Utility vis-à-vis Validity

Charles A. DePascale and Damian Betebenner
National Center for the Improvement of Educational Assessment, Inc.

Presented at the 41st Annual International Association for Educational Assessment Conference
Lawrence, Kansas
October 2015

The Three Most Important Considerations in Testing: Validity, Validity, Validity

Validity – the extent to which inferences drawn from test scores are appropriate – is by far the most important technical characteristic of a test. But because it is much more challenging to establish validity than other test desiderata, it has gotten short shrift in most testing programs. How can advances in technology, test development, psychometrics, and score reporting improve the validity of our testing programs?

In two simple sentences, the theme of this 2015 conference of the International Association for Educational Assessment (IAEA) presents the great paradox of educational assessment. Validity, widely accepted as *the most important technical characteristic of a test* is given *short shrift in most testing programs*. In the second sentence, the organizers posit that the relative difficulty in establishing validity is the reason that it does not receive the attention it deserves in testing programs. This premise logically leads **outward** to a search for solutions (i.e., advances in technology, test development, psychometrics, score reporting) that will decrease the challenge in establishing validity, making it more practical to give validity its due, and ultimately, to *improve the validity of our testing programs*.

In this paper, however, we propose that any effort to improve the validity of testing programs must begin by looking **inward** at the meaning of validity and validation. The seemingly straightforward definition of validity as *the extent to which inferences drawn from test scores are appropriate* belies not only the complexity of concept, but also the level of disagreement among scholars and theorists over fundamental aspects of what validity comprises and what constitutes the process of validation (Newton & Shaw, 2015). Over the last six decades, validity has been defined by a cycle of seminal works refining the concepts of validity and validation (Cronbach, 1971; Messick, 1989; Kane, 2006), countless articles, books, and often contentious debate attempting to interpret the latest views on validity (Ebel, 1961; Wainer & Braun, 1988; Newton & Shaw, 2015), and successive revisions of the *Joint Standards* (1984, 1999, 2014) attempting to codify the field's current understanding of validity.

Of course, theoretical debate about the meaning of validity and validation does not take place in a vacuum. Kane's 2006 chapter on Validation was published on the cusp of full implementation of the annual testing and accountability requirements of No Child Left Behind. The 2014 Joint

Standards (AERA, APA, NCME 2014) were published in the midst of debate about the high-stakes use of educational assessments for teacher accountability as well as school and student accountability, a desire to use assessment results to measure growth in addition to status, demands to measure more complex knowledge and high-level cognitive skills contained in Common Core State Standards (CCSS), and a call for assessments to provide more and better information to inform instruction .

As tests have taken on more and more prominent social uses, the concept of validity and the enterprise of test validation have expanded to accommodate consideration of the appropriateness and social consequences of those uses. However, in our own work, increasingly we find discussions of test validity extending to issues not traditionally included under the test validation framework. In particular, as tests have become essential components of high-stakes accountability systems, the concept of the validity of a test and test scores is often confounded with, to the point of being indistinguishable from, the validity of the testing program, the validity of the accountability program, and the efficacy of curricular, instructional, or other initiatives intended to improve instruction and student learning. In this regard, we find that the concerns expressed by Popham in 1997 were prescient:

> Although every right-thinking measurement person ought to be concerned about the consequences ensuing from a test's use, it does not follow that test-use consequences need to be linked to the now widely held view that validity is rooted in the accuracy of inferences we derive from examinees' test performances. I believe, therefore, that the reification of consequential validity is apt to be counterproductive. It will deflect us from the clarity we need when judging tests and the consequences of test-use. (p. 9)

As one example of the real problems caused by the lack of clarity resulting from the intersection of this expanded use of educational assessment and expanded framework for validity and validation we present the topic of this paper, utility vis-à-vis validity.

Utility within the Validity Framework

Messick (1979, 1989, 1990) presents a four-faceted view of validity in which the relevance and utility of a test plays a prominent role. Messick describes the four facets as

(1) An inductive summary of convergent and discriminant evidence that the test scores have a plausible meaning or construct interpretation,
(2) An appraisal of the value implications of the test interpretations,
(3) A rationale and evidence for the relevance of the construct and the utility of the scores in particular applications, and
(4) An appraisal of the potential social consequences of the proposed use and of the actual consequences when used.

Messick further categorizes these four facets based on the source of the justification (evidential or consequential) and the function or outcome of testing (interpretation or use).

Figure 1

Facets of test validity (Messick 1979, Figure 1)

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

In the iterative feedback loop that accompanies the table above, Messick indicates a clear and distinct need to gather evidence for the relevance and utility of test scores for their proposed use that is distinct from an appraisal and judgments on the value implications and social consequences of the interpretation and use of those test scores. Describing the relationship of relevance and utility to construct validity, Messick explains

> The empirical verification of this rational hypothesis [relating the construct to performance in the applied domain] contributes to the construct validity of both the measure and the criterion, and the utility of the applied relation supports the practicality of the proposed use. Thus, the evidential basis of test use is also construct validity, but elaborated to determine the relevance of the construct to the applied purpose and the utility of the measure in the applied setting. (pp. 20-21).

Utility Examples from Genetic Testing and Clinical Diagnostic Testing

Outside of educational assessment, the concept of utility as a distinct factor to be considered in the evaluation and selection of genetic and clinical diagnostic tests is gaining acceptance. Instead of focusing solely on the diagnostic accuracy of clinical tests, in selecting and approving tests for use policy makers are also seeking information on whether the use of the test leads to an improvement in final outcomes in typical populations (Bossuyt et al. 2012). Bossuyt et al. define the utility of a particular diagnostic test or testing program is as "the degree to which actual use of the corresponding test in healthcare is associated with changing health outcomes, such as preventing death and restoring or maintaining health."

A prime example is the ACCE framework developed by the Centers for Disease Control Office of Public Health Genomics which defines four main criteria for evaluating genetic tests (CDC, 2010):

- **A**nalytic validity – How accurately and reliably the test measures the genotype of interest.
- **C**linical validity – How consistently and accurately the test detects or predicts the intermediate or final outcomes of interest.
- **C**linical utility – How likely the test is to significantly improve patient outcomes.

- **E**thical, legal, and social implications – Ethical, legal, and social implications that may arise in the context of using the test.

In addition to the accuracy and reliability of the test results relative to both the construct and the criterion of interest (analytic validity and clinical validity), the ACCE model explicitly examines the likelihood that the information provided by the test will lead to improved patient outcomes. The model includes 44 targeted questions about the disorder itself and the four areas of interest, 16 of which are devoted to Clinical Utility.

Table 1
ACCE 16 Targeted Questions on Clinical Utility

1. What is the natural history of the disorder?
2. What is the impact of a positive (or negative) test on patient care?
3. If applicable, are diagnostic tests available?
4. Is there an effective remedy, acceptable action, or other measurable benefit?
5. Is there general access to that remedy or action?
6. Is the test being offered to a socially vulnerable population?
7. What quality assurance measures are in place?
8. What are the results of pilot trials?
9. What health risks can be identified for follow-up testing and/or intervention?
10. What are the financial costs associated with testing?
11. What are the economic benefits associated with actions resulting from testing?
12. What facilities/personnel are available or easily put in place?
13. What educational materials have been developed and validated and which of these are available?
14. Are there informed consent requirements?
15. What methods exist for long term monitoring?
16. What guidelines have been developed for evaluating program performance?

With even a cursory review of the 16 utility questions presented in Table 1, it is clear that such an approach would be relevant to the evaluation of policies related to the implementation and use of educational assessments for particular purposes. Assuming that the overarching purpose of most educational policies and initiatives is to lead to improved student learning, when implementing a testing program or a test-based accountability system, it would certainly be valuable to have answers to questions regarding how educators, students, or policy makers will be able to interpret and use the results of the tests that are administered.

Conceptions of Utility in Educational Assessment and Accountability

We argue that consideration of utility has been largely overwhelmed in the validity framework by questions of construct validity and content alignment or inappropriately subsumed under concerns about consequences of test use. Despite that, there have been conceptions of utility that

have managed to emerge with regard to the utility of tests, testing programs, and test-based accountability systems. In this section of the paper, we describe the way in which an argument for utility can be built that is, in fact, separate from other aspects of the validity and the validation process.

Gottfredson & Crouse (1986) address the question of validity versus utility in the context of the use of the SAT in the college admissions process. As was the case with the clinical examples above, Gottfredson & Crouse separate questions of the technical quality of the SAT in terms of construct validity or predictive validity from questions of its practical utility. In short, they describe Crouse's work as examining the general question, "even if a test is unbiased and predicts desired criteria well, how should it be used in practice, if at all?" In summarizing their utility argument, Gottfredson & Crouse present four questions to "provide a broad systematic perspective that can promote greater awareness of the ultimate consequences of different choices and this contribute to better informed decisions."

1. What is the main *purpose* for using the test in question?
2. In what sort of selection or appraisal *system* is the test embedded?
    a. For example, what other tools (e.g., high school record) are currently or potentially available for helping to meet the same goals? And how stable are the properties of these other tools with and without the presence of the test?
3. Who are the major *stakeholders* (e.g., individuals, colleges, secondary schools, professional groups, employers, minority groups, majority groups), and how might they react in both the short and long run to particular changes in test design and use?
4. What *criteria* (costs and benefits) should be considered in assessing overall utility?
    a. Criteria should be specific and measurable where possible.
    b. Choices depend on the values one attaches to the different potential costs and benefits.

[Adapted from Gottfredson & Crouse, p. 376]

The fundamental utility question being addressed is what value does the use of the SAT add to the college admissions process? Further, from a utility perspective, the value-added of the SAT is not considered solely on its own but in comparison to the use of no standardized tests or other possible standardized tests. This focus on *comparisons* is a central component of the utility question and is what makes utility distinct from other commonly applied aspects of the validation process. The emphasis on gathering evidence to support the utility argument is also consistent with Messick's placement of utility in the Evidential Basis dimension of his matrix.

A next logical step from considering utility in terms of gathering evidence and making comparisons is to frame the utility question as a formal exercise in experimental design. Briggs (2004) proposes the following in commenting on Kane's description of the interpretive argument:

> It seems clear that test validation must be seen as a comparative venture when it comes to evaluating causal inferences. We can conceive of taking a particular test as an experimental treatment, not taking a test or taking a different test as an experimental control, and the decision reached as an experimental outcome. (p. 172)

Considering the utility question from an experimental design perspective serves two important purposes. First, it forces the test user to think explicitly about the higher purpose of the proposed use of the test or testing program. As shown with the clinical testing example, the utility question focuses not on the limited purpose of the test (i.e., to make an accurate diagnosis), but rather on the decision to use the test for the broader purpose of improving health outcomes. Consequently, the test user must be able to define the desired outcome robustly enough to determine how it will be measured and how evidence will be collected and analyzed. Second, it forces the test user to think explicitly about alternatives to the use of this particular test. The user must consider the costs and benefits associated with the use of this test to accomplish the desired outcome compared to using an alternative test or perhaps using no test.

Briggs also notes that this type of evidence about the utility of a test cannot be collected until after a test has been administered. Most likely, if the desired outcome is meaningful and worth the investment necessary to implement a testing program, the test will have to be administered for multiple years before evidence of its effect on the outcome can be determined. Ho (2014), also acknowledged the need for multiple years of implementation in proposing a similar approach to investigating the utility of a proposed accountability system.

In Ho's thought experiment, the accountability system is considered the treatment. Alternative accountability systems, or perhaps no accountability system, would be the experimental control. The outcome variable would be one or more measurable factors directly related to the higher level purpose for implementing the accountability system in the first place (e.g., an increase in the percentage of students graduating college- and career-ready). At the end of a specified number of years, the utility of the accountability system would be determined by comparing its impact on the outcome variable to the impact of the control conditions.

As in the previous examples, the outcome variables used to evaluate the efficacy of the accountability system are related to the higher level purpose of the accountability program. Other factors such as cost, ease of administration of the accountability system and its component assessments, and consequences resulting from its implementation are all considered in concert with differences in the outcome variable when determining the overall utility of the accountability system. On the other hand, an understanding of the inner workings of the system such as the technical quality of the individual assessments administered falls outside of the immediate question addressing the utility of the accountability system. Of course, there are many reasons why it is useful and necessary to fully evaluate the technical quality of the system and its component parts, and we are not suggesting that the utility of the system is the only question that should be addressed. Separating the utility question from the question of technical quality, however, should increase the likelihood of accurately answering both questions.

Consequences and Utility

In practice, it can be quite difficult to separate the concepts of *consequences* and *utility*. In lay terms, Oxford online dictionaries defines *utility* as the state of being useful, profitable, or beneficial, and defines *consequences* as a result or effect of an action or condition. Clearly, there is an overlap between the terms when the state of being useful or beneficial comes about as the result or effect of an action. Turning to a more technical source, the Joint Standards defines consequences as follows:

> **Consequences**: The outcomes, intended or unintended, of using tests in particular ways in certain contexts and with certain populations. (p. 217)

Again, there is a clear overlap between this definition of consequences as outcomes and the ACCE definition of clinical utility cited previously: How likely the test is to significantly improve patient outcomes.

Messick (1990) adds to the fusing of utility and consequences with the explanation that the *Social Consequences* facet of his four-facet framework encompasses each of the other three facets:

> And once again, in recognition of the fact that the weighing of social consequences both presumes and contributes to evidence of score meaning, of relevance, of utility, and or values, this cell needs to include construct validity, relevance, and utility as well as social and value consequences.
>
> Thus construct validity appears in every cell, which is fitting because the construct validity of score meaning is the integrating force that unifies validity issues into a unitary concept. (p. 25)

In his conclusion, Messick further makes the case for the unitary concept of validity and for viewing his four facets, or four cells, as a single concept:

> As difference foci of emphasis are added to the basic construct validity appearing in each cell, this movement makes what at first glance was a simple four-fold classification appear more like a progressive matrix… One implication of this progressive-matrix formulation is that both meaning and values, as well as both test interpretation and test use, are intertwined in the validation process. Thus, validity and values are one imperative, not two, and test validation implicates both the science and the ethics of assessment. (p. 26)

Messick, therefore, with his four-facet framework breaks validity apart and puts it back together again. It is our view that all of this brings us back to the fact that there is virtually no room for disagreement with Popham's (1997) statement that *every right-thinking measurement person ought to be concerned about the consequences ensuing from a test's use*. From a practical perspective, however, there appears to be a clear choice to be made between the following

perspectives of Popham (1997) and Shepard (1997) on Messick's matrix and the value of considering all of its aspects jointly as validity.

Table 2

Alternative Views on the Consequences of the Messick Four Facet Framework and Unitary Concept of Validity

Shepard (1997) The Centrality of Test Use and Consequences for Test Validity	*Popham (1997) Consequential Validity: Right Concern – Wrong Concept*
In my view, the matrix was a mistake. Although Messick goes on at some length to emphasize that the facets cannot be pulled apart and considered independently and to remind us that construct validity resides in all cells, the temptation is too great to think that the traditional, scientific version of construct validity resides in the upper, left-hand cell and that consequences in the lower, right-hand cell are the business of moral philosophers and the politically correct.	Lumping our attention to the social consequences of test use with the concept of validity will not only muddy the validity waters for most educators, it may actually lead to less attention to the intended and unintended consequences of test use. Those consequences will be so masked by the subsuming and confusing framework of validity that they are likely to be overlooked. Such an unintended social consequence of Messick's framework would, of course, be genuinely unfortunate.

To some extent, perhaps both Shepard and Popham were correct. Few test developers and test users have established plans to collect and evaluate evidence on each and every one of the facets of validity described above. In particular, we find that questions of utility (i.e., intended outcomes) have been subsumed by a consideration of consequences (intended and unintended outcomes), in general. Moreover, by failing to address the utility question directly, test developer and test users have a tendency to focus on the narrow focus of consequential validity (i.e., the unintended consequences of implementing this particular program) rather than the broader focus of the concept of utility (i.e., the relative value of the implemented program compared to alternatives).

We also contend that the more serious problem is that the lack of focus on utility has led directly to a muddying of the waters between questions related to the validity of the use of a test and the validity of the use of the testing program or test-based accountability system for a particular purpose. Specifically, a reasoned consideration of utility with its emphasis on a) the higher-level purposes for administering the test and b) the consideration of alternatives to the implemented or proposed program, would inevitably lead to a broader consideration of the validity question, a more coherent and comprehensive theory of action, a more complete validation effort, and ultimately, to improved validity.

In the following section, we present three examples currently in the forefront of discussions in the United States about the validity of assessments to illustrate our concerns: testing time, school accountability, and teacher evaluation.

Significance of a Utility Mindset

In the preceding section, we argue that a lack of adequate attention to utility has led to a narrowing of the validity question and an inordinate focus on the test rather than the testing

program or accountability system of which the test is simply a component. The three examples presented here are intended to demonstrate how an insufficient focus on utility contributed to negative outcomes or limited the validity of the use of the test for the desired purpose.

Testing Time

As states across the United States implemented new assessments aligned to the Common Core State Standards (CCSS) or other college- and career-ready standards during the 2014-2015 school year, there was a tremendous backlash against the amount of time required by the new tests (Ujifusa, 2015). Accepting that testing time may be simply a proxy issue for a portion of those protesting the length of the tests, there is little question that testing time is a serious issue that could impact the long-term use of the new tests. Time and cost are routinely considered as constraints in the design of an assessment program, and state's new college- and career-readiness tests were no exception. Each of the national assessment consortia considered cost and time factors as they made design decisions almost from the beginning of the programs. Missing, however, was an adequate consideration of testing time from a broader utility perspective.

Large-scale state assessments serve an important function, but provide limited information that is directly useful to individual teachers and students (or provide limited information that individual teachers and students are prepared to use). In an effort to meet the requirements of construct validity (i.e., to fully measure the depth and breadth of the CCSS) and to measure student performance along the full proficiency continuum, test developers and test users did not adequately gather evidence to determine the point at which the burden of testing outweighed any perceived usefulness or benefits of the test results local educators. Deeper consideration of the utility issues prior to development may have led to earlier consideration of the alternative testing approaches designed to reduce testing time that are being examined now in several states.

School Accountability

Since the implementation of the assessment and accountability requirements of No Child Left Behind (NCLB), the 2002 reauthorization of the 1965 Elementary and Secondary Education Act, states have been required to implement high-stakes district and school accountability programs based primarily on student performance on annual Reading/Language Arts and Mathematics tests. Braun (2008) argues "consequential validity is the ultimate criterion by which we should judge an accountability system." In this case, Braun's use of the term consequential validity refers in large part to the extent to which the theory-of-action behind the implementation of the accountability system has been validated; that is, the extent to which the implementation of the accountability system has led to the intended higher level outcomes. This conception of the validity of an accountability system is consistent with the previously described Ho (2014) consideration of an accountability system as a treatment intended to produce a particular effect.

As NCLB was implemented, however, much of the validity discussion focused on issues related to construct validity or consequential validity of the use of tests for high-stakes accountability decisions (e.g., narrowing of the curriculum, focus on bubble students, and a lack of incentive to instruct high-performing students). Much less attention was paid to utility – the development of a rationale for and evidence of the ways in which teachers, administrators, and the public having the reading and mathematics test scores would lead to the desired higher-level outcomes such as

closing achievement gaps between students in identified racial/ethnic and socioeconomic subgroups and their peers.

The lack of attention to the utility of the NCLB tests-based accountability systems is ironic given that under the original 1965 ESEA those tests were intended as outcome measures to be used to evaluate the implementation and effectiveness of instructional and support programs designed specifically to achieve the desired higher-level outcomes (i.e., to gather evidence of utility). Specifically, the evaluation component of the 1965 act required

> that effective procedures, including provision for appropriate objective measurements of educational achievement, will be adopted for evaluating at least annually the effectiveness of the programs in meeting the special educational needs of educationally derived children…

Additionally, the use of tests was regarded "as a protection against the infusion of Title I funds into on-going school programs unlikely to upgrade the achievements of educationally disadvantaged children." (Bailey & Mosher, 1968).

Teacher Evaluation

Our third example of a fundamental disconnect in the current validity discussion is related to the use of student scores on state assessments in Reading/Language Arts and Mathematics to produce ratings of teacher effectiveness and to inform high stakes employment decisions. Let us begin by considering the construct that is being measured by the state mathematics assessment. Ideally, the assessment is measuring students' mathematics proficiency as defined by the state content standards and achievement expectations. We know that the test is not measuring teacher effectiveness. How then, are student test scores related to teacher effectiveness? Braun (2012) provides a thorough description of the steps that one must follow to get from student achievement to a test score to a measure of teacher effectiveness to a teacher effectiveness rating. He begins by with the reminder that the individual test score that results from the mathematics test is not used in the evaluation system until it has been combined with other test scores to produce a completely new statistic. Braun describes all of the pieces of the evaluation system that must be validated to support inferences drawn from the teacher effectiveness score. Questions that remain, however, are what is the role of that original mathematics test in the teacher evaluation system, what inferences are being made from that test score, and how must they be validated.

- Is it accurate to say that the mathematics test being used to determine teacher effectiveness?
- What aspects of the test design and reporting of student scores from the mathematics test are impacted by the use of the test in a teacher evaluation system?
- What inferences are being drawn from the student test score about teacher effectiveness?

Answering these questions correctly is more than an academic exercise. It is critical to establishing a strong validity argument for teacher evaluation systems.

For example, an incorrect response could lead to placing too much emphasis on validating the test score in isolation at the expense of validating the teacher evaluation system, the ratings (i.e., scores) that it produces, and the inferences drawn from those ratings about teacher effectiveness as a whole.

Distinguishing the validity of the test from validity of the evaluation system not only forces us to focus on preparing a strong argument to support the utility of the system, it also forces us to consider the validation of the test scores with regard to their use within and outside of the system. It may be the case that the use of the mathematics test as part of a high-stakes teacher evaluation system has a significant negative impact on the validity of student test scores and inferences drawn about individual student performance. In fact, some would argue that is a likely outcome. At the same time, however, dependent upon how those mathematics test scores are used in the teacher evaluation system, there may be little to no impact on the validity of teacher effectiveness ratings.

Conclusion

Throughout this paper, we have attempted to make the case that utility, a central component of the Messick framework, has been largely lost in discussions about the valid interpretation and use of the results of K-12 state assessment programs in the United States over the last two decades. In particular, there has been little systematic consideration of aspects of utility related to ensuring that systems are in place to maximize chances that the test results will be useful in attaining higher level purposes (e.g., improved learning, more effective teaching, and better schools) as demonstrated by the ACCE example.

We concede that it is likely that a major contributor to limited concern about utility in K-12 assessment is that the use of state assessments for district and school accountability is federally mandated and the parameters of those accountability systems have been tightly defined by the U.S. Department of Education (USED). When implementing an assessment and accountability system to comply with a federal law, there are few incentives to consider whether alternatives would be more effective or beneficial. In fact, with much of the validity argument reduced to "required by USED" it is little wonder that there has been an inordinate focus on aspects of validity such as content validity (i.e., alignment to standards) Further, the mere presence of the NCLB mandated tests also made it irresistible for the USED to expand the use of those test beyond school and district accountability to teacher accountability. Gottfredson and Crouse (1986) acknowledged a similar circumstance in discussing the hegemony of the College Board and the SAT.

Lost, however, has been the opportunity to develop a complete and coherent theory of action that describes not only the immediate and higher purposes of administering the test, but also addresses the role of the test and test results within the larger system designed to improve student, teacher, or school performance.

Our closing thought, or question, is what will it take to get to the point that all four facets in the Messick framework are considered appropriately in constructing a strong validity argument? Newton and Shaw (2015) ask whether it matters what we say is part of validity and how we describe validity. As part of their conclusion they state, “the fundamental question is whether

the consequences which result from using the word in one way are manifestly better than the consequences which result from using the word in any other." It seems to us, that is an empirical question. Indeed, it is a validity question. The time has come to validate the concept of validity. What is this construct of validity? What are the value implications and social consequences of portraying validity as a unified whole versus a collection of parts? Perhaps most important, from a utility perspective, which manner of portraying validity is more likely to lead to better validation, better tests, and the higher level purposes of better schools, more effective teaching, or improved student learning and achievement?

In Table 2 we presented opposing perspectives by Shepard and Popham on the impact on the understanding of validity and the validation process of portraying validity in a particular manner. Nearly 20 years later, it should be possible to evaluate, or is it validate, the utility of the positions offered by Shepard and Popham. Which of the two positions has led to a better understanding of validity and/or resulted in better validation of testing programs? Returning to the theme of the conference, one might conclude that Popham was indeed correct. If validity has *gotten short shrift in most testing programs*, perhaps that is because the validity waters have been become too muddy. On the other hand, to the extent that validity is addressed in many testing programs, particularly when addressed by the test contractor, the focus is most likely to be on the *traditional, scientific version of construct validity [that] resides in the upper, left-hand cell* as Shepard warned.

While we are waiting for a study to definitively determine which approach leads to better validity (or that neither does), it is important to acknowledge that the underlying problem is the lack of attention being paid to what Messick (1990) writes to connect his two thoughts quoted previously about the four facets of validity and his unitary concept of validity:

> It becomes clear that distinct aspects of construct validity need to be emphasized, in addition to the general mosaic of evidence, as one moves from appraisal of evidence for the construct interpretation per se, to appraisal of evidence supportive of a rational basis for test use, to appraisal of the value consequences of score interpretation as a basis for action, and finally to appraisal of the social consequences – or, more generally, of the functional worth – of test use. (pp. 25-26)

In particular, with regard to the importance of considering the utility of a proposed test, testing program, or accountability system, Kane (2004) offers this straightforward advice, "[i]f it is not possible to come up with a coherent interpretive argument that gets us from the observed performances to the proposed interpretations and uses, there is no point in proceeding further" (p. 141). That is the essence of utility.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. Washington DC: American Educational Research Association.

Bailey, S.K. & Mosher, E.K. (1968). ESEA The Office of Education Administers a Law. Syracuse, NY: Syracuse University Press.

Bossuyt, P.M.M., Reitsma, J.B., Linnet,K, & Moons, K.G.M. (2012). *Beyond Diagnostic Accuracy: The Clinical Utility of Diagnostic Tests*, Clinical Chemistry 58:12, 1636-1643.

Braun, H. (2008). *Vicissitudes of the Validators*. Presented at the 2008 Reidy Interactive Lecture Series, Portsmouth, NH.

Braun, H. (2012). *Conceptions of Validity: The Private and the Public*, Measurement: Interdisciplinary Research and Perspectives, 10:1-2, 46-49.

Briggs, D.C. (2004). *Comment: Making an Argument for Design Validity Before Interpretive Validity*. Measurement: Interdisciplinary Research and Perspectives, 2:3, 171-191

Center for Surveillance, Epidemiology and Laboratory Services (2010). ACCE Model Process for Evaluating Genetic Tests. Atlanta, GA: Centers for Disease Control and Prevention. Downloaded on October 1, 2015 from www.cdc.gov/genomics/gtesting/ACCE/

Cronbach, L.J. (1971). Test Validation. in R.L. Thorndike (ed.) Educational Measurement, Washington, DC: American Council on Education.

Ebel, R.L. (1961). *Must All Tests Be Valid?,* American Psychologist, 16, 640-647.

Gottfredson, L.S. & Crouse, J. (1986). *Validity versus Utility of Mental tests: Example of the SAT*, Journal of Vocational Behavior, 29, 363-378.

Ho, A. (2014). *Made to be Broken: The Paradox of Student Growth Prediction*. Invited address at the 45th annual conference of the Northeastern Educational Research Association, Trumbull, CT.

Kane, M. (2004). *Certification Testing as an Illustration of Argument-Based Validation*, Measurement: Interdisciplinary Research and Perspectives, 2:3, 135-170.

Kane, M. (2006). Validation. In R.L. Brennan (Ed.) Educational Measurement (4th edition). Westport, CT: American Council on Education and Praeger Publishers.

Newton, P.E. & Shaw, S.D. (2015). *Disagreement over the best way to use the word 'validity' and options for reaching consensus*. Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2015.1037241

Messick, S. (1979). Test Validity and the Ethics of Assessment. Princeton, NJ: Educational Testing Service.

Messick, S. (1989). Validity. In R.L. Linn (Ed.) Educational Measurement (3rd edition). Washington, DC: American Council on Education & National Council on Measurement in Education.

Messick, S. (1990). Validity of Test Interpretation and Use. Princeton, NJ: Educational Testing Service.

Popham, J.W. (1997). *Consequential Validity: Right Concern – Wrong Concept*, Educational Measurement: Issues and Practice, 16:2, 9-13.

Shepard, L.A. (1997) *The Centrality of Test Use and Consequences for Test Validity*. Educational Measurement: Issues and Practice, 16:2, 5-8, 13.

Ujifusa, A. (2015). *Amid Cries of Overtesting, a Crazy Quilt of State Responses*, Education Week, July 8, 2015. Bethesda, MD: Editorial Projects in Education. Downloaded on October 1, 2015 from www.edweek.org/ew/articles/2015/07/08/amid-cries-of-overtesting-a-crazy-quilt.html.

Wainer, H. & Braun, H.I. (1988). Test Validity. Hillsdale, NJ: Lawrence Erlbaum.