goodenough (1949) wrote about what became known as the commonly accepted definition of validity: A test is valid if it measures what it purports to measure. In the next two years, several authors extended this definition in various ways. For example, Gulliksen (1950) operationalized validity as the correlation with some criterion or true score, Cureton (1951) emphasized the need for expert judgment in defining these criteria, and Cronbach (1949) and Tyler (1949) referenced both of these ideas.

Tyler (1949) warned that correlating test scores with indirect criteria be done only if those criteria originate from tests that had FV in their own right. In the first *Educational Measurement* chapter on validity, Cureton (1951) referred to FV in this way: "A test is face-valid if it looks valid, particularly if it looks valid to laymen" (p. 672). He further stated that as a validity concept, FV reflects "inadequate and superficial analysis." These assertions refer to the concept of validity by assumption (Mosier, 1947) but not to other meanings of FV. For example, tests should exhibits high instructional validity (McClung, 1977), an alignment between their content and what was actually taught in the classroom. Relevant evidence to evaluate instructional validity can be based on students' perceptions of this alight net. Still, following Cureton's advice, researchers might wrongly dismiss this evidence as showing only FV and therefore inadequate for validation.

The flurry of intellectual activity around measurement issues, in particular validity, may have spawned the impetus for Adams (1950) to measure FV by asking government workers to judge the extent to which a set of tests had true validity. He found considerable differences between the judgments made by individuals about which tests had this sort of FV. Finally,

Adams concluded that FV exists because examinees consistently agreed that some of the tests appeared more valid and because their judgments correlated with the tests' actual criterion validity, meaning these perceptions of the tests' validity were relatively accurate.

Distinct Types of Validity

Thus far, there were two major approaches to conceptualizing validity: the role of judgment as evidence for content validity, and the criterion against which to validate the test. This conceptual disparity led the American Psychological Association (1954) to recommend as standard the distinction among four types of validity, each requiring different types of evidence and different interpretations: predictive, concurrent, content, and construct. Twelve years later, The *Standards for Educational and Psychological Tests and Manuals* (American Psychological Association [APA], American Educational Research Association [AERA], and National Council on Measurement in Education [NCME], 1966) subsumed predictive and concurrent validity into criterion-related validity. It was clear that at this point FV had been obliterated by the prevailing views of validity up until that time.

According to the new Standards, content validity provided evidence that items sample some definable universe of content, and hence lay the ground for interpreting their performance. In addition, the content validity of an achievement test was to be judged with respect to the goals of an educational program. This description of content validity was more broadly defined to include subject matter experts and behavioral or instructional objectives. Construct validity, however, is studied when there is no criterion or universe adequate for defining some quality for which measurement is desired. According to Cronbach and Meehl (1955), construct validity entails specifying a network of propositions which lead to the prediction of relationships among observables. As discussed previously, the concepts of construct validity and content validity are

intricately related. On one hand, content validity should relate to the desired goals of instruction rather than to the coverage of materials. On the other hand, the goals of instruction represent constructs inferred from the responses to items as operational definitions (Michael, 1961).

Criterion-Referenced Measurement

In the 1960s, the separation between the psychometric community and the judgment-oriented measurement and evaluation community began to expand. Glaser (1963) put forth the distinction between criterion-referenced measurement, which depends on absolute standards, and norm-referenced measurement, which depends on relative standards. Ebel (1962), however, suggested that because idiosyncrasies of test developers influence the content of the test, "…most objective tests rest on highly subjective foundations". (p. 21). In the 1970s, Cronbach's (1971) now-famous remark that "one validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447) revolutionized test validity. To validate interpretations of criterion-referenced test results, it is necessary to proceed beyond considerations of content validity and, according to Messick (1975), construct validity studies are necessary for establishing the meaning of measurement results.

Although there are problems with conceptualizing content validity using the framework of criterion-referenced measurement (see Fitzpatrick, 1983; Guion, 1977; Messick, 1989), studies employing the judgments of content specialists essentially address the nature of the test items and their congruence to objectives. Studies of construct validity, however, are intended more for the determination of the meaning of scores and not the meaning of items. In some sense, a problem that still exists for validation based on responses to the items is that their meaning may vary depending on who is making interpretations and for what purpose. When scores are interpreted in one way or another, inferences are made by the interpreter, which calls

for construct validation (Linn, 1979). Consequently, if interpretations vary, there are implications for the test's construct validity. In other words, the lack of consensus regarding score-based interpretations is, in fact, PBE regarding caveats in the interpretations that underlie validity.

In a more logical analysis, Turner (1979) deduced that face validity is a more fundamental concept than construct validity because some measures must be face valid in order to use the concept of construct validity. For example, suppose we develop a new test, test A, intended to measure creativity. If test A correlates well with a well-known creativity test, test B, we may use that as evidence for test A's construct validity. But how do we know test B measures creativity? It is hoped that when test B was developed, it too was correlated with some other well-known creativity test. How far can we take that chain of arguments? Turner claims that at some point there had to be a test for which validity was simply assumed to exist because there were no prior measures of creativity. There had to be a test that was face valid in order for all other tests to be construct valid.

Validity as Argument

Validation of interpretations, which emerged in the last quarter of the 20th century, left no room for FV, either. Consider the period starting in 1971 when Cronbach's argument for validating inferences challenged the trinitarian view of validity as: content, construct, and criterion (which still exist in practice; see Lissitz & Samuelsen, 2007). Validation of interpretation emerged from Messick's (1989) chapter in *Educational Measurement*, in which he refers to validity as "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" (p. 20).

The focus on validating the interpretations of test scores compelled researchers to think more thoroughly about the way in which arguments are developed and evaluated. Kane (2006) refers to Toulmin (1958) and House (1980) as the historical roots for the framework of validity argument. According to this approach, test validators should first construct an interpretive argument by specifying the network of assumptions and inferences that underlie the proposed interpretations and uses of test scores (Cronbach, 1988; Kane, 2006, 2013). That is, one should begin the process of validation by explicating the logical argument that leads from observed performance to conclusions about examinees and to any decisions and actions based on these conclusions. Then, validators should evaluate the interpretive argument by studying its clarity, coherence, and plausibility using evidence to support or refute its underlying assumptions and inferences. The evidence used for validation should be collected from five sources: (a) the test content, (b) its internal structure, (c) the underlying response processes, (d) relations to other variables, and (e) the consequences of testing.

The process of evaluating the accumulated evidence is the basis for developing the validity argument, which articulates the degree of confidence attributed to the proposed uses and interpretations of the test. To rigorously evaluate the plausibility of the interpretive argument, one must also consider the plausibility of alternative interpretations and uses of test scores. These alternatives attempt to uncover caveats in the proposed interpretations, as Cronbach (1980) suggested with respect to validation: “A proposition deserves some degree of trust only when it has survived serious attempts to falsify it” (p. 103). We believe that to accomplish this endeavor, validators need to collect both SBE and PBE.

Kane (2006) refers to the concepts conveyed by FV in the context of test critics and consequential validity. Kane’s description of FV also relates to the plausibility of an interpretive

argument. Efforts to strengthen FV are usually aimed at increasing the acceptance of the test among the examinees and other stakeholders. Similarly, Messick (1989) noted that lack of FV can influence the performance of examinees and the acceptance of the test by users and the public, and therefore, "… face invalidity should be avoided whenever possible". (p. 19). Face invalidity is the situation where various constituents do not think the adequacy and appropriateness of score-based inferences and actions are empirically or theoretically supported. Herein lays a contradiction: On one hand, FV is not evidence of validity, but on the other, lack of FV undermines the purposes of the test and, consequently, its validity. This begs the question "Can we support the validity of test scores by showing that they *do not lack* FV?"

The only relevant methodological treatment of FV can be found in Nevo (1985). Nevo argued that FV is important because it can affect (a) examinees' motivation to prepare and perform well; (b) the willingness of potential examinees to take the test; (c) the level of dissatisfaction of examinees with low scores; (d) the opinions of decision makers regarding the use of the test; and (e) the opinions of the general public, the media, and the judicial system. He then gave an operational definition of FV in the following mapping sentence: A RATER who is a(n) [testee/nonprofessional user/interested individual] RATES a [test item/test/battery] BY EMPLOYING a(n) [absolute/relative] TECHNIQUE AS [very suitable (relevant)… unsuitable (irrelevant)] FOR ITS INTENDED USE. Nevo's definition builds on validity as the appropriateness of IU of test scores. It indicates that useful evidence might stem from perceptions of various individuals about relevant attributes of the test. Although this mapping sentence was suggested more than 25 years ago, it has not been popularly applied.

Application of PBE in Test Development and Validation

Our conclusion from the historical review is that concepts underlying FV have been, regrettably, abandoned by the measurement community. FV is missing from the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999, 2014) and is treated minimally in the fourth edition of *Educational Measurement* (Brennan, 2006); also, most publications about validity theory lack any serious discussion of the topic. In fact, while the concept of validity has dramatically evolved over the past 70 years, the discussion of FV remained focused on what it is not (i.e., not real validity) rather than on what it could be. Consequently, we believe that both semantic confusion and pragmatic confusion exist regarding FV and other subjective judgments about attributes of a test. We explain the nature of this confusion and show how the concept of PBE helps elevate it.

The semantic confusion arises because researchers, as noted by Mosier and others, vary in how they interpret the term *face validity* and in how they use the concepts conveyed by it for validity research (see also, Newton and Shaw, 2014). This confusion stems from the definition of FV. Traditionally, FV is conveyed by the question "Does the test, on the face of it, measure what it is supposed to measure?" We argue that this definition is simplistic and outdated and needs to be revised to reflect current conceptions of validity. The traditional FV refers only to the appearance of the test and not to the IU of test scores. Moreover, the definition refers to only one claim from the complex chain of inferences that underlie the validity argument. Therefore, FV refers to a notion of validity that is no longer supported by the research community. In addition, the question raised by the traditional definition is too simplistic because it is unlikely to have a single answer. For example, people might think the test is satisfactory but object to how society uses the scores, or they might think the test measures what it purports to measure but it is too

long. In fact, opinions about the test could be as varied as the network of arguments that researchers aim to validate. Therefore, the modern conception of FV should be much broader than the traditional definition suggests.

Traditionally, FV refers to a holistic statement about the quality of the test, uttered by an untrained individual. A modern conception of FV would be more aligned with what PBE represents: evidence that evokes alternative interpretations and enables a reality check for the clarity and plausibility of validity arguments. Relevant PBE could cover perceptions about the content of the test items; examinees' response processes; the scoring procedure of the test; the way in which the results from the test are reported, interpreted, and used; and even their impact on individuals and societies. Theoretically, perceptions about any specific aspect of the test could be considered as relevant evidence for inferences made regarding that aspect of the test. Moreover, these perceptions can be culled from constituents with varying levels of familiarity with the test or expertise in the content area and thus reflect where disparities exist regarding the clarity and plausibility of the validity argument.

We also believe that there is pragmatic confusion about FV because the literature lacks research-based methodologies for collecting and using PBE. In the past 70 years, the measurement literature focused on SBE: which measures are appropriate and how they should be collected, analyzed, and used in a validation study. The pragmatic confusion arises when researchers wish to use PBE to enhance their validation analyses. They are left in the dark because, in the validity literature, there are no guidelines about what might be good measures to collect about people's perceptions, how best to measure these perceptions, and how to use this information in evaluating evidence for a validity argument. The only exception is the literature on content validation, where PBE is routinely collected. What the validity literature is lacking is

a comprehensive treatment explaining how PBE should be collected and used in various phases of test development and validation.

Our conclusion is that in an effort to dissuade practitioners and researchers from using FV as the *only* type of evidence of validity, measurement experts discredited FV but also, regrettably, the utility of PBE. Today, validity studies rely on more rigorous statistical analysis, and researchers rarely publish validation studies relying solely on FV. To illustrate this point, we reviewed papers published in relevant peer-reviewed journals[1] between 1995 and 2014 that had the term validity in their abstracts. Of the 6,334 papers we found, 109 (2%) also mentioned FV, but in none of them was it the only source of validity presented. We believe that researchers continue to use a term that is known to be misleading because they find this evidence useful, but they lack a better term to describe their work. We believe the term *perception-based evidence* is a better descriptor of such practices, and its collection can be accomplished in a manner that is both methodological and completely consistent with the modern validation framework.

Usefulness of PBE for the Validity Argument Framework

We see three ways in which PBE is relevant to the validity argument framework: (a) evaluating plausibility and clarity of the IU argument, (b) identifying alternative arguments, and (c) monitoring public opinion. PBE can originate from each of the five validity sources (content, internal structure, response processes, relations to other variables, and consequences). In addition, PBE plays a crucial role in test development, which we discuss later.

According to Kane (2013), the role of the validator is to evaluate the completeness, clarity, and plausibility of the argument of IU. To achieve this, validators need to evaluate the

[1] The search was conducted in Psychinfo under the journal classifications: educational measurement, testing, and educational psychology.

evidence for the proposed assumptions and inferences and identify the most problematic aspects of the interpretive argument. Based on that, the argument is rejected or adjusted until all inferences are plausible. The main reason PBE is relevant to this task is that perceptions are evidence regarding the extent to which the argument seems sufficiently plausible, clear, and coherent to relevant stakeholders: examinees, test users, and decision makers. Validators could compare expert and layperson perceptions regarding specific claims to identify points of agreement and disagreement. Issues where everyone agrees show support for a strong argument. Issues where perceptions differ are indicative of lines of argument where the claims are unclear or the inferences are not very plausible.

We believe that researchers need to evaluate their claims using both SBE and PBE. SBE should be used for establishing the psychometric soundness of the claims, and PBE should be used for establishing the plausibility and clarity of the claims. For example, people often reject tests in general because they believe all tests are biased. Not surprisingly, validators must show evidence to establish claims about the fairness of the test. Researchers can measure examinees' and test users' perceptions about test biasedness before and after exposing them to informatively designed evidence. The resulting PBE can be used in evaluating the plausibility of claims about the fairness of the test given the SBE at hand.

A second way in which PBE is useful for validation relates to how arguments are evaluated. To evaluate the plausibility of a proposed argument, test validators need to juxtapose their claims against various alternatives. A good source for alternatives can be the beliefs held by examinees, test users, and decision makers regarding the interpretations and uses of the test scores. The views of examinees and test users regarding what the test measures (and how well) can provide insights regarding construct deficiency or irrelevant variance. These constituents are

more likely to provide real alternatives (i.e., ones that significantly differ from the intended interpretations and uses) than the test developers who are limited by their intimate familiarity with the test. In that sense, collecting PBE can be seen as a process helping researchers develop alternative inferential networks. Essentially, validators need to provide SBE to justify the proposed interpretations and uses of the test scores as opposed to these alternatives.

For example, examinees often reject vocabulary items, claiming they measure only rote memory. This complaint can be phrased as an alternative claim regarding the extrapolation from test score to underlying construct. Researchers may design a study that compares the variation accounted for by language ability versus short-term memory and use the results to evaluate the examinees' claim against the test developer's claim. The study should present compelling evidence to support the interpretative argument regarding the construct being measured or otherwise make the appropriate adjustments to the test or the argument. Such an endeavor would improve the scientific creed of the test as well as its public relations.

Finally, as many have noted previously, face invalidity creates a unique threat to the existence of the test. If the public objects to using the test, or if test users seriously question its appropriateness, it is likely that the test will not be used in practice; in that case, there is no sense talking about the validity of IU of test scores. Test developers can do much to ensure that the test maintains desirable psychometric properties. This is crucial at a professional level and helps in avoiding many threats to validity in general. However, the public is usually unaware of or uninterested in the psychometric properties of the test. For laypeople, forming an opinion about the test is motivated more by satisfaction (or lack thereof) with their test outcomes and less by the test's reliability coefficient. Consequently, it is much easier and more common for the public

to criticize the test using the type of arguments that are perception based rather than to criticize its psychometric properties (i.e., qualities that depend on SBE).

Test developers must realize that the opinion of the public matters because it has the power to determine the fate of the test. For example, if examinees or users have a choice among multiple tests, they are likely to choose the test that is perceived to be more appropriate to the task at hand. If there is only one test, then examinees can protest or take other political action to advocate the development of a different test or of different criteria for decision making; they can even take steps to abolish the test completely. Similarly, people's perceptions about the test affect their motivation and preparation and consequently their performance. If people believe the test is not what it purports to be, they are likely not to be motivated to perform as well as they could (Nevo, 1985). The bottom line is that negative perceptions should not be overlooked because they undermine the basic assumptions of the validity argument. These validity threats cannot always be dealt with by increasing the psychometric rigor of test development. PBE could be useful for identifying the sources of such threats and helpful in finding ways to address them. Test validators should collect PBE to evaluate the extent to which perceptions held by the public regarding the interpretive argument may pose threats to the validity of test scores and possibly affect the fate of the test.

Use of PBE for Validation Throughout the Life Cycle of a Test

To ensure quality measurement, validity considerations should guide all phases of test development and use. We propose that PBE provides input for evaluating validity-related issues at four distinct stages of a test's life cycle: inception, test development, validation, and ongoing operation. If PBE is routinely documented, researchers will gain valuable evidence for facilitating test development and use (Secolsky, Wentland, & Denison, 2011).

Inception. Perception-based evidence is involved in the creation of a test from its inception as a measurement tool with a particular purpose. Theoretical propositions stemming from relationships among variables and the development of constructs require judgment. These judgments are based on researchers' perceptions of existing research and theories and are used in justifying the purpose of the test and the choice of constructs the test is intended to measure. An example of the use of PBE at this stage is that when a test developer is designing a battery for employment selection, his or her perceptions of relevant variables precede the collection of SBE. What we choose to measure is based on what we perceive to be important or relevant. Therefore, these perceptions should be documented in the conceptual assessment framework (see Mislevy & Riconscente, 2006) or the rationale for the test. These documentations are valuable for developing the argument of IU.

Another useful PBE related to inception is the documentation regarding the necessity of developing a test for a particular purpose, as given by test users, examinees, policy makers, the media, and the public. In addition, the expectations of these constituents regarding test design and use are likely to influence their perceptions of the actual test. Not only is this information useful for developing the actual test, but the validity argument should also benefit from it. Specifically, if validation means ensuring that the test scores are interpreted and used for their intended purposes, then test developers need to evaluate validity evidence not only with respect to how *they* perceived these purposes to be but also with respect to how society did. If views diverge, the reasons for these differences should be studied and efforts should be made to minimize the differences. Similarly, Jensen (2000) argues that such considerations should be emphasized in test construction in order to avoid the negative impacts of testing.

Test development. Researchers commonly use the perceptions of both experts and examinees in creating items and improving their qualities. Subject matter experts (SMEs) review items for inclusion on a test, evaluate their alignment with objectives and content standards, or comment on the technical adequacy of items. Researchers typically ask SMEs to rate each item's adequacy with respect to various criteria to ensure content validity and technical quality of items. When experts' ratings show adequate reliability, they can then be used alongside traditional item analysis (SBE) in deciding which items should be removed or changed.

If experts can have varying interpretations, then the variability of examinee interpretations is likely to be even greater, and perhaps independent of the item's psychometric properties (Secolsky, 1983). Researchers may choose to collect the perceptions of examinees regarding item difficulty, item design, score meaning, and perceived knowledge or skill demands of the test. Test developers also use their own perceptions when attempting to understand why items or people behave differently than expected. All these methodologies are in the realm of PBE and are routinely collected and analyzed for guiding revisions during the test development stage. The same PBE can also be used for validation by showing support to claims regarding the structure and content of the test.

Validation. The role of PBE in validation is tied to its role in test development. In many ways, these roles support a similar function: to provide evidence that the test and examinees are acting in accordance with expectations. In the case of test development, PBE is used for test improvement, and in the case of validation, PBE should be used in constructing and evaluating validity arguments. As explained earlier, this can be done by using PBE to generate alternative arguments and to evaluate the clarity and plausibility of specific chains of inference. Many forms of PBE can be used for these purposes. As now shown, PBE can enhance each of the five sources

of validity evidence listed in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014): test content, response processes, internal structure, relations to other variables, and consequences of testing.

Test content. SME perceptions about the adequacy of items can be used for establishing content validity (see also, Sireci & Faulkner-Bond, 2014). For example, the procedure described by Rovinelli and Hambleton (1977) aims to assess the extent to which content specialists deem items as measures of well-defined objectives. An item with a high index of item-objective congruence indicates that SMEs are in agreement their perception about the item-generating domain. Similarly, examinees' perceptions of the alignment between the instruction and the items can be used for establishing instructional validity. These PBE are essential for evaluating claims of generalizability (see also, Secolsky, Wentland, & Denison, 2011).

Underlying response processes. Perceptions of examinees regarding their response processes are commonly collected using various methodologies: think-aloud protocols, cognitive interviews, focus groups, member checks, and self-reports. These forms of PBE are commonly used for gaining insights about actual response processes and consequently establishing extrapolation claims regarding construct validity (e.g., Padilla & Benítez, 2014).

Internal structure. Perceptions of measurement experts regarding the dimensionality of the test are given by their choices in running and interpreting factor analyses. As Mislevy, Moss and Gee (2009) argue, there exists a qualitative frame that surrounds quantitative research on validity arguments. More specifically, the perceptions of test developers or SMEs regarding interconnections between items guide the design of confirmatory analyses, the interpretation of results, and the resulting modifications to the test. Therefore, the plausibility of inferences about test dimensionality and item quality should be supported by measures of consensus among

relevant experts. This PBE could also be used for developing a compelling argument for construct validity. These claims can be contrasted with alternative claims based on the perceptions of examinees and test users regarding what they think the test really measures. Relevant SBE should then be presented to support the claims in the interpretive argument.

Relations to other variables. The question on how to make inferences from SBE revolves around the individual researcher or user and, to a considerable extent, contains subjective judgments. For example, as one part of an inference chain (or a nomological net), is a correlation of .50 high enough to be considered sufficient evidence to support a particular claim of validity? The interpretation of any correlation matrix as showing convergent and discriminant validity mainly represents researchers' perceptions of the results, which may or may not coincide with other experts' views or with examinees' views. To evaluate such claims, validators need to show that their interpretations are plausible to both experts and laypeople. For example, Karelitz (2013) surveyed more than 8,000 past and future examinees regarding their perceptions of the admission test for higher education in Israel (Psychometric Entrance Test, or PET). Although the PET has consistently been shown to be a good predictor of academic performance over the past 30 years, (Oren, Kennet-Cohen, & Bronner, 2007), more than two-thirds of the survey respondents thought its predictive power was negligible. Because the ability to predict performance is a crucial piece of the argument for using the test, this evidence highlights a validity threat that needs further consideration.

Perception-based evidence is relevant as evidence regarding the consequences of testing, specifically for identifying validity threats based on the misuse or misinterpretation of test scores. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) state that researchers should "consider the perspectives of different interested parties,

existing experience with similar tests and contexts and the expected consequences of the proposed test use" (p. 12). Validators could study these perceptions by surveying examinees, test users, policy makers, the media, and relevant public groups and organizations. To construct a compelling argument, the claims given in the interpretive argument should be evaluated against relevant SBE but also against the relevant PBE.

In the study mentioned previously, Karelitz (2013) reports that roughly 70% of the respondents (mainly past and present examinees) thought that PET deters people from applying to higher education. This negative consequence seems to contradict the very purpose of the test and thus poses a serious threat to validity. The truth is that only about 50% of the Israeli students earn high school diplomas, which is the first requirement for admission to the universities. Because admission is based on a weighted average of high school graduation grades and the PET score, those who did not fare very well during high school are given a second chance by the PET. However, applicants do seem to avoid taking the PET because many of them apply to colleges that commonly require only a high school diploma. Consequently, policy makers forced universities to determine admissions cut points based on either the graduation exams or the PET.

Ongoing Operation. Many things can happen during the ongoing operation of a testing system, including logistic problems, issues of test security, rise of anti-test groups, and decline in the number of examinees. The way the testing institution and the relevant decision makers handle these issues will affect the system's public reputation, which can have a lasting effect on its success. In fact, it could very well determine whether the testing system will continue as is, change to adapt to current concerns, or cease to exist altogether.

Evidence based on the perceptions of various constituents can be used in identifying trends that may affect the future of the test. For example, test takers may find the test unfairly

more difficult than they believed is appropriate. Such perceptions can potentially affect the popularity of the test and consequently its very existence. In that sense, PBE can indicate whether a testing program will be sustainable and, if collected routinely, can alert developers and users about possible threats stemming from misinformation or misuse.

Summary

This article has two main claims. First, the term *face validity* is too simplistic, outdated, and negatively loaded to be used in current scientific discourse about validity (apart from historical reviews). Second, perceptions influence how the test is conceived, developed, evaluated, implemented, and accepted by society. We propose that researchers should routinely collect, analyze, and report evidence based on the perception of various constituents about aspects of the testing system. We contend that such data, which we call perception-based evidence, are useful for test development and validation. Specifically, PBE could be used for the following purposes:

1. To support decisions made during construction or modification of tests.
2. To enhance evidence of validity collected regarding test content, underlying response processes, internal structure, relations to other variables, and consequences of testing.
3. To generate alternative claims about the test and the interpretation and use of test scores.
4. To evaluate the clarity and plausibility of claims in the interpretive argument.
5. To identify threats to validity and evaluate the sustainability of the test.

We believe that many researchers are already collecting various types of PBE during test development and validation. This practice is desirable and should be expanded, and supported with proper methodological literature. We propose that, to gain the most from perception-based evidence, researchers use a variety of PBE in generating and evaluating claims during validation.

References

Adams, S. (1950). Does face validity exist? *Educational and Psychological Measurement, l0,* 320–328.

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

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

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: American Psychological Association.

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

Brennan, R. L. (Ed). (2006). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement: Measuring achievement, progress over a decade: No. 5* (pp. 99–108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

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

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

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

Ebel, R. L. (1962). Content standard test scores. *Educational and Psychological Measurement*, *22*(1), 15–25.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement*, *7*, 3–13. doi:10.1177/014662168300700102

Foran, T. G. (1930). *The meaning and measurement of validity*. Washington, DC: Catholic Education Press.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*(8), 519–521.

Goodenough, F. L. (1949). *Mental testing: Its history, principles, and applications*. New York: Rinehart.

Guilford, J. P. (1940). Human abilities. *Psychological Review*, *47*, 367–394.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *3*, 427–438.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist*, *5*, 511–517.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, *1*, 1–10.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: SAGE.

Jensen, A. R. (2000). Testing: The dilemma of group differences. *Psychology, Public Policy, and Law*, *6*, 121–127.

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

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000.

Karelitz, T. M. (2013). Using public opinion to inform the validation of test scores. *Research Report No. 387.* Jerusalem: NITE.

Lindquist, E. F. (1935). Objective achievement test construction. *Review of Educational Research*, *4*, 469–483.

Lindquist, E. F. (1936). The theory of test construction. In H. E. Hawkes, E. F. Lindquist, & C. R. Mann (Eds.), *The construction and use of achievement examinations: A manual for secondary school teachers* (pp.17–106). Cambridge, MA: Riverside Press.

Lindquist, E. F., & Cook, W. W. (1933). Experimental procedures in test evaluation. *Journal of Experimental Education*, *1*(3), 163–185.

Linn, R. L. (1979). Issues of validity in measurement for competency-based programs. In M. A. Bunda and J. R. Sanders (Eds.), *Practices and problem in competency-based measurement.* Washington, DC: National Council on Measurement in Education.

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

McClung, M. S. (1977). Competency testing: Potential for discrimination. Clear*inghouse Review*, *11*, 439–448.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966. doi:10.1037/0003-066X.30.10.955

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

Michael, W. B. (1961). Problems of validity for achievement tests. *18th Yearbook of the National Council on Measurement in Education*. 1–12.

Mislevy, R. J., Moss, P. A., & Gee, J. P. (2009). On qualitative and quantitative reasoning in validity. In K. Ercikan & W. M. Roth, (Eds.), *Generalizing from educational research: Beyond qualitative and quantitative polarization* (pp. 67–100). London, United Kingdom: Taylor and Francis.

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

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*, 191–205. doi: 10.1177/001316444700700201

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*(4), 287–293.

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

Oren, C., Kennet-Cohen, T. & Bronner, S. (2007). Aggregated data about the validity of the higher education selection system for predicting academic success in the first year (the 2003-2005 cohorts). *Research Report No. 342.* Jerusalem: NITE. [in Hebrew].

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144.

Rovinelli, R. J. & Hambleton, R. K., (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Tijdschrift voor Onderwijsresearch*, *2*, 49–60.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, *16*, 290–296.

Secolsky, C. (1983). Using examinee judgments for detecting invalid items on teacher-made criterion-referenced tests. *Journal of Educational Measurement*, *20*(1), 51–63.

Secolsky, C. (1987). On the direct measurement of face validity: A comment on Nevo. *Journal of Educational Measurement*, *24*(1), 82–83.

Secolsky, C., Wentland, E. & Denison, B. (2011). The need for documenting validation transactions: a qualitative component of the testing validation process. *Quality and Quantity*, *45*, 1303–1311.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Turner, S. P. (1979). The concept of face validity. *Quality and Quantity*, *13*, 85–90.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Turney, A. H. (1934). The concept of validity in mental and achievement testing. *Journal of Educational Psychology*, *25*(2), 81.

Vincent, E. L. (1924). A study of intelligence test elements (No. 152). Teachers college, Columbia University.