Living in a Post-Validity World: Cleaning Up Our Messick

Charles A. DePascale

NERA Presidential Address
Delivered at the 47th Annual Conference of the
Northeastern Educational Research Association
October 27, 2016
Trumbull, Connecticut

There are periods or events in our historical and personal timelines that are turning points; defining moments that in a significant way change all related events that follow. As we approach the presidential election and the events of the last month, one such turning point that comes to mind is the 1960 Kennedy-Nixon debate. The first televised presidential debate, it demonstrated the power of television as a visual medium to impact an election – and the rest, as they say, is history.

Sometimes, forces converge in remarkable ways and multiple defining moments occur within a single year. 1989 was such a year. In November 1989, we watched sections of the Berlin Wall being torn down by jubilant crowds, an iconic moment that signaled the end of the Soviet Union and the Cold War. Just one week later, in my personal timeline, I defended my doctoral dissertation and accepted a full-time job with a testing company. Of course, the following month, Taylor Swift was born in Reading, Pennsylvania. (One of those magical moments whose import and impact are only appreciated in the future.) For our field, however, the most important event of 1989 may have been the publication of the third edition of Educational Measurement with its chapter, Validity, by Samuel Messick.

Validity. Ninety pages that divide people in our field into two categories: people who have never actually completed the entire chapter and people who have pored over it incessantly in search of the deep and hidden meaning. Ninety pages that serve as a defining moment and demarcation in our validity timeline. There are pre-1989 and post-1989 concepts of validity.

Like many defining moments, Messick's Validity did not simply appear out of thin air in 1989. It was the culmination of a decade of intense thinking, debate, and shaping of the concept of validity; a decade that itself was the culmination of several decades of discussion of the meaning, importance, and usefulness of the validity. It could be described as an attempt to produce a Grand Unified Theory of Validity. It was an attempt to produce a theory which combined measurement, social, practical, and political concerns into a single definition of validity; our very own theory of everything.

We are not alone in our search for a unified theory – a theory that elegantly combines things that we partially understand and accounts for things that we do not really understand at all. Most of us, particularly in our field, are hard-wired to want things to fit together nicely. We yearn for causal connections. We want to be able to understand and explain everything. In particular, we want to be

able to explain things that we do not understand and are, in fact, unexplainable. We want to see the face of god and live.

And if measurement (or assessment) is our religion no one can question that we have established validity as our god. Ebel (1961) states "[v]alidity has long been one of the major deities in the pantheon of the psychometrician. Validity is the Alpha and the Omega; literally and figuratively. In the 2014 Standards, the first word of the first chapter, titled Validity, is Validity; and the final standard on page 213 (Standard 13.9) addresses the validity of an overall interpretation. As with any god, however, our ability as mere mortals to understand validity is limited. Even with Messick descending from the mountaintop, and a Shepard to guide us through his work, we have wandered the desert in search of our Promised Land for 27 years. As zealots and heretics clash through their writings and presentations, what are the consequences of their actions? Are we even sure what validity comprises? Can we communicate the meaning of validity to others?

The Standards begin with the following sentence describing the meaning of validity.

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.

The 21 words in that sentence appear to convey a relatively simple and straightforward thought. Why then have scribes devoted their lives to documenting the sacred thoughts and writings on validity? Why then can a civil discussion of validity quickly descend to the depths of sports talk radio, or worse, political discourse?

Well-intentioned educationists have built careers around interpreting the collective wisdom on validity for our children and policy makers. Yet, we are left with quotes such as these:

> Validity is "the most challenging class of the semester" and the hardest for "students to understand." "Despite my best attempts to describe the holy trinity, the unified framework, or argument-based approaches to validity, few students emerge from the class with confidence that they could evaluate validity when developing, using, or even selecting tests." (Gorin)

> Paramount among these imperfections is that the unitary conceptualization of validity as construct validity is extremely difficult to describe to lay audiences. (Sireci, 2007)

It has been 27 years since the publication of Messick's Validity. That means that most of the people in this room have spent our entire professional careers in a post-Validity world. What has been the impact? To paraphrase that great American philosopher, Ronald Reagan,

Are we better off now than we were 27 years ago?

Are we any closer now to understanding and being able to explain validity?

Or more to the point:

Are we building better tests than we were in 1989?

Are we making and promoting better interpretations of tests scores than we were in 1989?

Are we making and promoting better use of tests and testing programs than we were in 1989?

To the extent that we think that 'No' is the appropriate answer to any of those last three questions, to what extent is that outcome an unintended consequence of Validity, and in what ways can a better understanding of validity play in shifting the answer from No to Yes as we move forward.

In the next section of this address, I will outline what I perceive to be seven challenges to validation that currently exist in the field of K-12 assessment and education. All of the challenges are related in some way to Messick's Validity and our efforts to apply validity theory to an ever-changing and expanding universe of uses of testing. All of the challenges impact our ability to develop tests and implement testing programs; and also impact the ability of educators, policy makers, and the public to use assessment effectively to improve education.

Challenges

1. Our concept of validity is built around constructs and we have none.
2. Our treatment of validity creates a false sense of certainty that is inconsistent with our limitations.
3. It is no longer clear where, or whether, measurement fits within the theory.
4. A unified theory of validity masks real distinctions between parts of the field that are separate and should be considered separately.
5. Validity is described as a never-ending process, but our tests have a shelf life.
6. "Validity" becomes an end rather than the means to validity.
7. We may have replaced one holy trinity with another.

Our concept of validity is built around constructs and we have none

By default or design, building our theory of validity around construct validity turns a spotlight on the constructs that we intend to measure or the construct interpretations that we intend to make on the basis of test scores. The 2014 Standards for Educational and Psychological Testing, the *Standards,* define construct as "the concept or characteristic that a test is designed to measure." As examples of constructs "currently used in assessment" the *Standards* offer "mathematics achievement, general cognitive ability, racial identity attitudes, depression, and self-esteem." In addition to demonstrating that the scope of the *Standards* is much broader than K-12 educational testing, the examples suggest a concept of construct that is much narrower and well-defined than virtually all current uses of educational assessment. The assessments and testing programs that I deal with on a daily basis are intended to support inferences about as school quality, teacher effectiveness, college readiness, career

readiness, the politically expedient catch-all college-and-career readiness, and growth. None of those are reflected in the neat examples of constructs included in the *Standards*.

Our struggles with constructs and the influence of those struggles on the way we have framed validity and validation are not new. In 1989, the centerpiece of our field, the SAT, was embroiled in controversy over issues related to interpretation of scores and appropriate uses that are the core of validity. But what construct does the SAT measure? With the SAT, the concepts of test score, interpretation, and use are hopelessly entangled. It would difficult to imagine a worse example of an assessment against which to test this unified theory of validity than the SAT. Well, perhaps one could make an argument for the norm-reference standardized tests, which at the time were in the midst of their own crisis of identity having been called to task by Dr. John J Cannell for what became known as the Lake Wobegon Effect.

For the purposes of this address, we will skip over the 1990s – a decade which began with an "Assessment Spring" but ended with the industry turning the sharp blade of Validity against itself. In short, a decade in which we learned the answer to the question, Does Psychometrics Eat Its Young?

The 1990s led us to No Child Left Behind and the current era of assessment and accountability. Since the adoption of No Child Left Behind in 2002, a primary use of K-12 assessment has been school accountability. Although a test may be designed to measure an individual student's mathematics achievement, as Linn (2009) describes, "a key use of assessment results for NCLB is the identification of schools as either making or not making adequate yearly progress." Linn also explains that implicit in the use of assessments to determine whether a school has made adequate yearly progress (AYP) is the assumption that "the differences in observed school-to-school differences in student achievement are due to differences in school quality" and the resulting inference "that a school that makes adequate yearly progress (AYP) is better or more effective than a school that fails to make AYP."

What, then, is our construct – mathematics achievement, the ability to make adequate yearly progress, school quality/effectiveness? Most of those are clearly not the concept or characteristic that mathematics test was designed to measure. So, we may be able to answer 'Yes' to the question are we building better mathematics tests than we were in 1989. It may also be possible, however, to forge a chain that connects us directly from student achievement in mathematics to adequate yearly progress in mathematics. It becomes more difficult to extend that chain to inferences about school quality or effectiveness?

The use of the same mathematics test scores, or metrics derived from those test scores, as indicators of student achievement, teacher effectiveness, principal effectiveness, and school quality is an example of both the complexity and cloudiness of our constructs. We need to forge a chain linking the very same mathematics test scores to inferences about students, teachers, principals, and schools. Inevitably, that chain will snap as it is pulled in so many directions. When that chain does snap, we are left with the questions what is the construct that we intended to measure and what is our validity argument?

Are the validity arguments that we are building for basing inferences about teacher effectiveness, principal effectiveness, and school quality on student achievement in mathematics any stronger than the argument, "It is so because we said it is so?

Ebel (1961) shares the "story of three baseball umpires who were discussing their modes of operation and defending their integrity as umpires."

> "I call 'em as I see 'em" said the first. The second replied, "I call 'em as they are." The third said, What I call 'em makes 'em what they are."

One would expect that even the most cursory validity study would conclude that defining teacher or school effectiveness in terms of student achievement in mathematics is at best a case of construct under-representation. Is a policy maker or legislator able to wash away validity concerns simply by making the claim fit the evidence?

> *School quality includes more than mathematics achievement, but without mathematics achievement there cannot be school quality.*

Or can policy makers avoid the construct question altogether by returning to our old friend norm-referenced interpretations? When the Obama administration began to issue NCLB waivers in 2012, the accountability focus shifted from AYP and 100% proficiency in all schools to identifying the bottom 5% and 15% of all schools as identified through the accountability system as the schools most in need of assistance. ESSA, the 2015 reauthorization of the 1965 Elementary and Secondary Education Act, reinforces this focus on the lowest performing schools. Of course, the shift to norm-referenced interpretations does not totally eliminate the need to consider constructs; and it also raises the question would I design the same test to identify the bottom 5% of schools as I would to measure the effectiveness of all schools.

Sadly, our success at defining a construct is not much better when we limit our inferences to an individual student's mathematics achievement or further limit our inferences to that student's achievement of the grade 4 Common Core State Standards (CCSS) in mathematics. Studies across textbooks, school curricula, and various assessments "aligned to the CCSS" will contain some areas of overlap, but no consensus of the concept or characteristic of student achievement of the grade 4 CCSS. The result, as Madaus, Russell, and Higgins (2010) describe is that in most cases it is the assessment that defines rather than measures the construct "grade 4 mathematics achievement"; and that is an issue that raises additional challenges.

Our treatment of validity creates a false sense of certainty inconsistent with our limitations

> *The validity of an interpretation cannot be established by a research monograph or detailed manual. The aim for the report is to advance sensible discussion. Why should we wish for more? On matters before the public, the evidence usually is clouded. The institutions of the polity are geared to weigh up reasonable, partly persuasive, disputed arguments; and they can be tolerant when we acknowledge uncertainties. The more we learn, and the franker we are with ourselves and our clientele, the more valid the use of tests will become. (Cronbach, p. 107)*

The description of validity in the *Standards* may, in fact, be consistent with the cautionary conclusion to Cronbach's (1980) *Validity on Parole: How Can We Go Straight?* presented above. The words in the *Standards* and our use of them, however, do not convey the same sense of uncertainty and humility. It

is difficult to draw a direct connection between Cronbach's advice and the testing policies of NCLB and ESSA. It is much easier to do so with the *Standards* and other writings that emphasize the need for assessments, assessment programs, or assessment uses that are valid, reliable, and, now, fair.

We like to present assessment and validity in terms of technical quality in a way that connotes precision and truth. We prefer terms like measurement and reliability to terms such as estimation and probability. As practitioners, we generally do not like to acknowledge the existence alternative outcomes, explanations, or models even to the same extent as local meteorologists. This is not to suggest that we go to the other extreme and only provide a litany of alternatives as equally likely or plausible outcomes. Rather, we should strive to find the level of uncertainty that makes the best use of our professional expertise and allows policy makers and educators to effectively and appropriately integrate evidence from assessments into their decision making.

It is no longer clear where, or whether, measurement fits within the theory.

> *Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the **adequacy** and **appropriateness** of **inferences** and **actions** based on test scores or other modes of assessment. As will be delineated shortly, the term **test score** is used generically here in its broadest sense to mean any observed consistency, not just on tests as ordinarily conceived but on any means of observing or documenting consistent behaviors or attributes. (Validity, p. 13)*

In the beginning, if we viewed tests as measurement instruments and testing as a measurement process, it was easy to portray validity as a measurement issue, or at least a technical issue. From the opening paragraph of Messick's Validity presented above, however, the role of measurement in validity or the relationship between measurement and validity becomes less clear. As we now conceive of tests and testing programs as accountability systems, the connection is even less clear.

We can all accept that we moved on from defining validity and validation as an attempt to answer the question "Does the test measure what it purports to measure?" But, where did we go?

Shepard (1993) offered "Does the test do what it claims to do?" as "a more appropriate question to guide validity investigations." One can easily see the connection between Shepard's question and Braun's (2008) argument regarding the validity of accountability systems:

1. An accountability system is imposed or implemented with the intent that it will accomplish its desired goals (e.g., increased student achievement).
2. The mechanism by which this will happen is called the "theory-of-action".
3. In validating an accountability system, the theory-of-action plays the same role as does the construct in test validation.
4. Thus, consequential validity is the ultimate criterion by which we should judge an accountability system.

From Braun's argument, it is then a very short leap to the argument that Andrew Ho presented to this conference in 2014. Ho argued that in one sense we need not concern ourselves with the inner

workings or technical quality of the accountability system at all. From an experimental design perspective,

1. We can view the accountability system (or testing program) as a treatment that has been implemented with the intended goal of increasing student achievement.
2. We design an experiment to implement the desired treatments or perhaps alternate treatments for a period of time.
3. At the end of the experiment, we compare student learning in the treatment groups to a control group.

Based on the outcome of the experiment, we can judge the evaluative judgment about the impact of the accountability system on student achievement. And if we apply Braun's argument about the validity of accountability systems, we can evaluate the validity of the system based on the results of the experiment.

It is difficult to argue that Ho's experimental approach does not answer the Shepard's question "Does the test do what it claims to do?" But it is difficult to find a connection to measurement and it is difficult to find the sources of validity evidence delineated in the *Standards*.

In reality, however, perhaps the distinctions between Ho's experimental approach and the *Standards*, or an argument-based approach to validity are not that as great as they seem. We have taken a very high-level view of Ho's experiment. When one starts to actually design the experiment and account for all of the relevant variables, perhaps the experimental design process will produce something that looks very much like the collection of evidence that would results from an argument-based approach to validity. And perhaps a close focus on the technical quality of the measures or indicators within the system will result in better outcomes; that is, will result in stronger claims of validity. Nonetheless, it could be interesting to play out the consequences of considering validity from a non-measurement perspective.

A unified theory of validity masks real distinctions between parts of the field that are separate and should be considered separately

The beauty of the unified theory of Validity is that makes it clear that given the complexity of our constructs, one type of validity evidence is likely insufficient to make a strong validity argument. It is likely that evidence related to the construct, criterion, and content will all be necessary to some degree to establish a strong validity argument. To the extent that the construct requires interpretation of the test scores for an intended use, there will also be a need for consequential evidence.

On the other hand, the danger in a unified theory is that by shifting the focus from the various types of evidence to an overarching concept of Validity it may be make it easier for some people to think that unified means simple. No, it is not possible to reach that conclusion if you have read Validity or tried to craft a validity argument or theory-of-action, but who is responsible for those tasks. Who is now responsible for compiling validity evidence?

Shepard (1993) warns of the "danger that test makers will defer to test users to evaluate intended applications" of their tests and that "[o]ften, users are not qualified, or lack the necessary resources, to

conduct validity investigations." She explains that "this separation of responsibility would allow test makers to study only the "scientific" meaning of the test interpretation…leaving it to the user to evaluate the test for its intended purpose" (p. 445). Having had the opportunity to participate in countless meetings with test makers, test users, policy makers, etc. over the last 26 years, I can attest that Shepard was right to be concerned.

The issue, however, may be more complex than Shepard described in 1993. With the rise of customized state assessment, the line between test maker and test user has become much less clear. To what extent is the state department of education the test maker and to what extent are they the test user? In the complex contractual relationships that exist, the testing companies, traditionally considered the test maker, have been contracted to perform specific tasks within the test making process. Is it their responsibility to go beyond those tasks or to point out the need for someone to go beyond those tasks? That is, will they even feel a responsibility to study even what Shepard describes as the scientific meaning of the test interpretation?

There are also questions of conflict of interest in assembling and evaluating validity evidence. George Madaus has been a long-standing advocate for an independent board to evaluate validity of assessments, and the U.S. Department of Education has established a peer review process to evaluate evidence collected by state assessment programs. However, there may still be conflicts or implicit bias in what evidence is collected and/or how it is collected.

The unified theory of Validity, revealed the complex and multi-faceted nature of validity that always existed. Training test user to recognize and deal with that complexity and finding the resources to allow them to do so, however, remains a daunting task.

Validity is described as a never-ending process, but our tests have a shelf life.

> *Inevitably, then, validity is an evolving property and validation is a continuing process. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both the current use of the test and current research to advance understanding of what the test scores mean. (Validity, p. 13)*

From the beginning, one of the anticipated consequences of a unified theory of validity was that it was too big and too complex, effectively absolving individuals from the responsibility of a rigorous process to collect evidence of validity. As Shepard (1993) warned

> Finally, the complexity of Messick's model and chapter creates the same difficulty as nearly every other treatise on construct validity before his. Each emphasizes that construct validation is a never-ending process, because there are so many hypotheses to be tested across so many settings and populations and because the generalizability of findings decays over time. While the never-concluding nature of construct validation is a truism, the sense that the task is insurmountable allow practitioners to think that a little bit of evidence of whatever type will suffice. (p. 429)

Exacerbating the dilemma that validation is a never-ending process is our current view of tests as disposable or ephemeral. Historically, we viewed a test as a product that was carefully constructed over the course of several years and intended to be used for several years as a single form or a set of interchangeable parallel forms. As states shifted to custom development, it was not uncommon for 25%, 50%, of even 100% of the items on the operational test form to be released after each test administration. A new "test" was developed and administered each year. As we move from paper-and-pencil fixed form tests to computer adaptive testing, a test may be a unique event administered only one time to an individual student. How do we reconcile this view of a test with a never-ending process of validation?

The logical answer, of course, is to orient the validation discussion to the testing program rather than to a particular test form. Unfortunately, the life cycle of many state assessment programs has become so short that such a focus gains little.

"Validity" becomes an end rather than the means to validity.

One of the dangers inherent in an unwillingness to accept uncertainty is that we corrupt the concept of validity. In 2014, Michael Kane addressed this conference on the importance of asking the right questions. One of the things that he cautioned against was our tendency to reduce complex situations to questions that we can answer. In validity terms, limiting ourselves to questions that can be answered results in construct underrepresentation. When evaluating a test, we would be quick to identify and point out a lack of alignment between the test items and either the content or cognitive processes that the test is designed to measure. At the same time, when compiling evidence of validity, it is so easy to skim through the list of 100-200 proposed validity analyses and select the studies that can be done quickly and easily (preferably with available data) and avoid those that would require additional resources.

As suggested in the previous discussion of constructs, a more serious threat to validity would be reframing our validity questions until those questions match our available evidence. One could think of this as the equivalent of enhancing reliability at the sake of validity by eliminating test items simply on the basis on low inter-item correlations. In the end, we may have a high coefficient alpha or generalizability coefficient, but in reality, we have neither reliability nor validity.

Alternatively, one could consider the pressure to achieve validity in the same manner that we view high-stakes testing and the pressure to achieve a certain score. When the test score becomes the focus of rather than the knowledge and skills that the score is supposed to reflect, there is a tendency to engage in test preparation activities or test taking practices that are likely to reduce validity. The same can be true for those charged with designing and implementing a validity study.

Note that there is a critical difference between the behaviors described above and reframing or limiting claims about test scores to only those interpretations and uses of test scores which can be supported by evidence. The former reflects an intention to distort validity, while the latter reflects an intention to promote validity.

We may have replaced one holy trinity with another

A primary purpose and accomplishment of Messick's Validity was to solidify the view of validity as a unitary concept. Validity is regarded as "the most fundamental consideration" in developing tests and evaluating the interpretation of test scores for proposed uses. A unified theory of validity, therefore, strengthens educational testing.

To a large extent, however, we never really succeeded in establishing validity as a unitary or preeminent concept with the public and to a certain extent within the field. Whether in textbooks, research papers, or laws, it is rare to see the term valid without its sidekick reliable close by its side. For example, in the Every Student Succeeds Act, the terms valid and reliable appear together in some form at least nineteen times. This invites the question of whether one can have validity without reliability, which may be the intended purpose; and does little to suggest the tradeoffs that are inevitable in considering both validity and reliability. Also, as we have seen on multiple occasions over the last two decades, it may have the unintended consequence of placing an inordinate emphasis on reliability at the expense of validity.

The publication of the 2014 *Standards* further complicates the issue by adding "Fairness" to the mix by isolating Validity, Reliability, and Fairness as the foundations of educational and psychological testing. Therefore, the question now becomes can one have validity without reliability or without fairness? Like Reliability, the *Standards* position Fairness as "a fundamental validity issue" to be considered throughout the testing process.

In practical terms, however, have we simply recreated a Trinitarian model, replacing the holy trinity of "content, criterion, and construct validity" with our new trinity of "Validity, Reliability, and Fairness"? Will the special treatment afforded to Reliability and Fairness enhance or detract from the process of establishing validity?

Where do we go from here?

If those are the challenges that we face, where do we go from here as psychometricians, researchers, academicians, teacher educators and K-12 educators to ensure that we can confidently answer 'Yes' to the three questions I posed at the beginning of the address:

Are we building better tests than we were in 1989?

Are we making and promoting better interpretations of tests scores than we were in 1989?

Are we making and promoting better use of tests and testing programs than we were in 1989?

As a starting point, we must keep our focus on the simple idea expressed at the beginning of the *Standards*:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.

From there, I believe that there are five areas in which we can improve our practice:

1. We need to consider the user
2. We need to build an appreciation for and expectation of uncertainty
3. We need a shared understanding of the basic requirements to call something a construct
4. We need to figure out the role of the test in validity
5. We need to explicitly broaden the conversation about validity from tests and test scores

We need to consider the user

Cronbach (1980) stated "our task is not to judge *for* nonprofessionals but to clarify, so that in their judgments they can use their power perceptively." When communicating information about tests or testing programs we need to consider the users:

a. What do they need to know to be able to decide whether to select a test?
b. What do they need to know to be able to use test information appropriately?
c. What do certain users need to know to be able to build strong validity arguments?

We need to develop solid examples or models of validity arguments for tests or testing programs designed for a particular purpose. These models should not be exhaustive validity plans that account for every possible source of evidence and potential use covered in the *Standards*. They need to be tailored to specific needs of the users and demonstrate acceptable, reasonable validity arguments, not ideal gold standard collections of evidence that are beyond the reach of the typical test user.

We need to build an appreciation for and expectation of uncertainty

Although most of our professional and academic writing is replete with caveats and cautious conclusions, we tend to convey a sense of precision and certainty when we couch our discussions of validity and reliability in measurement jargon. We need to better communicate that validity is a matter of degree and that validation is an ongoing process based on the application of best practices and the collection of evidence. We need to be clear that often evidence will be incomplete, while avoiding the impression that it is impossible to make an informed decision. Virtually everything that we do with tests is built around providing information to inform professional judgment.

Cronbach (1980) correctly points out that "the courts are properly impatient when asked to take seminar-room abstractions as a basis for settling concrete cases." The same is true of policy makers and the general public. All of those parties, however, are able to make informed decisions based on evidence if they know that is what they are being called upon to do.

We need a shared understanding of the basic requirements to call something a construct

We need to be honest in communicating about constructs and our ability to "measure" them. I believe that there are two important factors to consider with regard to what we call constructs and how we describe tests and test scores designed to "measure" those constructs.

First, I believe that the constructs that we are assessing must exist and be observable (at least indirectly) in the real world ; that is, outside of the test. Boorsboom, Mellenbergh, and Van Heerden (2004) offered what they described as an "exceedingly simple" argument defining the "very basic concept" of validity:

> If something does not exist, then one cannot measure it.
>
> If it exists but does not causally produce variations in the outcomes of the measurement procedure, then one is either measuring nothing at all or something different altogether.
>
> Thus, a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure. (p. 1061)

Whether one agrees with Boorsboom et al. that the argument above is sufficient to define validity, I argue that for virtually all of our intended uses of K-12 assessment, we must be able to demonstrate that our tests are measuring something that exists, is observable, and can be described even without the test. That is, there must be other means in addition to the test for identifying students who are proficient in mathematics, teachers who are effective, or high-quality schools. Only if we accept that requirement, then it will be possible to evaluate whether differences in the construct "causally produce variations in the outcomes of the measurement procedure." If we do not require the existence of such external evidence then we are at too great a risk of the test score becoming the construct, which renders our claims and interpretations to be nothing more than tautologies, making validation impossible.

Second, we have to recognize, appreciate, and communicate the complexity of the constructs that we are attempting to measure and our limitations in measuring them. In her 2012 presidential address to this conference, Lynn Shelley shared "The Rat Story" and the description of the complex interactions of personality, social, and emotional characteristics that impact children's nonacademic and academic success; as well as the importance of understanding basic child development and brain development in creating learning environments and interpreting student performance. Mislevy (2016), in an article in the current edition of the Journal of Educational Measurement describes just how complex our constructs could become if we attempt to fully account for all of the forces influencing a student's performance on a complex task measuring higher-order skills. In our claims and in our validation processes we have to find a balance between the inherent complexity and the uselessness of having to create validity arguments that are so conditional that they are so unique that they apply only to an individual student at a given point in time.

We need to figure out the role of the test in validity?

Yes, we have moved beyond the notion that validity can be determined simply by answering the question, "Does the test measure what it purports to measure?" But under what conditions is it important that the test measures what it purports to measure and that we understand what it measures?

Is it simply assumed that the test measuring what it purports to measure was a necessary, but not sufficient component of the validity argument? Are there situations or conditions when that assumption is false?

The first question in my list of three questions asked whether we are building better tests than we were in 1989. On one hand, progress in the development of content standards, performance level descriptors, the use of information from IRT models throughout the development process, and the application of processes such as Evidence-Centered Design would support a clear answer of 'Yes' to that question. On the other hand, the shift in focus to producing student-level scores has on most state assessments reduced the number of questions administered to assess school performance within a content area from several hundred to 40-50.

We have established that an excellent test of student achievement in mathematics might be a poor indicator of school quality. However, we need to be clear about our expectations for that test as a measure of mathematics achievement; and we need to be clear about who is responsible for providing the evidence to support the validity argument related to its use as a measure of mathematics achievement.

We need to explicitly broaden the conversation about validity from tests and test scores

Historically, we have spoken about tests and test scores when discussing validity or validation. Our reality, however, is that most of the validity issues that users are addressing do not involve a simple test and test score. They involve results from a testing program which administers multiple forms of a test over multiple years; and they involve results from accountability systems that comprise multiple indicators which often are metrics derived from one or more test scores. In many cases, the same test score is the basis for multiple indicators. Our language surrounding validity and validation must reflect that reality.

Looking ahead to the very near future, our concept about validity will have to expand to include interpretations, inferences, and decisions based on data mining and learning analytics. As past president April Zenisky demonstrated in her 2015 presidential address the amount of data that we have access to and the uses to which that data can be applied are growing exponentially. We cannot wait for the next revision of the *Standards* to determine how to apply validity to those sources of information.

Conclusion

As I mentioned in the opening of this address, in my personal timeline 1989 was the year that I defended by doctoral dissertation at the University of Minnesota and entered the assessment profession. I think that it is important to note, however, that my education and training at The U established a broad foundation in educational research, program evaluation, as well as theoretical and applied educational measurement. My area of specialization within educational psychology was measurement and evaluation – with a heavy emphasis on program evaluation. My advisor, John Stecklein, was a leader in the field of institutional research. One of my mentors, Stan Deno, worked in the special education department and found me a school-based research assistantship. I believe that background combined with my experiences working with policymakers is what enables me to take a

holistic view of validity while still appreciating the detail of the individual components that validity comprises.

Ebel (1961) had this to say about validity: "It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development." Cronbach titled his address to the 1979 ETS conference on new directions in assessment "Validity on Parole: How Can We Go Straight?" Newton and Shaw (2015) questioned whether validation would be better served by retiring the word validity altogether. I am optimistic, however, that the future of validity is bright. As we move forward, I am confident that if we keep the focus on helping professionals to gather sufficient evidence to support their instructional and policy decisions that validity will be fine.