A practitioner's guide to validation framework development

Chad W. Buckendahl

Alpine Testing Solutions

April 30, 2013

Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA.

A practitioner's guide to validation framework development

As the measurement community gets muddled in the depths of esoteric discussions of the definition and meaning of validity, practitioners struggle with the day-to-day responsibility of operationalizing the concept, often without an understanding of what evidence to collect or why they have been advised to pursue a particular line of inquiry. Without better guidance for how to identify, prioritize, collect, and evaluate evidence, the educational measurement community runs the risk of further distancing itself from the stakeholders that could most benefit from information about how to transfer the concept to implementation.

Many practitioners can dutifully recite from the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) that "validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by the proposed uses of tests" (p. 9). It sounds deceptively simple: ensure that the accumulation of evidence and theory for your testing program support how you interpret scores for an intended purpose. In transitioning from recommended concept to actual practice, one's vision can get quickly blurred due to the lack of practical guidance as to how one can operationalize this concept.

The purpose of this paper is to elaborate a proposed organizing framework for guiding testing programs in the validation process. The concepts and practices described herein are based on a recommendation that emerged from an evaluation of the *National Assessment of Educational Progress* (Buckendahl et al., 2009) coupled with elements of change management strategies discussed from a psychometric prescriptive in Henderson-Montero and Buckendahl (2010). The goal of this paper is to suggest a usable template that testing programs can adopt or adapt to design the validation plan and evaluate the evidence needed to support interpretations and uses of test scores. A secondary goal is to demonstrate how using this organizing framework

can assist testing programs in communicating key elements and processes of their program by promoting transparency of applied measurement science to a broader group of stakeholders. As part of this paper, an application of the framework from a large-scale licensure testing program is included. However, the core tenets of the framework can be generalized across sectors of the testing industry because the development is based on the context of the respective program.

Brief overview of validity theory

In attempting to bridge theory and practice, it important to begin with the theoretical foundation on which educational and psychological measurement is based. The field has strived to more clearly articulate this seemingly opaque concept over the decades. The history of validity theory evolution as defined in the first version of the *Standards* in 1954 has been driven by the shift from considering validity evidence as separate types (e.g., content, criterion, construct) to a unified vision that characterizes construct evidence as an umbrella under which convergent sources of evidence are shielded from the storms of irrelevant variance. Cronbach and Meehl (1955) started the next phase of thinking on validity and validation as a process. Cronbach's (1971) discussion of validity as dependent on the intended interpretation of scores then further clarified the field's definition. This concept was elaborated in Messick (1989) to articulate five sources of evidence that form the basis of the chapter on validity in the *Standards* (AERA, APA, & NCME, 1999). With many years of evolution, clarification, and definition, it would seem that the field would be very refined in its conceptualization of validity. However, practitioners continue to be at a loss with respect to how to meaningfully implement Messick's recommendations. This scatterplot of interpretation was illustrated in Newton & Shaw (2012) who illustrated the dozens of representations of "validity" that can be found in the literature (e.g., face validity, procedural validity, divergent validity).

Influenced perhaps by his experiences with a number of operational testing programs, Kane (2006) has attempted to bring us closer to a utilitarian view of validation as an argument based process in which the collection of evidence is continually evaluated against the intended interpretations and uses of scores. This iterative, contextual process requires users to evaluate and adapt as needed. These steps have been laudable. Further, in his description of multi-level validity, Zumbo (2007) reinforced the commonly held belief that validity is a matter of degree and not an either/or proposition. He further argues that the same test scores may require differing interpretative arguments depending on the intended use by a given stakeholder group. This concept was highlighted in Buckendahl et al. (2009) in that scores from the U.S.'s *National Assessment of Educational Progress* may have differential interpretations and uses depending on whether you are a classroom educator, a parent, an administrator, or a policymaker.

Extending the discussion about the influence of context when defining validity in practice, Cizek (2012) suggested disentanglement of score interpretations from justifications for specific uses of the scores. Would this shift in thinking clarify our understanding and application of the concept or unintentionally provide a restatement of other core principles that we want user to internalize? Responding to Newton's (2013) and Sireci's (2013) requests for greater simplification of the validity and interpretation/use argument approach, Kane (2013) pushed back, suggesting that the interpretation/use argument is necessary because of the lack of specificity by programs regarding these concepts and how they guide the validation activities that contribute to the validity argument.

In the friendly confines of psychometric dialogues, we are probably comfortable with a certain amount of ambiguity. As a profession, psychometricians can answer almost any question posed to them with a response of "it depends." Not many of us will laugh when hearing that

because we have all used that response because recommended policies and practices do depend on number of factors, not all of them purely psychometric. However, to many practitioners, these philosophical discussions do not provide appropriate guidance for how to identify, prioritize, collect, and evaluate the evidence needed for their programs. And

Because it is a safe assertion that the population of practitioners is larger than the population of theorists on the topic of validity, Brennan's (2006) indirect call for greater guidance on conducting validation research and Newton's (2013) and Sireci's (2013) suggestions to provide a more parsimonious design for users served as motivation for this work.

Overview of organizing validation framework

As a result of the diverse stakeholder groups, interpretations, and uses – intended and unintended – Buckendahl et al. (2009) suggested three overarching categories that could be used as an organizational template to develop a comprehensive validation framework: operational, policy, and innovation. However, beyond these high level descriptions, they did not provide additional input regarding how such a framework would be constructed and what would be included in an output of such an effort. In this paper, the overarching organization is retained but with guidance about how it can be converted into an operational tool for use by practitioners. For purposes of this paper, practitioners are considered individuals responsible for design, execution, and evaluation of operational testing program activities. Because many of these practitioners are not psychometricians, the framework is intended for use by broader stakeholder groups than earlier efforts. Each of these organizing categories is described here.

Operational

The operational element of the framework represents the largest component and the one that practitioners generally consider when identifying, prioritizing, and collecting evidence

because of its proximity to the pragmatic needs of a testing program. This element can then be further subdivided into five sources of evidence suggested by the *Standards* (AERA, APA, & NCME, 1999) and recommended by Sireci (2012): test content, response processes, internal structure, relations to other variables, and consequences of testing. However, it is also important to reinforce that each testing program's validation framework will be unique with respect to defined interpretation and use. Therefore, these sources should not be interpreted as being equally weighted or having fixed set of weights that would generalize across programs. Rather, each source of evidence will contribute differentially as it aligns with the intended uses and interpretations. This is particularly important when evaluating the evidence collected in each of these phases of the framework. Because not all sources of evidence are equal in their contribution to the validity argument, holistic conclusions about the compliance of the program with expectations need to consider the differential impact of these sources (see, for example, Buckendahl & Plake, 2006).

Policy

The policy element of the validation framework is included as a placeholder and reminder of the dynamic nature of policies that can impact practices for testing programs. It is generally the most malleable of the sources of evidence that may be requested or needed because of the uncertain nature under which these occur. Although we can anticipate some policy changes or needs that may necessitate expanding, the intended interpretations and uses, policymakers have demonstrated that we cannot possibly anticipate them all. Thus, this element of the framework is intended to acknowledge that during the course of a testing program's development, validation, and maintenance, there will be unexpected instances where validation research is needed – whether it is directly requested or not. Henderson-Montero and Buckendahl

(2010) note that programs' and specifically, psychometricians' ability to adapt to such changing environments can greatly impact the success of a program.

Innovation

Whereas policy may challenge programs to adapt rapidly to changing validation research needs, the innovation section of the validation framework is intended to focus on the macro level design elements of a testing program that may occur parallel to the core operational needs. In an effort to move from a current to future state, testing programs often neglect the strategic design features and related validation research that is necessary to move them to the next phase of their program's evolution. For example, testing programs transitioning from paper-pencil to computer based testing or from linear to adaptive delivery requires organized validation efforts that may occur over multiple operational cycles before fully migrating to the new phase. Without consideration for the future validation needs, testing programs can be caught in a situation of unpreparedness that leaves them exposed to challenges that they might not have had under the legacy design.

Under the organizational framework described above, testing programs can begin to identify, prioritize, collect, and evaluate evidence that contributes to their intended interpretation and use of test scores. How would a practitioner use this template to develop a useful framework for their program? The next section of the paper provides an example of how this might occur in practice.

Application of the validation framework

The illustration described in this section of how this validation framework can be applied in practice is based on a testing program whose scores are used as part of determining licensure eligibility. The program includes a series of tests, some written and some performance (e.g.,

clinical, practical, psychomotor, skills) examinations. The use of this illustration is intended to demonstrate the utility of the validation framework for an operational testing program that has a range of stakeholders (e.g., candidates, training programs, regulatory authorities, profession, public). The range of stakeholders described below is similar to what PreK-12 education assessment programs face (i.e., a large constituency of stakeholders desiring multiple, potentially divergent uses). A second, intended message of this illustration is to communicate that even with an intentional validation design the resulting framework may not be considered comprehensive.

Rather than separating the interpretation/use argument as recommended by Kane (2013), this paper applies it as a prerequisite to the larger organizing framework so that it can serve as a reference point when developing the elements and evaluating the resulting evidence. For the licensure testing program that we will use through this section, the first step in applying the framework begins with an articulation of intended interpretations and uses of test scores from the program, a rationale for each, and a description of unintended interpretations and uses. These are described here.

Intended interpretations and uses of test scores

The primary intended interpretation and use of test scores from the program's Clinical Examinations in Dentistry is to provide assurance to the public that dentists have demonstrated the job-related, clinical judgments and skills that are necessary for safe, independent, entry-level practice. A secondary intended interpretation and use of test scores from the program is to provide descriptive feedback to dental training programs regarding the extent to which their students have demonstrated the entry-level clinical judgments and skills related to licensure within the major sub-domains that define the profession.

Passing each clinical examination of dentistry indicates that the candidate has demonstrated minimally competent clinical judgments and skills in domains necessary to safely enter independent dental practice.

Rationale for each intended use

Multiple factors provide justification for the first intended use of scores including an interest by the profession in protecting the integrity of the field. However, the program'primary purpose is supported through state level legislation that requires state boards of dentistry to regulate entry into the profession. Expectations for the profession are then articulated, sometimes vaguely, through state level Practice Acts. These legislative acts may delegate the responsibility to how the regulation occurs directly to the board. In many states, dentisty has a specific psychomotor component of the profession that cannot be delegated to an auxiliary or another member of the patient care team. This aspect of the legislation has an impact on how a testing program is ultimately developed.

A justification of the additional uses of scores from the program is that training programs in the field seek feedback on the effectiveness of the curriculum they have adopted and outcomes of the instructional practices. Therefore, providing descriptive feedback on candidates', and by default students', performance can serve as feedback to these institutions. However, the secondary uses of scores are more difficult to justify in practice and are discussed in the next section.

Anticipated, unintended interpretations and uses of test scores

Like many licensure testing programs where formal educational programs are designed to lead to the profession, there are often requests to use results from these programs for evidence of programmatic outcomes for accreditation purposes. This is where the desire by programs to

receive the descriptive information noted above can morph into an unintended consequence. Specifically, the results from licensure testing programs are rarely designed to be an instrument for comprehensive program evaluation or as an indicator of educator effectiveness. Because licensure targets the knowledge, skills, abilities, and judgments needed for public protection, all components of an educational training program that might prepare a student will not necessarily contribute to that intent. As a result there is often a desire by faculty members and policymakers who want to extrapolate the descriptive level of results that the testing program may be willing to support beyond the intended use. Noting these limitations up front and communicating them transparently through program documents, candidate information sources, and score reports can be used as strategies for mitigating unintended uses.

Defining stakeholder subpopulations

Another important component is to define and describe the range of stakeholders of the program. Table 1 illustrates how these stakeholders were defined for the licensure testing program that serves as the common example in this paper.

Table 1. Stakeholders of an illustrative program's clinical examinations in dentistry

Role/Entity	Category	Rationale for inclusion
New graduates	Candidate Population	Licensure is required to enter practice in all jurisdictions.
Dentists whose license has lapsed, been suspended or revoked	Candidate Population	Maintenance of a license requires demonstration of knowledge of current practice in the field. Practice related reasons may have contributed to a lapse or enforcement action.
Dentists who do not qualify for licensure by credential	Candidate Population	Licensure is a state function and not all states accept results from all examination programs. A state may not accept a particular exam's results and the dentist may not have sufficient experience to qualify for licensure by credential.
State dental licensing boards	Regulatory Authority	Policy body charged with public protection within a state. Responsible for making decisions about initial licensure, maintenance, and enforcement.
Dental training faculty and institutions	Education Providers	Required to balance historical, current, and emerging practice in their curriculum and instruction.
Community members	End Users	Because licensure exists for protection of the public, members of the general public are the ultimate stakeholder if incompetent dentists are knowingly allowed into practice.

The stakeholders described in Table 1 are necessarily inclusive of not only the candidate (i.e., examinee) population, but also the educators, regulatory authority, and general public. Because the validation framework for the program includes educators, policymakers, and the public, understanding their role in the broader system can be valuable, particularly when thinking about the evidence of consequences that is part of the operational component. In the next section, the resultant validation activities that were designed for this program are outlined.

Validation framework for the clinical examinations in dentistry

Using the framework described above, Table 2 shows the validation activities that the program has chosen to include in its design. For purposes of the confidentiality and proprietary nature of the specific framework, the studies described in Table 2 should be considered representative of the validation activities undertaken by the program, but should not be interpreted as communicating the program's long term programmatic strategies.

As a brief overview of the tests within the program, the clinical examinations in dentistry include both written and performance testing components. The written examination is intended to measure candidates' ability to diagnosis conditions and develop treatment plans in a range of domains. The four separately scored performance examinations represent a combination of simulated and patient based demonstrations of skills in four domains (i.e., operative, endodontics, fixed prosthodontics, periodontics). The performance examinations are scored by trained and calibrated examiners who physically evaluate candidates' performance on the respective tasks. In Table 2, these examinations are distinguished as clinical judgments (written examination) and clinical skills (performance examinations).

Table 2. Validation activities for clinical examinations in dentistry by each component of the organizing framework

Operational validation activities

a. Content

1. Update practice analysis – regularly scheduled activity that occurs every 5-7 years to represent the current practices in the field. [Clinical judgments and skills examinations]
2. Task development/redevelopment informed by results of job task analysis – scheduled annually. [Clinical skills examinations]
3. Item development/redevelopment based on evaluation of gap analysis of functioning items in the bank – scheduled annually. [Clinical judgment examination]

4. Content validation studies to independently evaluate content, cognitive demand, and performance level relative to the results of the practice analysis – scheduled every 5-7 years to align with practice analysis and annually to evaluate new content that is developed. [Clinical judgment and skills examinations]
5. Standard setting to reset the passing score for examinations – scheduled every 5-7 years in conjunction with redevelopment that occurs after the practice analysis. [Clinical judgment and skills examinations]

b. Response processes

1. Alignment of test design with job-related expectations of the profession (e.g., cognitive tasks and level of processing, psychomotor skills) – scheduled every 5-7 years as part of practice analysis design and evaluation. [Clinical judgment and clinical skills examinations]
2. Consideration of cognitive demand and performance demand during content validation studies for item and task development/redevelopment. [Clinical judgment and clinical skills examinations]
3. Psychometric analysis of rater performance and decision consistency – scheduled semi-annually to provide interim feedback for formative purposes and annually for formative and summative purposes. [Clinical skills examinations]

c. Internal structure

1. Psychometric analysis of item and form performance, including Differential Item Functioning to evaluate item drift – scheduled semi-annually to provide interim information and annually to provide information for forms re-assembly. [Clinical judgment examination]
2. Linking, equating, and scaling analysis to maintain the interpretation of the cut score across years and forms of the examinations – scheduled annually. [Clinical judgment examination]
3. Item drift analysis – scheduled annually to evaluate stability of item performance over time. [Clinical judgment examination]

d. Relations to other variables

1. No studies currently scheduled.

e. Consequences of testing

1. Evaluation of the number of practice based disciplinary activities – scheduled annually as summary of state level information for participating states (e.g., complaints, censure, suspension, revocation). [Clinical judgment and skills examinations]
2. Impact of testing program on curriculum and instruction in training programs – scheduled every 5-7 years as an adjunct to the practice analysis [Clinical judgment and skills examinations]

Policy validation activities

a. Anticipate questions regarding evaluation of candidates' performance on licensure examinations to GPA/class rank of dental training programs. [Clinical judgments and skills examinations]
b. Anticipate questions regarding evaluation of the alignment of dental school curriculum with domain representation. [Clinical judgment and skills examinations]
c. Anticipate questions regarding the potential overlap of the written examination with Parts I and II of the national examinations administered by the joint commission – normally scheduled when either testing program undergoes substantive changes. [Clinical judgment examination]
d. Anticipate questions regarding evaluation of current policies regarding examination practices and the use of patients including state Practice Acts, organizational bylaws. [Clinical judgment and skills examinations]

Innovation validation activities

a. Mode effects evaluation of the comparability of natural versus simulated teeth for the procedures measured in the endodontics examination. [Clinical skills examinations]
b. Mode effects evaluation of the comparability of candidates' performance on patients versus simulated patients on the operative examinations. [Clinical skills examinations]
c. Evaluation of the predictive validity of students' mock board examination performance on operational examination performance. [Clinical skills examinations]
d. Evaluation of alternative item types for the written examination (e.g., case studies, technology enhanced performance). [Clinical judgment examination]
e. Evaluation of emerging technologies for simulated performance testing. [Clinical skills examination]

Discussion

As practitioners in the measurement community continue to seek guidance for how to ethically and appropriately apply the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) this paper suggests an organizing validation framework for how to identify, prioritize, collect, and evaluate validity evidence for a testing program using the intended interpretation and use of scores as a guiding principle. Though not intended to contradict current philosophical discussions on the topic, the simplification of validity as a concept and validation as a tangible process will be undoubtedly perceived by some in the field

as too cursory. This is a necessary tradeoff for communicating broader understanding and use of the core concepts of validity in practice. Without usable, operational guidance, theorists and practitioners will continue to be frustrated by the gap between the factions.

With this attempt to promote usability also comes responsibility. Specifically, we cannot compromise threshold elements that would be necessary to support interpretation and use of scores; therefore, practitioners will need to meet philosophy somewhere in the middle. This means that a certain level of assessment literacy is necessary for practitioners to understand and apply the concepts described in the proposed framework. However, we would expect that understanding to evolve over time as users become more sophisticated in their knowledge and application.

For the framework to be useful to contribute to strategy, communicate key elements of the program's validation efforts, and ultimately guide the execution, the creation needs to occur at the outset of the program or during a meaningful phase of redevelopment. This recommendation provides programs with the best opportunity for the validation framework to guide the program and not be used to retrofit or cobble together evidence that may not meaningfully contribute to the broader evaluation of evidence for the program's or users' interpretation and use of test scores.

As a final comment on the use of the organizing validation framework described in this paper, it is important for programs to remember that it is unlikely that a comprehensive plan will emerge at the outset of an effort to development one. Rather, such a framework will be expanded and refined as the program matures. This strategy is encouraged as it reminds practitioners and users alike that the process of validation is dynamic. Therefore it should be no surprise that the framework that guides these processes is also intended to be evaluated and revised.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for education and psychological testing*. Washington, D.C.: American Educational Research Association.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 1–16). Westport, CT: American Council on Education and Praeger.

Buckendahl, C. W., Davis, S. L., Plake, B. S., Sireci, S. G., Hambleton, R. K., Zenisky, A. L, & Wells, C. S. (2009). *Evaluation of the National Assessment of Educational Progress: Final report*. Washington, D.C.: U.S. Department of Education.

Buckendahl, C. W. & Plake, B. S. (2006). Evaluating tests. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 725-738). Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological methods, 17*(1), 31-43.

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

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

Henderson-Montero, D. & Buckendahl, C. W. (2010, May). *Psychometric challenges in a rapidly changing environment*. Paper presented at the annual meeting of the National Council on Measurement in Education. Denver, CO.

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of educational measurement, 50*(1), 115-122

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

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

Newton, P. E. (2013). Two kinds of argument? *Journal of educational measurement, 50*(1), 105-109.

Newton, P. E. & Shaw, S. D. (2012, April). *We need to talk about validity*. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, BC.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of educational measurement, 50*(1), 99-104.

Sireci, S. G. (2012, April). *De-"constructing" test validation*. Paper presented at the annual meeting of the National Council on Measurement in Education. Vancouver, BC.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (pp. 45–79). Amsterdam, NL: Elsevier Science.