Flexibility and Utility in Assessment and Validation

Michael Kane, ETS

EALTA

Dublin, 2018

Main Points

- A validity model can provide a general framework, or specific guidelines, or both.
- In interpreting and using assessment scores, we make claims, and we have an obligation to justify, or *validate* these claims.
- A first step in the process of *validation* is to state the claims being made clearly and completely, and to develop an assessment that supports the claims (the development stage).
- At some point, claims need to be challenged (the critical stage).
- The *validity* of the proposed *interpretation and use* of the assessment scores depends on positive and negative evidence from both phases.

Outline

- Interpretation/use-specific conceptions of validity and general conceptions of validity.
- An argument-based approach to validity
- Developing and Refining an Assessment System – *The Development Phase*
- Checking Interpretations and Uses of Scores - *The Critical Phase*
- Overall Evaluation of the Evidence for Validity
- Some Sleights of Hand and Fallacies
- Concluding Comments

Interpretation/Use-specific Conceptions and General Conceptions of Validity

I/U-Specific and General Conceptions of Validity

- By an I/U-specific conception of validity, I mean one that assumes a specific interpretation and/or use for scores (e.g., prediction).
- I/U-specific models tend to be relatively straightforward, but they are limited to a specific I/U.
- A general conception of validity does not assume any particular I/U, but rather, provides a framework for validating a range of I/Us.
- A general conception tends to be more flexible, but it requires more work to define the I/U and to identify the kinds of evidence relevant to that I/U.

Specific Conception – 1
Cureton (1951)

- In the late 1950s, in the U.S., “predictive validity” was the dominant paradigm.
- “A more direct method of investigation which is always to be preferred wherever feasible, is to … see how well the test performances agree with the task performances.” (Cureton, 1951, p. 622-23)
- The I/U-specific models are straightforward, but are not necessarily easy to implement (e.g., getting a good criterion is not easy).

Specific Conception – 2
Cronbach and Meehl (1955)

- C&M suggested that CV would be, “involved whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined” (p. 282),
- and for “attributes for which there is no adequate criterion” (p. 299).
- CV was presented as an alternative to the criterion and content models, to be used for “constructs” defined by a theory.

Specific Conception – 3
Lissitz and Samuelsen (2007, p.446)

- Validity is defined as, “the study of the test construction process, including the specification of the psychometric theory associated with the assessment device.”
- “the test definition and development process (what is currently known as content validity) and test stability (what is currently known as reliability, …) become the critical descriptors of the test.
- They exist independent of, or regardless of, the application of the test ...”

Specific Conception – 4
Borsboom, et al. (2004)

- "A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes. …"
  - Borsboom, Mellenbergh, and Van Heerden (2004)

- This is a very strong and explicit claim.
- It is not easy to establish a causal claim, and it is not always necessary!

Some Comments on Specific Conceptions of Validity

- For I/U-specific models, the interpretations and uses are built into the model, and therefore don't need to be specified, or even mentioned.
- The kinds of evidence needed to support validity can be listed in advance.
  - For a predictive model, evaluations of accuracy of the predictions, …
  - For a construct model, evaluations of the adequacy of the theory, …
- If the I/U defining the I/U-specific model matches the proposed I/U, the I/U-specific mode should work well.
- If the I/U in the model does not match the proposed I/U, the I/U-specific mode will usually not be satisfactory.

General Conception 1
Cronbach (1971)

- "Narrowly considered, *validation* is the process of examining the accuracy of a specific prediction or inference made from a test score." (p. 443)
- "More broadly, validation examines the soundness of all the interpretations of a test – descriptive and explanatory interpretations as well as situation-bound predictions." (p. 443)
- "The phrase *validation of a test* is a source of much misunderstanding. One validates, not a test, but an *interpretation of data arising from a specific procedure*." (p. 447)

General Conception 2
Messick (1989)

- "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment." [italics in original] (p.13)

- Note that Messick framed his discussion in terms of "construct validity", but between 1955 and 1989, CV had evolved from a specific conception in terms of theoretical constructs, to a general conception with a very broad and open definition of the term "construct".

General Conception 3
Kane (2006)

- “To validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses. Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified.” (p. 17)

Comments on General Conceptions of Validity

- It is not a coincidence that my examples of general conceptions of validity come from different editions of a general reference work (*Educational Measurement*) on assessment.
- For a conception of validity to apply to the many different kinds of assessments and interpretations/uses of scores discussed in EM's chapters, it has to be defined broadly.
- A more specific, and narrowly defined, I/U-specific conception of validity could work quite well in some chapters, but would not work at all well in others.

Prescriptive and Contingent Approaches

- The I/U-Specific models for validity are *prescriptive*, in that they assume a *kind of interpretation/use* and as a result, the evidence needed for validation can be stated in advance.
  - For a predictive model, accuracy of scoring and predictions, generalizability and fairness of scores, appropriateness and fairness of outcomes.
- The General models for validity are *contingent*, in that they consist of frameworks for validating a range of possible interpretations and uses, and for the kinds of evidence relevant to specific claims.
- For the general models, the evidence needed for validation depends on the particular interpretation/use proposed.

What Do We Validate?

- We validate ***interpretations and uses*** of the scores generated by an ***assessment*** for a ***population***.
- What would it mean to validate an interpretation or use without specifying the assessment, or to validate an assessment without an interpretation or use.
- For an I/U-specific model, it can be easy to think that one is validating an assessment, as such, and not an interpretation or use, but that is because the I/U is taken for granted. The I/U is part of the background!
- For example, under Cureton's conception, the main question is how well the **test** predicts some (given) criterion.

The Argument-based Approach as a General Conception of Validity

The argument-based approach does not provide an algorithm for validation, but it does provide a framework for designing and implementing validation efforts that address the claims based on test scores. It requires that the inferences and assumptions inherent in the proposed interpretation and use be specified (the IUA) and that these inferences and assumptions be critically evaluated (the validity argument).

Validation as a Pragmatic, Scientific Activity

Kane (2013, JEM, p.121)

A General Framework for Validity

- The argument-based approach does not assume any particular kind of interpretation or use.
- This makes it more flexible, but it makes it harder to use in some cases, because it requires that the interpretation/use be clearly formulated.
- As a result, validation does not follow a pre-specified formula:
  - Validation “is doing your damnedest with your mind –no holds barred.” (Cronbach, 1988)
- Validation requires that the claims based on scores (i.e., interpretations and uses) be stated (e.g., in an IUA) and that the claims be evaluated (e.g., in the validity argument).

Paraphrasing
Waking Ned Divine,

the argument-based approach “has its faults!”

In particular, it may be too flexible!

A General, Argument-based Approach and I/U-specific Models

- I/U-specific models are special cases of the argument-based approach.
- If an IUA focuses on prediction, the VA will focus on predictive evidence (as in Cureton, 1951).
- If an IUA focuses on the role of a construct in a theory, the VA will examine how well the scores fit this role (C&M, 1954).
- If an IUA focuses on a content domain, the VA will focus on domain relevance and reliability (Lissitz and Samuelsen, 2007).
- If an IUA focuses on causal claims, the VA will evaluate these causal claims (Borsboom, et al., 2004).

Developing and Evaluating Assessment Systems
The Development Phase

Developing and Refining an Assessment System

- Specify the intended interpretation and use .
- Design an assessment that is likely to fit the intended interpretation and use.
- Identify likely challenges, given the intended interpretation and use (e.g., for a performance test, poor reliability; for a MC test, construct underrepresentation), and make revisions if needed.
- Examine the design, materials, procedures, etc. for sources of bias or irrelevant variance, and make revisions if needed.
- Develop an argument (e.g., an IUA) leading from scores to the interpretation/use.

Refining an Assessment System

- The development phase is formative and confirmationist in the sense that any problems that are identified may be fixed, by revising the assessment or the interpretation/use.
- If the assessment itself (e.g., tasks, format, instructions, time limits) does not work well in pilot studies, it can be revised in various ways (e.g., clarify instructions or items).
- For example, if the format of the tasks is unusual, it may be necessary to provide some practice tasks.
- One may need to renegotiate some constraints (e.g., time limits, costs) or limit the claims being made.

Refining an Assessment System to Fit a Psychometric Model

- It is not unusual, to identify items for review, and possible revision or deletion based on fit to a psychometric model.
- This can be a reasonable thing to do, (using classical methods or IRT) but I urge caution and restraint.
- If the test scores are intended to assess a broad, complex content domain, items from some subdomains may have a high probability of being “flagged”, and removing such items may result in an under-representation of those sub-domains.
- For example, in licensure tests in health professions, items on psycho-social aspects of care are hard to write, and they can represent a distinct dimension.

The Development Phase Should Produce Three Major Products

- First, an assessment consisting of more-or-less standardized materials and procedures that will be used to collect data and generate scores based on the assessment takers' performances.
- Second, a statement of the claims (interpretations and uses) to be based on the scores
- Third, an IUA that lays out the inferences and assumptions leading from assessment scores to the interpretations and uses based on these scores.

Evaluating Interpretations and Uses of Scores

The critical Phase

Critical Evaluations
"Kicking the Tires"

- The *development phase* produces evidence that it is reasonable to expect that the assessment will support the proposed score interpretations and uses.
- In the critical phase (or "appraisal phase"), the operating characteristics are evaluated more thoroughly and more realistically in operational contexts.
- In addition, the critical phase provides opportunities to empirically evaluate intended and unintended consequences.

Construct Validation after Thirty Years Cronbach (1989)

- “Despite many statements calling for a focus on rival hypotheses, most of those who undertake CV have remained confirmationist. Falsification, obviously, is something we prefer to do unto the constructions of others.” (p. 153)
- “Besides, as Kuhn (1962) taught us, falsification does not quite work. Theorists have a wonderful power to shake off lethal doses of it. (Serlin & Lapsley, 1985)”
- But, negative findings should give us pause.

The Need for a Critical Phase

- The development phase can yield strong evidence for validity, but it tends to be confirmationist.
- At some point, it is advisable to take a more critical stance, by identifying and empirically evaluating the most questionable assumptions and claims being made.
- Some issues that cannot be fully addressed during development (e.g., construct representation, generalizability, bias), because of limited sample sizes, should be rigorously evaluated.
- Many claims (efficacy, washback, criterion relatedness) cannot be realistically addressed until the assessment is used operationally.

Overall Evaluation of the Evidence for Validity

Questions to be Addressed During the Development Phase

- Is the assessment adequately described?
- Is the interpretation and use specified?
- Is the assessment plausibly related to the interpretation (e.g., in terms of sampling, theory, a research base)?
- Are potential sources of bias or irrelevant variance identified and considered?

Questions to be Addressed During the Critical Phase

- Are the scores to be reported generalizable enough, given the intended use?
- Are the most likely kinds of bias to be expected adequately addressed?
- Are any relationships (e. g. criterion-related or concurrent) inherent in the interpretation adequately supported?
- In general, are the claims based on the scores supported by adequate evidence?

Some Sleights of Hand and Fallacies (or my ‘Pet Peeves’)

“Begging the Question”
Cherry Picking Validity Evidence

- In Logic, to “beg the question” is to assume a large part of what is to be shown.
- In educational assessment, we beg a lot of questions:
  - A high alpha reliability indicates consistency over items; it says nothing about stability over occasions or contexts.
  - A high alpha or good model fit, in itself, says nothing about extrapolation to ‘real life’.
  - Sub-scores are not necessarily instructionally useful.
- In particular, the fact that an assessment is designed to achieve some goal does not imply that it will achieve that goal.

Straw-man Criticisms
Michell on Requirements for "Measurement"

- Michell (2008) has argued that psychometrics is a "pathological science", because "measurements" must satisfy Holder's Axioms (which describe the properties of physical quantities like length), and psychometric analyses do not check these assumption.
- Holder's axioms are essentially a very abstract, formal model for combining values of attributes:
  - E.g., Holder's eight axiom says that, (for any levels of the attribute, *a* and *b*, there is a level of the attribute, *c*, such that $c = a+b$).
- There is no reason to worry about Holder's Axioms as long as they are not part of the IUA, explicitly or implicitly.

Concluding Comments

Main Points 1

- In any project, it is good to know what you are trying to do and why you are trying to do it, or to figure this out early on.
- The argument-based approach to validity is intended to encourage this kind of self awareness.
- It is not intended to provide a set of instructions, or a recipe.
- In the development phase, the claims to be based on scores are specified as an IUA, and the assessment is developed, in an iterative process.
- In addition, potential threats to validity (e.g., sources of irrelevant variance) are identified and, to the extent possible, evaluated and removed.

Main Points 2

- In the critical phase, the most questionable inferences and assumptions are to be critically evaluated.
- Once we have an operational assessment in use, concerns about generalizability and bias can be evaluated more thoroughly (e.g., larger samples).
- Other issues like predictive accuracy and intended and unintended consequences can begin to be evaluated empirically.
- The I/U-specific models may be particularly useful at this stage, because they can identify the issues (e.g., common sources of bias) that need to be addressed

Thank You.

mkane@ets.org