International Journal of Engineering Education Vol. 32, No. 5(A), pp. 1960–1971, 2016
Printed in Great Britain

0949-149X/91 $3.00+0.00

Reliability, Validity, and Fairness: A Content Analysis of Assessment Development Publications in Major Engineering Education Journals*

KERRIE A. DOUGLAS, ANASTASIA RYNEARSON and ŞENAY PURZER
School of Engineering Education, Purdue University, Seng Liang Wang Hall, 516 Northwestern Ave., West Lafayette, IN 47906, USA.
E-mail: douglask@purdue.edu, anastasia.rynearson@gmail.com, purzer@purdue.edu

JOHANNES STROBEL
School of Information Science & Learning Technologies, University of Missouri, 211B Townsend Hall, Columbia MO 65211, USA.
E-mail: strobelj@missouri.edu

After more than a decade of efforts to enhance the quality of engineering education research, including assessment development, it is timely to explore what types of validity evidence are frequently reported in assessment research articles. According to the *Standards for Educational and Psychological Testing*, the foundation of quality assessment rests on evidence of reliability, validity, and fairness. The purpose of this study was to explore what aspects of reliability, validity, and fairness evidence are provided in assessment instrument development publications in major engineering education journals since 2005. Using quantitative content analysis, the authors reviewed twenty-nine articles published in four major engineering education journals between 2005 and 2015. A coding scheme, based on Messick's Unified Theory of Validity and the *Standards for Educational and Psychological Testing* was developed to code the aspects of reliability, validity, and fairness provided in each article. Frequencies for each code are reported. Engineering education articles on instrument development most frequently reported evidence related to aspects of internal reliability, content-related validity, and substantive aspects of validity. However, studies of generalizability, consequences, and fairness were largely void. In addition, reliability was most frequently studied through internal reliability coefficients, while other forms, such as test-retest were less frequently reported.

Keywords: validity; instrument development; assessment; measurement; reliability; fairness

1. Introduction

Scientific progress is in many ways determined by the accuracy and precision of the measurements used. Just as poor instruments can lead to inaccurate readings and false results, assessment instruments that are poorly constructed or incorrectly applied can lead to erroneous findings and fallacious conclusions for engineering education researchers, instructors, and administrators. Over the last ten years there have been large shifts towards higher rigor and calls for increased quality of research in engineering education, including assessment research. In tandem with conversations at engineering education conferences and meetings, then IJEE editor, Michael Wald explicitly positioned the research journal by stating the role of the journal is to "promote pioneering and research based ideas for the future of engineering education" [1, p. 1]. Similarly, the *Journal of Engineering Education* published a special issue devoted to raising the bar for research in engineering education including assessment [2]. In subsequent issues, several guest editorials and articles reinforced the need to improve engineering education through high-quality research and reiterated that quality assessment instruments are crucial to the advancement of engineering education as a field [3, 4]. Recently, Douglas and Purzer revisited the 2005–2006 calls with a focus on validity and quality assessment in engineering education research and argued the need to once again foster conversation about assessment within the engineering education research community [5].

In parallel with reforms in engineering education, assessment experts have challenged researchers to "rethink" common assessment development practices in order to more clearly align what is measured with advances in the learning sciences [6, 7]. In the United States, the National Academies of Sciences, Engineering, and Medicine along with the National Science Foundation held a Symposium on Assessing Hard to Measure Cognitive, Intrapersonal, and Interpersonal Competencies in December 2015. A major theme of the symposium was the argument that quality assessment can be a means for education reform [8]. Indeed, assessment can be a very powerful tool in advancing engineering education as how and what is assessed largely influence both teaching and learning.

The purpose of this research is to examine what aspects of quality assessment, defined by the *Stan-*

* Accepted 28 May 2016.

dards for Educational and Psychological Testing [9], are most frequently reported in engineering education research since the 2005 call for increased rigor in assessment [2]. It must be emphasized that validity is not a checklist of tasks to complete or a set list of evidence every assessment instrument must have [10]. The evidence required is dependent upon use; where the more important uses require higher levels of evidence. Furthermore, it is not possible, nor desirable, to publish in one manuscript every aspect of validity sought for an instrument. The chief aim of this study is to foster conversation regarding what types of evidence are commonly reported, where there are gaps, and what these findings mean for use.

Scale development texts [e.g., 11] define the steps for assessment instrument development based on general use cases. The focus in this research is to examine what aspects of reliability, validity, and fairness are commonly reported in top-tier engineering education journals' assessment publications. We also discuss what test results mean for use.

2. Literature review

2.1 What makes an assessment instrument high quality?

With each new high quality assessment instrument, the engineering education community adds another tool to its repertoire, an activity that is critical for advancing large-scale research. In the last decade, several authors have reported the development and validation studies of instruments specific to assessment in engineering education contexts [12–14] including a review of methods used in entrepreneurial engineering [15]. Additionally, Jorion and colleagues recently created a framework for evaluating the claims of concept inventories as an aid to potential users [16]. While having a framework for individual users to apply when selecting a concept inventory is of practical importance, more broadly, there is also a need for common understanding about what type of information justifies an assessment instrument as being of high-quality for use in engineering education research.

Unlike traditional fields of engineering, researchers in engineering education depend on measurement tools that are not fully objective. Yet, standards regarding how to choose, apply, and assess the quality of an instrument for a specific use is critical in every field of engineering and education. One would not choose an ohmmeter to test the mass of an object; neither would one use a scale accurate to one gram when precision to one-tenth of a gram is necessary. Furthermore, instruments must be calibrated before their results can be considered valid for a study. Both in engineering and educational measurement, there is a rationale for how an instrument is expected to perform in certain situations and then evidence is collected to test that functioning. In essence, validity is an evaluation framework. Put another way, "validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use," [17, p. 742]. Kane [18] articulated this process as argumentation; where sources of evidence to support use are determined and tested. In recent years, the relationship between what evidence is collected and the use of the instrument continues to be central. High quality assessment instruments have alignment between evidence (i.e., what evidence is collected and how that evidence is obtained) and intended use of the instrument. Conceptualizations of validity as an argument for use continue to deepen, with Evidence-Centered Design [19] as an example of an approach to developing and testing assessment instruments in terms of evidentiary arguments.

A high quality assessment instrument can take five, ten, or even twenty years to develop. During that time, there may be appropriate uses for the instrument while developers continue to work toward a more precise measure. Validity is not a dichotomous variable of all or nothing; it is a matter of degree [20]. With this in mind, there is inherently a developmental nature to assessment instruments. Careful selection and understanding of appropriate use of the assessment instruments are crucial to ensure accurate conclusions from any study. Therefore, common understanding of what constitutes high quality assessment in engineering education is needed. According to the 2014 *Standards*, the cornerstones of quality assessment are reliability, validity, and fairness [9].

Validity does not reside only in the hands of the assessment instrument developers. Rather, it is the user of an assessment instrument who holds primary responsibility for providing evidence the instrument is used in a valid way [9]. As a community that openly shares assessment instruments, we have a joint responsibility to establish evaluative norms regarding what aspects of validity substantiate high quality assessment for use in research, educational evaluation, and instructional support. From this perspective, it is worthwhile for the engineering education research community to have a conversation about what publication norms would support appropriate use of assessments in engineering education research. In other fields, such as chemistry education, counselor education, and industrial organization psychology, reviews of common practices in instrument development have been published in major journals [21–23]. These works

serve to foster the discussion regarding acceptable evidence needed to deem an assessment instrument as high quality within a research community.

The *Standards* [9] emphasize that the higher the stakes associated with the use of an assessment, the higher level of evidence is required. Within engineering education broadly, there are five general reasons an assessment might be used: (1) assessment of student learning in a specific course (both formative, summative), (2) educational evaluation (i.e., program level), (3) educational research, (4) instructional support, and (5) admission decisions. Within each broad category of purpose, there are variances in the stakes involved. For example, a quiz does not contribute to a students' course grade as much as a final exam. Some measures used for program accreditation may be weighted more heavily than others. In addition, some assessments may be used dually — for example, a materials science lab report was used to assess student learning and also informed curricular revisions to the tensile testing simulation lab [24]. It is important that educators are aware of specific and multiple uses of such data and able to distinguish appropriate uses of an assessment instrument and the data it provides.

3. Theoretical framework

Our approach to validity comes from Messick's Unified Theory of Validity [17]. The specific aspects of validity continue to evolve and the role of evidence has become more central in recent times [19]. The American Educational Research Association, American Psychological Association, and National Council on Measurement in Education joint committee discusses and translates theory and research into the *Standards for Educational and Psychological Testing* [9]. The *Standards* are intended to guide sound assessment research and as a resource to evaluate assessment practices [9].

Based on the *Standards* conceptualizations, high quality assessment instruments are very specific in purpose, have evidence of consistency (i.e., reliability), a clear argument for use based on rationale and evidence, and evidence of fairness. While there is not an exhaustive list of sources of evidence (as researchers will continue to develop new methods) the *Standards* provide some guidance on what types of evidence can be collected to argue reliability, validity, and fairness. A brief overview of each cornerstone of high quality assessment is provided.

3.1 Reliability

Reliability is the degree to which the instrument can be trusted as consistent. There are many ways to assess reliability; the form of reliability examined is dependent upon the relevant sources of variance [25] (See Table 1). Based on Classical Test Theory, there are three traditional categories of reliability coefficients: (1) alternate-form coefficients: correlations derived from administering alternative forms of a test are used to evaluate error based on the sample of items and potential of the assessment instrument to generalize to a broader domain; (2) internal consistency coefficients: correlations based on the scores of individual items for subsets of the assessment instrument or total score which takes into account the variance attributable to subjects and interaction between subjects and items; (3) test-retest coefficients: correlations between scores for the same assessment instrument administered to the same person/group at differing times, used to consider error factors associated with time lapse [26, 27]. In addition, inter-rater reliability is often used in cases of open-ended responses to determine the

Table 1. Common Types of Reliability

Types of Reliability	What it Evaluates	Questions Asked	Examples of Evidence
Alternate-Form	Consistency between different versions of same test	Are different versions of a test interchangeable?	Reliability coefficient, coefficient of stability
Internal Consistency	Statistical interrelationship between responses to items	Do items written as theoretically related show interrelationships?	Coefficient alpha, Kuder-Richardson Formula 20, 21, Split-halves
Test-Retest	Consistency/correlation between scores over time (with no intervention)	How similar are responses at different administrations?	Coefficient of stability
Item Response Theory/ Precision	Items' ability to differentiate among persons	How consistently does the item measure persons based on ability?	Item information functions
	Variation in observed score due to measurement error.	How precisely is the instrument measuring?	Standard error of measurement
	Systematic and unsystematic sources of error variation.	What are the sources of measurement error?	Generalizability coefficient, Dependability Index

consistency between ratings of more than one person.

Item Response Theory (IRT) treats reliability more broadly as precision. Under the IRT model, reliability can be assessed in many ways, including standard errors, reliability coefficients, item information functions, and generalizability coefficient [9, 28]. Reliability is essential because consistency of scores reduces measurement error.

3.2 Historical and contemporary perspectives of validity

The historical view of assessment development notes that three distinct types of validity should be evidenced: content-related, criterion-related, and construct-related [29]. While three types of validity are still commonly referred to and found in measurement textbooks, current conceptualizations of validity are far more comprehensive. One drawback to the historical model of validity is the overemphasis on statistical procedures and lack of explicit attention to the foundations of assessment, evidence that the assessment items are accurately representing a true competency.

In the late 1980's and early 90's, validity was reconceptualized as a unified theory where validity is reasoned from evidence [5]. Messick proposed a comprehensive model for construct validity, where all other sources of validity information were subsumed (and expanded to include additional aspects) as part of construct validity in a mosaic of evidence [17, 20, 30, 31]. For example, content, criteria, and consequential aspects of validity were integrated into a construct validity framework used to test hypotheses about score meaning and use. Validity, however, does not require any one form of evidence, nor can it be asserted based on just one type of evidence [32]. Messick stated, "What ***is*** required is a compelling argument that the available evidence justifies the test interpretation and use, even though some pertinent evidence had to be forgone. Hence, validity becomes a unified concept, and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity," [17, p. 744].

The most recent version of the *Standards* [9] further emphasizes validity as the integration of evidence and its key role in assessment. The Unified Theory of Validity framework lays out seven functional aspects of validity, used to gather evidence and inform use of test scores: (4) generalizability, (5) external, (6) fairness, and (7) consequential" to read as "(4) external, (5) generalizability, (6) consequential, and (7) fairness. These aspects of validity identified by Messick expand the types of evidence provided by the original three types (content, criterion, and construct) and integrate the evidence to make an argument about specific use of an assessment. Table 2 lists each aspect of validity, then describes what the aspect is concerned with, the type of question that would lead to its' study, and

Table 2. Description of Aspects of Validity, Uses, and Evidence

Aspect of Validity	What it Evaluates	Types of Questions Asked	Examples of Evidence
Content	Technical quality, relevance, and content representativeness, face validity/appearance*	How well does the table of specifications or blueprint match the intended purpose of the assessment? What is the level of alignment between test objectives and actual items?	Expert review, correlation with similar assessments, item difficulty and discrimination
Substantive	Respondents engage with, read, and understand the assessment items as intended	Is the group of interest interpreting the items as intended? Are the cognitive processes the test is designed to measure being assessed?	Verbal protocol analysis, observations, semi-structured interviews
Structural	Fidelity of scoring structure. Items can be summed together in a scale and labeled as a single construct.	Is the internal structure of the instrument congruent with the structure of the construct domain?	Factor Analysis, Item Response Theory scaling procedures
External	Scores are convergent or discriminate with other variables as hypothesized.	Do the scores correlate with other variables as expected, either convergent or discriminant?	Correlational studies with external variables
Generalizability	Extent to which technical qualities of instrument generalize to a group, across groups, tasks, and contexts	Can the scale be generalized to other situations under which it will be used?	Meta-analysis, Generalizability theory techniques
Consequential	Potential and social implications of using the results are in alignment with purpose and ethical	What is the evidence that the consequences of the test scores are justifiable? Who will determine the usage of the test scores?	Follow-up studies of use cases

Note. * Face validity is also often referred to as an aspect of validity, we include it here as part of the content.

some examples of types of evidence collected to evaluate that aspect. Faireness is discussed in the following section.

3.3 Fairness

One of the major updates in the newest version of the *Standards* is that fairness has been elevated to the same importance as reliability and validity [9]. During the Public Briefing, Barbara Plake stated, "We fundamentally believe that fairness is one of the major foundational constructs that need to be attended to in order for a test to be of high quality" [33]. The *Standards* articulate that fairness includes: (1) providing access for all examinees in the intended population, (2) identifying and removing irrelevant sources of performance, and (3) supporting appropriate reporting of results. There are many statistical methods for identifying and removing potential bias from the test scores and items. These are important and necessary, however, fairness is a very complex issue that extends beyond statistical approaches to studying test bias [6].

Any conversation about fair assessment in the context of engineering education must consider the historical and current reality that the majority of engineers are white males. Despite years of calls and initiatives for diversity, it is no ground-breaking news to report that in 2013, women made up 18.6% of engineering undergraduates in the United States; Hispanics 9.9%, Blacks 5.1%, and Native Americans comprised less than 1% [34]. The American Society of Engineering Education identified diversifying engineering as a core value and designated 2014–2015 as the Year of Action on Diversity [35]. The issue of diversity also involves language, where tests may be taken in a language that is not native to the test taker. Certainly, instruments intended as measures of engineering competencies should have some evidence of measurement invariance across groups and a rationale of how to use the assessment fairly to measure diverse groups. Furthermore, as assessments of engineering learning are used in high-stakes situations, issues such as fairness related to opportunities to learn content and opportunities to take assessments must be considered [6]. In addition, consequences should be evaluated from the perspective of fairness.

4. Methods

To review how reliability, validity, and fairness of assessment use is argued in high-quality engineering education journals, a content analysis was conducted [36]. Whereas a systematic review is a search and synthesis over multiple databases [37], the authors purposefully chose to only synthesize the highest-ranking journals specific to engineering education, as it is understood the evidence provided in lower ranked journals would not require the same level of validity evidence as a top-tier journal. The final body of research articles was analyzed using a structured coding protocol [36]. We specifically examined how arguments for reliability, validity, and fairness were made in the sample pool of research articles. Exemplars for aspects of reliability, validity, and fairness were noted and are given in the results.

4.1 Identifying articles

The researchers conducted a purposeful sampling strategy for the selection of journals [36, 38]. We referred to Van Epps' holistic analysis of journals in engineering education [39]. Van Epps identified the following as the top four engineering education specific journals: (1) *Journal of Engineering Education*, (2) the *International Journal of Engineering Education*, (3) *European Journal of Engineering Education*, and (4) *IEEE Transactions on Education* [39].

The Scopus database, which houses articles published in all four journals, was used to search within the journals, with the keywords "(scale OR instrument OR survey OR "concept inventory") AND development" in the title, keywords, and abstracts. The search was limited to articles published since 2005 when many calls were published for an increase in the rigor of engineering education research. This search process resulted in 257 articles published between January 2005 and December 2015. Next, all abstracts were read and those that discussed instrument development specifically or inferred development by discussing validity or reliability were considered for analysis. Forty-two articles remained for analysis. Four researchers read these articles in full. After reading the articles in full, 15 articles did not discuss issues of reliability, validity, or fairness. The resulting dataset of 29 articles are listed in the Appendix. Three assessment development articles were published in the *IEEE Transactions on Education,* six articles were published in the *International Journal of Engineering Education (IJEE)*, and 20 were published in the *Journal of Engineering Education (JEE)*. Assessment validation publications were not found in the *European Journal of Engineering Education (EJEE)*.

4.2 Data analysis

From a deductive approach Messick's presentation of the Unified Theory of Validity [17] and the *Standards* [9] were used to create a coding scheme. For reliability, Table 1 column *Types of Reliability* was used as the coding scheme. For validity, Table 2 presenting *Aspects of Validity* was used as the coding scheme. For fairness, articles were coded if

they included an explicit discussion of fairness or bias, whether addressed through item development or statistical procedure.

Two researchers independently coded each article based on evidence of reliability, validity, and fairness. Each article was then reviewed for consistency in how it had been coded. When differences were found, definitions based on Messick [17] and the *Standards* [9] were used to guide the final coding decision and researchers came to agreement. Articles were coded based on evidence provided; not according to the quality or quantity of the scores or evidence. The purpose was not to evaluate authors' work (which has already been done through editorial and peer review), but rather to examine the types of reliability, validity, and fairness evidence reported in instrument development articles in engineering education journals and provide recommendations for common practice. For example, if an article reported the findings of an Exploratory Factor Analysis, it was coded as having evidence of structural validity even if later in the article the authors concluded that the factor structure was less than ideal. For another example, generalizability can refer to a range of differing aspects as discussed earlier, this code was applied if the authors presented an argument or rationale that their results would generalize in some way beyond the sample the instrument was tested in.

5. Results

5.1 What type of reliability evidence is commonly reported?

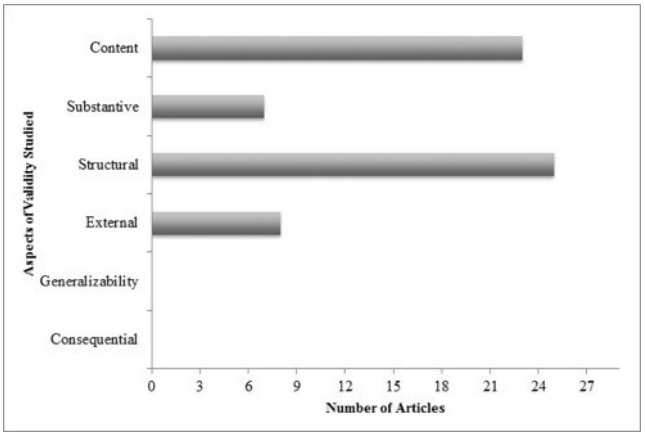
Ninety-three percent (27 out of 29) of the journal articles reviewed reported types of internal consistency through Cronbach's alpha or Kuder-Richardson Formula 20 (KR-20). One article of the 27 [40], reported inter-rater reliability and test-retest reliability. No other articles in the sample reported test-retest reliability or alternate (parallel) form reliability. Forms of reliability under Item Response Theory were not found.

5.2 What evidence of validity is reported?

The areas of content and structural validity are the most frequently evidenced aspects of validity in engineering education assessment instrument publications. Less routinely reported are evidence of substantive, generalizability, external, and consequential aspects of validity. Fig. 1 shows the number of articles that evidenced each aspect of validity. The two aspects of validity that were reported the most frequently are content and structural. Seventy-nine percent (23 out of 29) of the articles provided evidence of content aspects of validity. Evidence included processes such as identification of scope of domain through qualitative research and expert review (often referred to as face validity). Twenty-four percent (seven) of the articles provided evidence of the substantive aspects of validity through processes such as the assessment triangle and Item Response Theory. Eighty-six percent (25) of the articles provided evidence of structural aspects of validity, through Factor Analysis (20) and Item Response Theory (five). Twenty-eight percent (eight) of the articles provided evidence of external aspects of validity, such as convergent or discriminant correlations with other variables. One article provided evidence regarding generalizability of findings. No article explicitly studied consequential aspects of validity.

5.3 What evidence of fairness is reported?

One article (3%) examined potential bias in the assessment instrument based on gender and demographic information. Of note, another article stated



Number of Articles

Fig. 1. Articles Providing Validity Evidence.

the need for examination of bias as an area for future research. No other explicit mentions of bias or fairness were found in the articles.

6. Discussion

6.1 Reliability, validity and fairness

Reliability, or consistency in results, is commonly reported in engineering education as internal consistency measured by Cronbach's alpha or Kuder-Richardson Formula 20. Nearly all researchers calculated these statistics when presenting their instrument analysis. Reliability over time (test-retest reliability) is rarely reported, and consistency between similar instruments measuring the same constructs (alternate or parallel form reliability) was not reported in this sample. Internal consistency of instruments is valued as the mark of reliability by the engineering education community, though a complete picture of consistency including over time or across instruments is not commonly reported in engineering education research at this time. Test-retest reliability information would be helpful for use of instruments multiple times in a study. It would also be helpful for studies to show how stable the traits under investigation are. IRT methods of reliability would provide greater confidence in the precision of the assessment. By paying greater attention to additional forms of reliability, beyond internal consistency, potential users would be more informed about the consistency and precision of scores.

Validity is an argument made based on evidence for a purpose, not simply a checklist of tasks to be done. Researchers should take care when using the words, *valid*, *validated*, and *validity*. As stated by Kane, "Validation research is assumed to involve a systematic effort to improve (1) the accuracy of conclusions based on test scores, (2) the appropriateness of the uses made of these scores, and (3) the quality of the data-collection procedures designed to support the proposed conclusions and uses," [18, p. 3]. Douglas and Purzer further spell out misuses of the term validity [5].

The two aspects of validity most commonly reported are content and structure. Content aspects of validity give confidence in the process developers went through from identifying what construct is desired to be measured and the items written and chosen to represent the construct. By providing evidence that the test items clearly match the intended purpose of the assessment instrument, researchers show how well a scale named after a construct has items corresponding to that construct. In terms of structural aspects, Factor Analysis in conjunction with Cronbach's alpha provide a more accurate examination of dimensionality and thus, provide further evidence that individual items in the same scale can be appropriately scored together [25]. As is the case with all measurement, regardless of what phenomenon is being studied, a foundational principal is to measure one aspect at a time. Studies on the structural aspects of validity in assessment instrument investigate whether this principal is met and that the structure of the instrument matches the theoretical framework.

Less routinely reported are external, substantive, generalizability, and consequential aspects of validity. External relates to outside variables that are conceptually related. For example, before using an assessment instrument to evaluate effective instruction, there must be evidence the instrument has instructional sensitivity [41]. Another example would be to test the hypothesized relationship between the constructs purported as measured in the instrument and the desired related or unrelated constructs (e.g., social desirability, overall course performance). Substantive refers to the evidence that assessed persons are cognitively involved, as intended by the developers. As pointed out by the National Research Council report, "assessments do not offer a direct pipeline into a student's mind" [7, p. 42]. Without studying substantive aspects, there is limited information regarding whether learners read and interpret the items as intended. For an instrument to be widely used for large-scale research or for high stake decisions, generalizability is very important. Measurement properties can change when sample sizes are small. Consequential aspects are important when an assessment's intended use will include decisions about the person or group assessed. Arguably, most, if not all, of the assessment instruments published in our reviewed selection of journals are for research purposes, not educational decision-making. It is understood, that the consequences of use would likely be related to research or curriculum decisions rather than direct consequences to those assessed.

Considering the dominant group in engineering has historically been white male, the finding that none of the published instruments examine test bias or measurement invariance between groups is concerning, but also a reflection of historical values. While fairness of educational testing is not a new concept, the latest *Standards* [9] have placed this aspect of quality on par with validity and reliability. Just as the concept of validity has undergone major transformation, conceptualization and empirical study of fairness will continue to advance the community forward. Unfortunately, inadequate assessment is a significant barrier to progress in diversity [42]. It is a mistake to assume that any given assessment instrument measures all people the same. Practically speaking, simply reporting mean

differences between sub-groups *does not* sufficiently address whether the differences are due to bias in the instrument, opportunities to learn, or actual differing ability between groups. Groups can be differentiated along many dimensions, including gender, race/ethnicity, poverty or socioeconomic background, international or domestic status, or status as English language learners among others. Stating that an instrument has found differences in gender and concluding that there is a difference in affect, understanding, or performance by gender without first ensuring fairness in the instrument is problematic. Furthermore, as pointed out by others, fairness in assessment goes beyond psychometric studies. There are many ways to explore social justice and equity in how students are assessed. The heightened call for fairness in conjunction with the present research demonstrates there is a pressing need is for engineering education researchers to examine the evidence that instruments can be used to fairly assess diverse groups.

There are many potential explanations for the findings that aspects of internal reliability, content-related validity, and substantive aspects of validity are the most commonly reported types of evidence in engineering education assessment publications. One likely contributor is that these areas can be studied through psychometric analysis and map to two of the three historical types of validity; namely, criterion and construct. These areas have been understood as crucial aspects for any assessment instrument designed for widespread use. For example, calculating the internal consistency is prerequisite to many statistical techniques, such as Factor Analysis. It is very straightforward to calculate and requires far less effort than administering an instrument twice to the same sample weeks apart.

Another potential reason is that researchers are more familiar with methods they use and read about in other published articles. Just as this research found several articles describing internal consistency and aspects of content and structure, so also those seeking to publish an instrument would also find those areas of evidence as potential 'templates' of what to include to be accepted in a particular journal.

6.2 Recommendations for assessment development researchers and users

Assessment instruments need to be designed for very specific purposes and it would not be appropriate to recommend one "right" method. Determining the quality assessment instruments is very nuanced and requires critical thinking on the part of the user. The questions asked must be: (1) What is this instrument designed to measure? (2) How did the authors go about evidencing the appropriateness of that use? What evidence was collected and why? (3) How similar is my disired use to the intended use of the developers? (4) What additional evidence is needed to support my intended use? and (5) What should be the expected and intended consequences of the test?

More generally, we offer the following recommendations are offered as a basis for what evidence to include when reporting newly developed assessment instruments designed for research purposes in engineering education.

6.3 Use and limitations

To demonstrate evidence for the reliability, validity, and fairness of this research, the researchers provide transparent details. In terms of reliability of our coding process, the search strategy to locate manuscripts is included with sufficient detail so that others could attempt to replicate this work, adding to the consistency of the findings. In addition, two researchers independently reviewed each publication at each stage of the review process, to increase the trustworthiness of the results. In terms of validity of use, the researchers have provided evidence of what is commonly reported in high quality engineering education journals and compared the results to how validity is discussed by educational

Table 3. Recommendations for Assessment Development Publications

Area	Recommendation
Content	• Include a rich description of the theory, construct definitions, domain, and scope being assessed. • Include a detailed description of how items were developed and evaluated.
Substantive	• Study target audiences' cognitive process when reading and answering items.
Structural	• Include guiding theory for studying technical quality.
Reliability	• Include a description of what potential sources of variance are of concern and how addressed.
Fairness	• Describe who specifically the instrument is intended to assess and how fairness was considered.
Use	• Provide a detailed description of what provided evidence means for specific use. • Provide additional sources of evidence as justified by intended use. (e.g., for use as a measure for effective instruction, an experimental study demonstrating ability to differentiate between control and treatment groups) • Provide appropriate use cases of the instrument, including limitations and generalizability of results.

assessment specialists. Based on these findings, an appropriate, or valid, use of this research would be to consider what evidence one has to substantiate the use of an assessment instrument for a given use. Lastly, the researchers considered fairness in reviewing others work and how to report findings. Reviewing respected colleagues' work and providing fair synthesis was of utmost importance to the researchers. The researchers approached the review from a perspective that assessment research is developmental [5] and there is no "checklist" that all assessment publications must include [10]. In this vein, the authors chose to focus solely on higher-tier engineering education journals, rather than conference proceedings or less cited journals. It would be unfair to expect authors to provide the same amount of evidence as expected in higher tier dissemination outlets. Furthermore, the authors subjected their own published works in the dataset to the same level of review as other works. The authors did not review their own individual published assessment works, rather, members of the team not involved in the publications reviewed those papers.

This paper is intended to create an overview of what type of evidence is commonly reported in engineering education research instrument validation publications but does not approach the technical quality of how well the evidence is gathered. As such, it can be used as a foundation for a more in-depth investigation specific to the areas that are frequently addressed in publications. For example, many authors report aspects of structural and content validity. Future research may consider the appropriateness of methods such as procedural decisions when conducting Factor Analysis for structural validity claims or how well supported claims of content validity are. The authors recognize that some articles discussing instrument development may not have been found using the sampling approach and search method, however, the team consulted with an engineering librarian to minimize this limitation. The full sample of articles used in this study is listed in the appendix.

7. Conclusion

It is unrealistic to expect that any assessment validation article would empirically examine all aspects of reliability, validity, and fairness. Yet, these three areas are foundational and necessary for high quality assessment. Any assessment instrument designed to be of high importance deserves careful attention to all three areas. It can take years and several iterations before an assessment instrument demonstrates appropriate levels of validity evidence for a particular use. Validity is an ongoing process of collecting evidence for every intended purpose. Published assessment instruments must be viewed as what they are: works in progress. Validity is the responsibility of an entire community of researchers who seek to build knowledge on a solid foundation. The purpose of this research is to present common reporting practice in engineering education assessment development among select journals and to provide recommendations of what type of information to report. In conclusion, within select engineering education research journals, the most frequently reported evidence of assessment validation is in the form of internal consistency, structural, and content properties. Other types of reliability, such as test-retest or measures of precision are largely not reported. Also less frequently reported are external and substantive aspects of validity. Of note, evidence of consequences, generalizability, and fairness were largely void. Additionally, the authors have provided recommendations regarding the types of evidence to report in validation studies of newly developed instruments. These recommendations serve not as a checklist to stamp assessments as "validated", but rather as practical guidelines for what information is needed so that potential users can evaluate whether an instrument is appropriate for their purpose.

Acknowledgements—This project is partially supported by the National Science Foundation TUES Grant No. 1245998. 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.

The authors would like to thank Amy Van Epps for providing consultation regarding search strategy and methods to identify articles.

References

1. M. Wald, Editorial, *International Journal of Engineering Education*, **22**(1), 2006, p. 1.
2. J. R. Lohmann, Building a community of scholars: The role of the Journal of Engineering Education as a research journal, *Journal of Engineering Education*, **94**(1), 2005, pp. 1–6.
3. R. A. Streveler and K. A. Smith, Conducting rigorous research in engineering education, *Journal of Engineering Education*, **95**(2), 2006, pp. 103–105.
4. B. M. Olds, B. M. Moskal and R. L. Miller, Assessment in engineering education: Evolution, approaches and future collaborations, *Journal of Engineering Education*, **94**(1), 2005, pp. 13–25.
5. K. A. Douglas and Ş. Purzer, Validity: Meaning and relevancy in assessment for engineering education research, *Journal of Engineering Education*, **104**(2), 2015, pp. 108–118.
6. J. W. Pellegrino, Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggests, *Commissioned by the National Center on Education and the Economy for the New Commission on the Skills of the American Workforce*, 2006.
7. National Research Council, *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academies Press, Washington, D.C., 2001.
8. F. Oswald and T. Chavous, Reflections on the day, *presented at the National Academies Syposium on Assessing Hard-to-Measure Cognitive, Intrapersonal and Interpersonal Competencies*, Washington, DC, US, 16-Dec-2005.

9. American Educational Research Association, American Psychological Association, National Council on Measurement in Education and Joint Committee on Standards for Educational and Psychological Testing, *Standards for Educational and Psychological Testing 2014 edition*, AERA, Washington DC, 2014.
10. N. B. Songer and M. A. Ruiz-Primo, Assessment and science education: Our essential new priority?, *Journal of Research in Science Teaching*, **49**(6), 2012, pp. 683–690.
11. R. G. Netemeyer, W. O. Bearden and S. Sharma, *Scaling Procedures: Issues and Applications*, SAGE Publications, Thousand Oaks, CA, 2003.
12. A. R. Carberry, H.-S. Lee and M. W. Ohland, Measuring engineering design self-efficacy, *Journal of Engineering Education*, **99**(1), 2010, pp. 71–79.
13. Q. Li, D. B. McCoach, H. Swaminathan and J. Tang, Development of an instrument to measure perspectives of engineering education among college students, *Journal of Engineering Education*, **97**(1), 2008, pp. 47–56.
14. L. R. Lattuca, D. Knight and I. Bergom, Developing a measure of interdisciplinary competence, *International Journal of Engineering Education*, **29**(3), 2013, pp. 726–739.
15. Ş. Purzer, N. Fila and K. Nataraja, Evaluation of current assessment methods in entrepreneurial engineering entrepreneurship education, *Advances in Engineering Education*, **5**(1), 2016, pp. 1-27.
16. N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, *Journal of Engineering Education*, **104**(4), 2015, pp. 454–496.
17. S. Messick, Validity of psychological assessment, *American Psychologist*, **50**(9), 1995, pp. 741–749.
18. M. T. Kane, *An Argument-based Approach to Validation*, The American College Testing Program, Iowa City, 1990, pp. ii–44.
19. R. J. Mislevy, R. G. Almond and J. F. Lukas, A brief introduction to evidence-centered design, *ETS Research Report Series*, **2003**(1), 2003, pp. i–29.
20. S. Messick, *Validity of Test Interpretation and Use*, Educational Testing Service, Princeton, NJ, 1990, pp. 1–29.
21. T. R. Hinkin, A review of scale development practices in the study of organizations, *Journal of Management*, **21**(5), 1995, pp. 967–988.
22. R. L. Worthington and T. A. Whittaker, Scale development research: A content analysis and recommendations for best practices, *The Counseling Psychologist*, **34**(6), 2006, pp. 806–838.
23. J. A. Arjoon, X. Xu and J. E. Lewis, Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence, *Journal of Chemical Education*, **90**(5), 2013, pp. 536–545.
24. A. Coughlan, D. Johnson, H. A. Diefes-Dux, K. A. Douglas, K. Erk, T. A. Faltens and A. Strachan, Enhanced learning of mechanical behavior of materials via combined experiments and nanoHUB simulations: Learning modules for sophomore MSE students, *Proceedings of the Materials Research Society Symposium*, Boston, MA, January 2015, 18 February 2015.
25. J. M. Cortina, What is coefficient alpha? An examination of theory and applications, *Journal of Applied Psychology*, **78**(1), 1993, pp. 98–104.
26. J. Nunnally, *Psychometric Methods*, McGraw Hill, New York, 1978.
27. L. J. Cronbach, Test validation, in R. L. Thorndike (ed), *Educational Measurement*, 2nd edn, American Council on Education, Washington, D. C., 1971, pp. 443–507.
28. R. J. Shavelson, N. M. Webb and G. L. Rowley, Generalizability theory, *American Psychologist*, **44**(6), 1989, pp. 922–932.
29. L. M. Crocker and J. Algina, *Introduction to Classical and Modern Test Theory*, Cengage Learning, Mason, OH, 2008.
30. S. Messick, Meaning and values in test validation: The science and ethics of assessment, *Educational Researcher*, **18**(2), 1989, pp. 5–11.
31. S. Messick, Test validity: A matter of consequence, *Social Indicators Research*, **45**(1), 1998, pp. 35–44.
32. S. Messick, The interplay of evidence and consequences in the validation of performance assessments, *Educational Researcher*, **23**(2), 1994, pp. 13–23.
33. B. Plake, Major changes in the 2014 standards, *presented at the Standards Public Briefing on Standards for Educational and Psychological Testing: Essential Guidance and Key Developments in a New Era of Testing*, Washington, D.C., 12 September 2014.
34. National Center for Science and Engineering Statistics and National Science Foundation, *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2013*, National Science Foundation, Arlington, VA, 2013.
35. ASEE: Year of action on diversity, http://diversity.asee.org/about, Accessed 7 October 2014.
36. K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, SAGE Publications, Thousand Oaks, CA, 2012.
37. M. Borrego, M. J. Foster and J. E. Froyd, What is the state of the art of systematic review in engineering education?, *Journal of Engineering Education*, **104**(2), 2015, pp. 212–242.
38. M. Q. Patton, *Qualitative Research and Evaluation Methods*, SAGE Publications, Thousand Oaks, CA, 2001.
39. A. S. Van Epps, Beyond JEE: Finding publication venues to get your message to the 'right' audience, *presented at the ASEE Annual Conference*, Atlanta, GA, 2013.
40. C. Charyton, R. J. Jagacinski, J. A. Merrill, W. Clifton and S. DeDios, Assessing creativity specific to engineering with the revised creative engineering design assessment, *Journal of Engineering Education*, **100**(4), 2011, pp. 778–799.
41. M. A. Ruiz-Primo, R. J. Shavelson, L. Hamilton and S. Klein, On the evaluation of systemic science education reform: Searching for instructional sensitivity, *Journal of Research in Science Teaching*, **39**(5), 2002, pp. 369–393.
42. National Academy of Engineering and American Society for Engineering Education, *Surmounting the Barriers: Ethnic Diversity in Engineering Education: Summary of a Workshop*, National Academies Press, Washington D.C., 2014.
43. L. Wise, Major changes in the 2014 standards, *presented at the Standards Public Briefing on Standards for Educational and Psychological Testing: Essential Guidance and Key Developments in a New Era of Testing*, Washington, D.C., 12 September 2014.

APPENDIX

The articles used in the analysis ordered chronologically and alphabetically.

Year	Article
2005	R. M. Felder and J. Spurlin, Applications, reliability and validity of the index of learning styles, *International Journal of Engineering Education*, **21**(1), 2005, pp. 103–112.
	P. S. Steif and J. A. Dantzler, A statics concept inventory: Development and psychometric analysis, *Journal of Engineering Education*, **94**(4), 2005, pp. 363–371.
	K. E. Wage, J. R. Buck, C. H. G. Wright and T. B. Welch, The signals and systems concept inventory, *IEEE Transactions on Education*, **48**(3), 2005, pp. 448–461.

2006 Ş. Yaşar, D. Baker, S. Robinson-Kurpius, S. Krause and C. Roberts, Development of a survey to assess K-12 teachers' perceptions of engineers and familiarity with teaching design, engineering, and technology, *Journal of Engineering Education*, **95**(3), 2006, pp. 205–216.

2007 S. Gallardo, F. J. Barrero, M. R. Martinez-Torres, S. L. Toral and M. J. Duran, Addressing learner satisfaction outcomes in electronic instrumentation and measurement laboratory course organization, *IEEE Transactions on Education*, **50**(2), 2007, pp. 129–136.

T. A. Litzinger, S. H. Lee, J. C. Wise, and R. M. Felder, A psychometric study of the Index of Learning Styles©, *Journal of Engineering Education*, **96**(4), 2007, pp. 309–319.

2008 Q. Li, D. B. McCoach, H. Swaminathan and J. Tang, Development of an instrument to measure perspectives of engineering education among college students, *Journal of Engineering Education*, **97**(1), 2008, pp. 47–56.

D. M. Qualters, T. C. Sheahan, E. J. Mason, D. S. Navick and M. Dixon, Improving learning in first-year engineering courses through interdisciplinary collaborative assessment, *Journal of Engineering Education*, **97**(1), 2008, pp. 37–45.

2009 C.-C. Lin and C.-C. Tsai, The relationships between students' conceptions of learning engineering and their preferences for classroom and laboratory learning environments, *Journal of Engineering Education*, **98**(2), 2009, pp. 193–204.

2010 A. R. Carberry, H.-S. Lee and M. W. Ohland, Measuring engineering design self-efficacy, *Journal of Engineering Education*, **99**(1), 2010, pp. 71–79.

R. M. Felder and R. Brent, The National Effective Teaching Institute: Assessment of impact and implications for faculty development, *Journal of Engineering Education*, **99**(2), 2010, pp. 121–134.

M. J. Nathan, N. A. Tran, A. K. Atwood, A. Prevost and L. A. Phelps, Beliefs and expectations about engineering preparation exhibited by high school STEM teachers, *Journal of Engineering Education*, **99**(4), 2010, pp. 409–426.

M. J. Prince, M. Vigeant and K. Nottis, Assessing misconceptions of undergraduate engineering students in the thermal sciences, *International Journal of Engineering Education*, **26**(4), 2010, pp. 880–890.

2011 C. Charyton, R. J. Jagacinski, J. A. Merrill, W. Clifton and S. DeDios, Assessing creativity specific to engineering with the revised creative engineering design assessment, *Journal of Engineering Education*, **100**(4), 2011, pp. 778–799.

T. Hong, Ş. Purzer and M. E. Cardella, A psychometric re-evaluation of the design, engineering and technology (DET) survey, *Journal of Engineering Education*, **100**(4), 2011, pp. 800–818.

S. A. Lathem, M. D. Neumann and N. Hayden, The socially responsible engineer: Assessing student attitudes of roles and responsibilities, *Journal of Engineering Education*, **100**(3), 2011, pp. 444–474.

R. A. Streveler, R. L. Miller, A. I. Santiago-Román, M. A. Nelson, M. R. Geist and B. M. Olds, Rigorous methodology for concept inventory development: Using the 'assessment triangle' to develop and test the thermal and transport science concept inventory (TTCI), *International Journal of Engineering Education*, **27**(5), 2011, p. 968.

2012 B. M. Capobianco, B. F. French and H. A. Diefes-Dux, Engineering identity development among pre-adolescent learners, *Journal of Engineering Education*, **101**(4), 2012, pp. 698–716.

M. Prince, M. Vigeant and K. Nottis, Development of the heat and energy concept inventory: Preliminary results on the prevalence and persistence of engineering students' misconceptions, *Journal of Engineering Education*, **101**(3), 2012, pp. 412–438.

S. P. Schaffer, X. Chen, X. Zhu and W. C. Oakes, Self-efficacy for cross-disciplinary learning in project-based teams, *Journal of Engineering Education*, **101**(1), 2012, pp. 82–94.

2013 L. R. Lattuca, D. Knight and I. Bergom, Developing a measure of interdisciplinary competence, *International Journal of Engineering Education*, **29**(3), 2013, pp. 726–739.

Y. Maeda, S. Y. Yoon, K. Kim-Kang and P. K. Imbrie, Psychometric properties of the revised PSVT:R for measuring first year engineering students' spatial ability, *International Journal of Engineering Education: Special Issue: Human Computer Interaction in Engineering Education*, **29**(3), 2013, pp. 763–776.

J. Zhu, Y. Li, M. F. Cox, J. London, J. Hahn and B. Ahn, Validation of a survey for graduate teaching assistants: Translating theory to practice, *Journal of Engineering Education*, **102**(3), 2013, pp. 426–443.

2014 B. Ahn, M. F. Cox, J. London, O. Cekic and J. Zhu, Creating an instrument to measure leadership, change, and synthesis in engineering undergraduates, *Journal of Engineering Education*, **103**(1), 2014, pp. 115–136.

W. C. Lee, H. M. Matusovich and P. R. Brown, Measuring underrepresented student perceptions of inclusion within engineering departments and universities, *International Journal of Engineering Education*, **30**(1), 2014, pp. 150–165.

S. Y. Yoon, M. G. Evans and J. Strobel, Validation of the teaching engineering self-efficacy scale for K-12 teachers: A structural equation modeling approach, *Journal of Engineering Education*, **103**(3), 2014, pp. 463–485.

2015 I. E. Esparragoza, S. Lascano Farak, J. R. Ocampo, J. Nuñez Segovia, R. Viganò, J. Duque-Rivera and C. A. Rodriguez, Assessment of students' interactions in multinational collaborative design projects, *International Journal of Engineering Education*, **31**(5), 2015, pp. 1255–1269.

H. K. Ro, D. Merson, L. R. Lattuca and P. T. Terenzini, Validity of the contextual competence scale for engineering students, *Journal of Engineering Education*, **104**(1), 2015, pp. 35–54.

K. G. Wilkins, B. L. Bernstein and J. M. Bekki, Measuring communication skills: The STEM interpersonal communication skills assessment battery, *Journal of Engineering Education*, **104**(4), 2015, pp. 433–453.

Kerrie A. Douglas is an Assistant Professor in the School of Engineering Education at Purdue University. She earned her PhD in Educational Psychology, with a concentration on evaluation and assessment, from Purdue University in 2012. Her research is focused on decreasing barriers to high-quality assessment practice in engineering education. This focus includes what evidence and rationale are used to justify educational data use and the consequences of that intended use. She studies how to combine multiple sources and types of data to provide a deeper assessment of learners and how assessments can be used to support learning.

Şenay Purzer is an Associate Professor in the School of Engineering Education at Purdue University and the Director of Assessment Research at the INSPIRE Institute for Pre-college Engineering Research. Şenay is a NAE/CASEE New Faculty Fellow and the recipient of a 2012 NSF CAREER award. Her research focuses on assessment of design learning with a specific focus on information literacy, innovation, and decision-making processes. She received a B.S.E with distinction in Engineering at Arizona State University in 2009 as well as a B.S. degree in Physics Education in 1999. Her Ph.D. degree is in Science Education with a focus on engineering education from Arizona State University.

Johannes Strobel is a Full Professor in the School of Information Science & Learning Technologies, University of Missouri. His research is in STEM Education, particularly engineering education. He studies empathy and care in technical fields; teachers' engineering knowledge and design; and the use of maker spaces to teach mathematics.

Anastasia Rynearson is a Post-Doctoral Research Assistant in the School of Engineering Education at Purdue University. She received her PhD from Purdue University in Engineering Education (2016) and holds Bachelor of Science and Master of Engineering degrees in Mechanical Engineering from the Rochester Institute of Technology (2008). Her research focuses on early P-12 STEM education including engineering identity development.