Guest Editorial

Validity: Meaning and Relevancy in Assessment for Engineering Education Research
Kerrie Anna Douglas and Şenay Purzer
Purdue University

Introduction

Ten years ago marked the beginning of a transformative time for engineering education. The January 2005 special issue of this Journal, “The Art and Science of Engineering Education Research,” was characterized as a “milestone event” in the “process of reinventing” engineering education as a discipline (Streveler & Smith, 2006, p. 103). The issue ignited a conversation regarding how the engineering education community would identify itself and make greater contributions as a research discipline, along with calls for increasing the rigor of fundamental research in engineering education (e.g., Gabriele, 2005; Haghighi, 2005; Kerns, 2005; Fortenberry, 2006; Radcliffe, 2006; Streveler & Smith, 2006; Wormley, 2006).

During the 2005 and 2006 time period of high momentum, researchers identified the development of assessment strategies and high-quality engineering education assessment instruments as integral parts of the community’s transformation (Lohmann, 2005; Olds, Moskal, & Miller, 2005; Radcliffe, 2006; Streveler & Smith, 2006). Perhaps Olds et al. (2005) stated the importance most clearly:

The advancement of engineering education in many ways depends on assessment. High-quality assessments can provide educators with information they can use to move the field forward. Inadequate or poorly constructed assessments can cause educators to pursue ineffective paths, resulting in the loss of time, money, and energy. (p. 13)

In addition to the discussion regarding the development of high-quality assessment instruments, there was a call to conduct research on methods of assessment that may be unique to engineering: “Ultimately, we need to understand the body of knowledge needed by engineering educators to develop, apply, and evaluate assessment methods and tools as well as define the training necessary to accomplish this goal” (NEERC, 2006, p. 261).

Since 2005, the need for research on science, technology, engineering, and mathematics (STEM) assessments in general has increased substantially. The last few years have also brought about reform in science education policy and practice, which has had direct and consequential influences on engineering education. Pellegrino (2012) characterized the environment as “Living in interesting times.” In light of recent changes in educational standards, how to assess what students know and can do has generated much conversation in science education. For example, the Journal of Research in Science Teaching dedicated an entire issue to assessment issues, with an editorial calling for assessment to be the new priority for science education (Songer & Ruiz-Primo, 2012). With the current educational climate in mind, we argue that assessment should also be the priority of the engineering education research community. We
invite others to contribute to the methodological advancements in educational assessment in areas that are unique to engineering.

Exemplars of assessment instruments have been developed specifically for engineering. The Thermal and Transport Science Concept Inventory (TTCI) is used to measure students’ understanding of fundamental concepts in heat transfer, thermodynamics, and fluid mechanics (Streveler et al., 2011). In addition to providing a tool for community use by publishing the TTCI, Streveler et al. (2011) provided an example of how to practically follow assessment guidelines put out by the National Research Council (2001). Streveler et al. combined both qualitative and sophisticated psychometric techniques to develop and study aspects of validity for using TTCI. Another notable assessment instrument is the Teaching Engineering Self-Efficacy Scale (Yoon, Evans, & Strobel, 2014). Yoon, Evans, and Strobel (2014) demonstrated a clear rationale about the areas of engineering self-efficacy covered in the instrument. Informed by a theoretical understanding of self-efficacy, they reviewed how other researchers have measured teacher self-efficacy, and reviewed research on teacher professional development for engineering. Yoon et al. then empirically tested whether the items function as they hypothesized through exploratory and confirmatory factor analysis.

Despite these and other exemplars, discourse within the engineering education research community is lacking about what exactly constitutes high-quality engineering education assessment and how to provide evidence of the meaning and relevance (that is, valid interpretation) of the resulting scores. In order to improve the quality of engineering education assessment, there must be a conversation regarding validity in the context of assessment instrument research and use.

Our goal here is to stimulate conversation within the engineering education research community encompassing what constitutes meaning and relevancy of assessment instruments. We discuss the concept of validity in the context of assessment and then identify common misconceptions regarding validity as well as explain contemporary conceptualizations of the term.

What Is Validity?

Validity derives meaning from the root word valid, which at the very broadest sense means to be “well-grounded or justifiable: being at once relevant and meaningful” (Validity, n.d.). The National Council of Measurement in Education (NCME) defines validity as “the degree to which the evidence obtained through validation supports the score interpretations and uses . . . from a certain test administered to a certain person or group on a specific occasion” (Validation, n.d.). In other words, the heart of validity is the extent to which stakeholders can justify the appropriateness of using an assessment instrument for a specific purpose. Validation is the process of studying evidence to support use and interpretation of an assessment instrument (Wise, 2014; AERA, APA, & NCME, 2014).

The purpose of assessment is to observe “students’ behavior and produce data that can be used to draw reasonable inferences about what students know” (Pellegrino, 2012, p. 833). Within the education literature, terms such as tests, assessments, and measurement instruments are used interchangeably. We prefer the terms assessment instruments or (shortened) instruments to emphasize the point that people assess, instruments do not. Tests, quizzes, and surveys are examples of tools that people use to make assessments.

Engineers readily understand the importance of precise measurement and go through processes to reduce measurement error that are analogous to validation. Yet, this same underlying
drive to reduce “noise” to ensure meaning and relevancy can seem a bit daunting when the instruments are tests or self-report surveys, and the measured variables are educational or psychological. As noted by Wankat, Felder, Smith, and Oreovicz (2002), it is arguably more difficult to conduct educational research where variables are largely uncontrollable and subjective than to conduct research where variables can be clearly identified and manipulated.

Although educational research has many uncontrollable factors, it would be a mistake to infer that measurement error is unavoidable and has no real meaning for research results. Nor is it the case that because educational assessment is inherently subjective, interpretations made from scores are somehow arbitrary or lack empirical evidence. This attitude can be seen in statements such as, “Well, it’s just self-report; it’s not really empirical data.” There are self-report surveys that are poorly constructed and very limited in terms of what can actually be inferred from the results, but there are also very high-quality self-report instruments that can be used diagnostically.

There are varying degrees of assessment instrument quality and how one argues for a particular meaning of the results (that is, the validity of score interpretation; see Kane, 2006) should be based on a chain of reasoning from multiple sources of evidence. For a researcher to assert a particular instrument is validated, the researcher must have gone through the necessary processes of collecting evidence and formulating a rational argument about the instrument’s appropriate use.

**Validity: Misconceptions and Clarifications**

We have observed five misconceptions regarding validity within engineering education research journals and conference proceedings. We briefly state each misconception, discuss what is problematic with the statement, and then offer a clarification of current validity treatment.

**Misconception 1: There are three types of validity: content, construct, and criterion**

Previous conceptualizations of validity asserted that there were three distinct types of validity evidence (content, construct, and criterion), which all assessment instruments should have (Cronbach & Meehl, 1955). In the 1990s, however, the education community began to move away from this trinitarian view and move toward a unitarian view of validity theory in which there are unlimited sources of evidence, all of which is subsumed under construct validity for the purpose of making an evaluative judgment (Messick, 1989, 1990).

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) redefined and have continued to reiterate (AERA, APA, & NCME, 2014) validity as a unifying concept with considerations beyond what is covered in the historical three types. Continuing to refer to three types of validity is not only outdated, but contributes to confusion that validity is something fragmented and quite limited in scope. For example, an instrument might have evidence to claim content validity, yet still not actually assess the intended objectives (NRC, 2001). Further adding to the confusion are otherwise excellent textbooks on educational research methods that have not been updated since the *Standards’* recognition of validity as a unifying concept. Navigating through multiple aspects of validity evidence, determining which type of evidence is needed to support use, and then triangulating evidence are far more complex than simply gathering the three types of evidence and asserting the results as evidence of an instrument being highly validated. Modern approaches to validity consider multiple sources of evidence used to make an evaluative judgment regarding an assessment
instrument’s use, interpretation, and consequences (Messick, 1995; Mislevy, 1994). The evidence collected and rationale to argue for a particular use extend beyond the trinitarian view.

Clarification 1: There are multiple aspects to consider, but only one validity Researchers, educational administrators, and other stakeholders evaluate validity of score interpretations, uses, and consequences of an instrument based on rationale and reasoning from multiple sources of evidence (AERA, APA, & NCME, 2014; Messick, 1994, 1995; Mislevy, 1994). The goal is to collect the most useful and unambiguous information possible to make a decision about the meaning and relevancy of the instrument for the desired purpose (Landy, 1986). Stakeholders use the weight of all the evidence to formulate a chain of reasoning whereby appropriate use of an instrument can be judiciously argued (Kane, 1992). Therefore, validity is an argument for score use, interpretation, and consequences, based on evidentiary reasoning (AERA, APA, & NCME, 2014; Kane, 1992; Messick, 1990, 1994; Mislevy, 1994). In other words, validity is not a property of an instrument but a description of the use, interpretations, and consequences of an instrument. Put more directly, there is no such thing as a valid assessment, only valid uses (Wise, 2014; AERA, APA, NCME, 2014).

Misconception 2: If an assessment instrument is published, it is a valid assessment While any empirical research published in a high-quality journal goes through a peer-review process and must meet certain editorial standards, one cannot assume that any assessment instrument published is automatically of high quality or will function similarly in a different setting. When published and validated assessment instruments are equated, the publication process is treated as a one-time stamp of approval, rather than as a milestone in the developmental process of instrument development (Douglas, Yale, Bennett, Haugan, & Bryan, 2014). By treating the publication process as a stamp of approval, one places all responsibility for interpretation and use of the instrument onto the developers and editorial process and away from the researchers who will use the instrument for their own purposes. This placement of responsibility is the exact opposite of how NCME conceptualizes validation in the statement “Though the burden generally is on the user of test scores to provide the evidence, often the test developer furnishes information” (Validation, n.d.). It is important and expected that developers study evidence of validity for their intended use, but as others use the assessment instrument for their own purposes, evidence should also be given to justify that use. Furthermore, it is not possible to collect all aspects of validity for an instrument to be used for a wide range of purposes, nor is it possible to provide all aspects of validity evidence into one publication. Collecting evidence to inform intended use should be ongoing.

Clarification 2: There is a developmental nature to assessments Validity is never quite over. It is a goal we strive for, but given the nature of educational variables, the process of reevaluating the appropriateness of an instrument’s use is ongoing. In the public debriefing on the new 2014 Standards, Wise (2014) stated, “Validation is an open-ended process, you’re never quite done. . . . Validity evidence should be collected prior to the initial test use and further data analyzed as the test continues in operational use.” This means there is developmental nature to assessment instruments. A developmental perspective of instruments means that evidence is collected both to inform future improvements to the instrument and to provide evidence of appropriate use.

Validation is not a simple open-and-closed process for publication. In reality, very few researchers go through the years of multiple sites, large sample sizes, and iterations needed to have a stable instrument for use in diverse settings. Few have the sizeable external funding that would be required for that level of study. More commonly, researchers have funding for a few years to develop and test an assessment instrument, or the instrument...
development occurs as part of a larger project. A developmental approach would be to collect the evidence needed for intended use during that funding period, and then publish the instrument findings supportive of use, along with limitations and future areas for work. Future researchers would then be free to use the instrument as they deem appropriate and study other aspects of validity to further enhance the quality of the instrument. In this developmental approach, the published instrument continues to be refined by its developers and future collaborators.

When developers assert broadly that their instrument is valid, or has high validity, the developmental process is closed down. There is a clear message, “No further evidence is wanted.” This frequent claim is counterproductive to the advancement of research, where replicability enhances the trustworthiness of results and mixed findings should be further explored to be understood. “Inferences are hypotheses,” writes Messick, “and the validation of inferences is hypothesis testing” (1990, p. 4). Rather than making a claim about the validity of an instrument, it would be helpful for instrument developers to clearly articulate their argument for what are appropriate interpretations of scores, uses, and consequences, as well as their rationale and evidence that supports that argument of use, interpretation, and consequences. Users of the assessment instruments should be encouraged to also consider evidence from their own study contexts to support their use. Validity, as an evaluative argument, is a joint responsibility between the developers and users of an instrument.

An exemplar in the ongoing pursuit of validity, is the work conducted by Yukiko Maeda and So Yoon Yoon regarding the Revised Purdue Spatial Visualization Tests: Visualization of Rotations (Revised PSVT:R; Maeda, Yoon, Kim-Kang, & Imbrie, 2013; Maeda & Yoon, 2013). First, Maeda and Yoon conducted a study to examine the psychometric properties of the original PSVT:R (Guay, 1976; Bodner & Guay, 1997); then they suggested revisions to the instrument based on evidence (Maeda et al., 2013) and continue to examine different aspects of validity to support appropriate use of the Revised PSVT:R with engineering students (Maeda & Yoon, 2013; Maeda & Yoon, in-progress). In each article, Maeda and Yoon have identified both potential threats to validity and future research for the instrument. Readers can see clearly the strengths and limitations of the instrument and what future work needs to be done to improve its use.

**Misconception 3: Psychometrics = validity** Another common error is putting too much emphasis on psychometrics alone for justification of validity. This over-reliance on psychometrics happens when an article’s discussion emphasizes procedures, such as factor analysis or applications of item response theory, without giving equal consideration of other vital pieces of information, such as construct representativeness, instructional sensitivity, or theoretical bases. Such articles frequently lack a clear discussion about what hypotheses are being tested through the psychometric procedures and what can be directly inferred from the results. Why is this evidence collected? What other evidence of validity was collected before a large-scale data collection? How well do the study participants match those the scale was intended for? These threats to validity could be addressed through many procedures, in addition to psychometrics. Multiple sources of evidence can be collected as part of a chain of reasoning that is used to argue valid interpretations (Kane, 1992). One application of this argument-based approach is the Evidence Centered Design method of instrument development (Mislevy, Steinberg, & Almond, 2003).

**Clarification 3: Validation starts long before any data are collected or analyzed** In the development stage, it is important to consider the most common threats to validity, that is, what issues commonly jeopardize an instrument’s quality. Messick (1995) pointed out the two biggest threats: (1) Not actually representing what the instrument is designed to measure, known as
construct underrepresentation, happens when a name is given to an instrument, but there is inadequate evidence that the group of items can be labeled and used to represent an unobservable variable. (2) Construct irrelevant variance, when something other than the actual measured trait is influencing results, occurs when a person’s performance on an instrument is influenced by something other than his or her actual skill or attribute. Poor wording of the questions or use of language unfamiliar to certain groups are examples of how construct irrelevant variance is created. However, at times, the sources of irrelevant variance can be difficult to determine.

There are three foundational aspects to all assessment instruments (NRC, 2001): (1) cognition – theory or understanding of cognition and learning related to the content domain; (2) observations – the type of tasks students are asked to do in the assessment based on a set of principles; and (3) interpretation – the methods and tools used to determine how to interpret the results. These three aspects are referred to as the assessment triangle (NRC, 2001). Qualitative studies can be designed for the purpose of instrument development, to inform construct representativeness, and to learn useful information on potential sources of construct irrelevant variance. The assessment triangle ensures appropriate content is covered in an assessment instrument. In the case of assessing learning-related outcomes (for example, concept inventories or knowledge tests), the development of the instruments should be guided by first understanding how students learn and what tasks demonstrate student learning (Streveler et al., 2011). In the case of noncognitive assessment, such as of self-efficacy, identity, or self-directed learning, etc., the assessment triangle still has utility, with slight modification. The cognition corner of the assessment triangle could be replaced with the theory or understanding of what is being assessed, and the observation corner would be how this attribute is manifested. The interpretation corner includes studies of psychometric evidence to empirically investigate the instrument, but equally important are theories behind what is being measured and what type of tasks or questions should be asked. The psychometric evidence is interpreted in light of the theoretical framework underlying the instrument and the principles of how the attribute is expressed. An example of this approach is the work conducted by Ahn, Cox, London, Cekic, and Zhu (2014), who utilized a mixed methods approach to develop a measure for engineering leadership, change, and synthesis.

In terms of construct irrelevant variance (that is, extraneous, unintended variables that affect assessment outcomes), cognitive interviews and think-aloud protocols can provide evidence that students read and understand the questions in the way the creators intended. Issues of fairness can be considered. For example, in the creation of the Hopes and Goals Survey for STEM education (Douglas & Strobel, 2014), third-grade students in a high-minority, low-resource, low-performing urban school district read each statement out loud and explained how they chose their answers. Having evidence that students read and understand both items and answer choices as intended by the developers provides insight into sources of irrelevant variance and informs revision. Without this qualitative evidence, it is hard to distinguish whether some students answer incorrectly due to not understanding certain words instead of not understanding the content area being targeted by the question. Considering the sizeable enrollment of international students in U.S. engineering programs, evidence is needed to support the rationale that all students read and understand the instrument in the same way.

**Misconception 4: Validation does not require theory** In terms of validity of interpretations for research purposes, this misconception is problematic in two ways: there is an assumption that it is acceptable to assume a group of items represents some general idea, and it is possible to make an interpretation without a theoretical basis. For example, what is the significance of students having positive or negative attitudes towards a program? This confusion about
the importance of theory in noncognitive assessments likely arises from difficulty of identifying information needed for an evaluative or practical purposes and that needed for making research contributions with a theoretical value.

**Clarification 4: Theory is essential in the validation process** As previously mentioned, the process of validation is hypothesis testing based on theory. As with all scientific study, engineering education research studies should be grounded in theory. At a minimum, all researchers involved in assessment instrument development should specify a theory, employ methods to measure the theoretical constructs, and empirically investigate how well the items function as hypothesized (Cronbach & Meehl, 1955). Theory informs the questions we ask and how we interpret the findings. Hence, researchers start the process of assessment by identifying the precise domain and scope being studied. For assessment instruments to be used for broad–level research and of community-wide interest, there should be a theoretically based rationale for the selection and use of an instrument. Unless there is a theoretical reason for why constructs are being measured, the contributions to research would be quite limited.

Recent advances in assessment continue to make theory central to assessment. For example, the NRC (2001) recommendations of the assessment triangle expand the previous recommendations by Cronbach and Meehl (1955) by explicitly stating the integration of cognitive theory, principles regarding how students represent their knowledge, and the inferences from scores. When we align cognition, observations, and interpretations, theory is not only informing our assessment research, but also being influenced by the findings. It is imperative that we align theory, instrument design, and the statistics used in order to test how well the scores can be used as intended and whether the instrument measures constructs such as student learning or noncognitive attributes such as identity, self-efficacy, or leadership skills.

**Misconception 5: Cronbach’s coefficient alpha is always or never appropriate** A common misunderstanding is that coefficient alpha is an essential calculation for all instruments. It is simple to calculate and, at times, is abused to claim validity. Having a high alpha value does not equate to validity; it means that items have a high degree of internal consistency. An instrument can have a high coefficient alpha and still be measuring something that differs from what it is intended for. An analogous example would be a bathroom scale that consistently measures someone 10 pounds heavier than that person’s actual weight. The number shown on the scale is not a valid representation of the person’s weight, even though it is consistent in its measurements. Similarly, it is entirely possible for an instrument to have a high alpha level and still not have adequate information to justify valid uses, inferences, or consequences.

Another misunderstanding is to assert coefficient alpha is not appropriate in situations where groups of items are being scored together as a single category. One reason for this misunderstanding is that the number of items does contribute to the calculation. However, variance of the item and variance of the overall instrument (or section of the instrument) also contribute to alpha calculation.

Educational measurement experts have extensively discussed the appropriate use of the coefficient alpha as a gauge for internal consistency and ways to increase consistency (Cortina, 1993; Schmitt, 1996; Weng, 2004). While there is considerable literature concerning reliability, the central purpose of coefficient alpha is to empirically test whether items that are written to reflect a basic concept are consistent with each other. Before an instrument can have valid uses, it must first measure consistently.

**Clarification 5: Reliability is not about validity but about the degree of consistency** While reliability is a critical aspect of validity, the method of evaluating reliability should be aligned with the purpose of the instrument. Calculating the coefficient alpha is not the only option, although
it can be used as an important first step in examining the dimensionality of an instrument (that is, determining whether a specific construct or a narrow set of related constructs is measured). There are, however, important reliability questions beyond what the coefficient alpha can calculate. For example, if a student takes an instrument multiple times, would the scores have a high correlation each time? If the same student takes two different versions of an instrument, would the versions have similar or correlated results?

We offer the following three rules of thumb, based on the principles and discussion on reliability found in Kubiszyn and Borich (2007): (1) When internal reliability is calculated for assessment instruments that contain items not representative of another variable and not necessarily grouped into categories to be scored, there is no value added by knowing the coefficient alpha level. For example, consider a final exam in a course on introduction to materials engineering that covers a range of topics including atomic structure, phase transformations, and material selection for a given design. In such a situation, low internal reliability may lead erroneously to the conclusion that the instrument is deficient. (2) Calculate item means first to determine the average scores. Very easy or very difficult tests tend to have biased internal consistency because of the nature of the difficulty, not the nature of what is measured. (3) Similar to the difficulty issue, it is important to consider the diversity of those assessed. Groups with a wide range of skill levels tend to produce lower levels of internal consistency, whereas homogenous groups tend to produce higher levels of internal consistency. For example, scores from an honors course would likely produce a higher coefficient alpha than scores from the same instrument used in a general course.

Conclusion

While much of what we provide here could be referred to as just sound practices for development and use of assessment instruments, the engineering education research community has needed a conversation at this level. Exemplars in assessment instrument development can be found, yet more researchers need to involve themselves in the laborious work of assessment. Furthermore, engineering education researchers must be leaders in research to create methods of assessment that can be used to measure constructs essential or valued in engineering education. Such leadership involves collaboration with those who have formal training in assessment and the learning sciences.

It is an exciting time for engineering education research. Yet, to capitalize on the opportunities in front of us, validity must be the concern of the whole community.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grants DUE-1245998 and EHR-1348547. We thank Dr. So Yoon Yoon for providing feedback on earlier drafts of this editorial. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


**Authors**

Kerrie Anna Douglas is a visiting assistant professor in the School of Engineering Education at Purdue University, 516 Northwestern Avenue, Wang Hall 4541, West Lafayette, IN 47906; douglask@purdue.edu.

Senay Purzer is an assistant professor in the School of Engineering Education at Purdue University, 516 Northwestern Avenue, Wang Hall 4545, West Lafayette, IN 47906; spurzer@purdue.edu.