Kane 2013: Validating the Interpretations and Uses of Test Scores

* Study focusing on validating a given ‘interpretation’ of a test score (e.g., evaluate how plausible claims of X are based on these scores.
* Argument based approach: Claims based on test scores should be outlined as an ARGUMENT that specifies the INFERENCES and SUPPORTING ASSUMPTIONS needed to get from score X to score based interpretation Y and use Z
  + Validation is an evaluation of the coherence/completeness of the argument for this interpretation and use.
  + Also, how plausible these inferences/assumptions are based on the information provided.
* Eight general points:
  + Proposed score interpretations and uses are what is validated, NOT the test or test scores itself (e.g., a broader generalization of the material)
  + The validity of a proposed interpretation/use depends on how well the evidence supports the claims made.
  + More ambitious claims need MORE evidence than less ambitious claims
  + More ambitious claims (e.g., construct interpretations) are generally more useful than less-ambitious claims (but harder to validate!)
  + Interpretations/uses can change over time based on new needs and new understandings. This can result in new evidence being needed for validation!
  + Evaluation of score uses requires evaluation of the CONSEQUENCES of the proposed uses.
    - Negative consequences can render a score use unacceptable.
  + Rejection of a score use does not necessarily invalidate a prior, underlying score interpretation.
  + Validation of the score interpretation on which a score use is based does NOT validate (more generally) the use of the score.
* Test scores often support claims that go beyond observed performances – these claims are NOT self-evident, and merit evaluation

Instead, we try a validation framework that focuses on test uses. Proposed uses are specified in an ASSESSMENT USE ARGUMENT (AUA). Use is validated by evaluating the plausibility of the AUA.

* Interpretation Use Argument (IUA) includes all claims based on the test scores (e.g., the network of inferences and assumptions inherent in the proposed interpretation/use)
  + Some focus IUA’s focus on a particular use, others involve interpretation in terms of a skill/disposition to behave in a given way, allowing for a range of uses.
* Validation of proposed interpretations/use of test scores can be defined by plausibility and appropriateness of the proposed interpretation/use.
  + A proposed use/interpretation can be valid insofar as the IUA is coherent/complete and it’s assumptions are highly plausible a priori or adequately supported by evidence.
* Less ambitious claims: Students w/ high scores on a test can GENERALLY perform the kinds of tasks included in the test.
  + Requires not much empirical support outside of data supporting generalizability.
* More ambitious claims: Inferences about some theoretical construct.
  + Requires more evidence (e.g., evidence evaluating the theory, and the consistency of test scores within the theory) supporting additional claims being made.
* If scores are used to predict future outcomes (employment/placement context) evidence indicating that these predictions are accurate are needed.
  + If no predictions are made… this evidence however is irrelevant
  + Moral Conviction predicts more difficulty with attitude change.
* Validity is a continuum, thus, uses that make sense and have evidence = high validity, but those who are not supported/contradicted = low validity.
  + Scores on a test can be valid for some uses, but not others!

Argument-Based Approach to Validation

* Have a fairly and complete statement of the claims being made, s/t what is to be evaluated is clear.
  + Develop an IUA that lays out the reasoning inherent in the proposed interpretations and uses of the scores. Specified as a network of inferences and assumptions leading from test performances to conclusions/decisions based on conclusions!
  + IUA is an explicit statement and the FRAMEWORK for validation!
* This lets us avoid the need for fully formed ‘formal’ theories needed for strong construct validity and avoid open-endedness/ambiguity for weak construct validity!
  + The evidence needed for validation is what is needed to evaluate the inferences/assumptions outlined in the IUA!
* Begin by specifying the IUA, proceed by evaluating the coherence/completeness of the IUA and the plausibility of it’s inferences and assumptions. Stop after these inferences/assumptions have been evaluated!

IUA

* The IUA plays the role of the formal theory in the strong program of CV.
  + To claim that a proposed interpretation/use is valid, is claiming that the IUA is clear, coherent, and complete, that it’s inferences are REASONABLE, and that it’s ASSUMPTIONS are plausible.
* Most tests require takers to perform some tasks (or respond to stimuli), and results are used to draw conclusions/make decisions.
  + Score the observed performances and combine the scores in some way.
    - ‘Scoring Inferences’ take us away from observed performances to an observed score. This assumes the scoring criteria is appropriate.
  + Note, the scores themselves aren’t seen as mattering too much, but instead we want to make claims about how this test taker performs in a larger domain of tasks, which involves a ‘generalization inference’.
  + We typically extend this interpretation further – using these scores to predict future performance in other contexts or draw inferences on how the test taker performs on different tasks in different contexts.
* Note: Inferences take the general form of ‘if-then’ rules.
  + If the scoring inference states that the observed performance has certain characteristics, it gets a certain score!
    - Then, extrapolation indicates that if a given score has a certain value, the criterion variable is expected to have some other value.
  + These are rules for making claims based on available data.
  + A presumptive inference establishes a presumption in favor of a claim, but doesn’t prove it, which may fail in a given case b/c of special circumstances.
    - E.g. if a regression equation is based on third graders, it’s applicability would be restricted to that population.
* Inferences in the IUA generally are qualified to various degrees, thus, the argument as a whole will be qualified!
  + Qualifiers applied to the IUA will depend on the qualifiers of the different inferences, and how these fit into the argument as a whole.
* The IUA provides an explicit statement of the reasoning inherent in the interpretations and uses of test scores – specifies the steps involved in getting from the observed test performances to the claims based on test scores.
  + Also, this includes an exposition of the intended uses of the test scores.

The Validity Argument

* This is an overall evaluation of the claims made in the IUA.
  + Although not ‘objectively prove-able’ the IUA can be evaluated with regards to clarity, coherence, and plausibility.
* Four Criteria for deciding on the studies to be pursued by test evaluator:
  + Is this issue genuinely in doubt?
  + How much uncertainty will remain at the end of a feasible study?
  + How expensive is the investigation in time/dollars?
  + How critical is the information for achieving consensus in the relevant audience?
* This approach is straightforward! STATE THE CLAIMS > EVALUATE PLAUSBILITY OF CLAIMS
  + E.g. Interpretation of math scores w.r.t. ability to solve similar problems as those on the test.
    - This is very limited, involves FEW inferences/assumptions, and thus only requires modest support for validation.
    - Interpreting these scores in terms of GENERAL mathematics aptitude and readiness for educational programs requires MORE inferences and assumptions, and thus requires MORE evidence/more KINDS of evidence in it’s validation.

Developing the Test, the IUA, and the Validity Argument

* We develop tests with a given purpose in mind – guiding the development of both the test and the IUA.
  + If the purpose to select for an educational program, it would be reasonable to begin by identifying some of the core skills and aptitudes associated with success in the program, and then develop a test that measures these attributes.
* Begin by getting a testing program, and developing an IUA that represents the proposed interpretation and use of the scores, and is CONSISTENT with the characteristics of the test.
  + If these assumptions are untenable, then modify the test/IUA/both.
  + E.g., if generalization is inadequate, you can lengthen the test, to improve generalizability
* Then, appraise the IUA, identifying and examining challenges that a skeptic may pose, especially identifying any HIDDEN assumptions or ALTERNATIVE explanations.
* Separate the specifications of the claims being made from the evaluation of the claims themselves!
  + Specification of an IUA is meant to lead to careful/complete specification of the proposed interpretation and use. Then, it is clearly useful to know what is being claimed, before evaluating these claims themselves.

Generalizations

* When interpreting test scores we typically generalize over the tasks included in the test/over test forms.
  + E.g. We don’t say, mary did well on form 26B of the reading test, we say, mary did well on the reading test.
  + We assume that the particular test form used does NOT matter, and the results would generalize over different tests!
  + We are also generalizing over other factors (test setting, time of day, etc.) that could ALL have impacted the test scores.
  + You should indicate how WIDELY the interpretation is to be generalized!
  + Don’t just assess cronbach’s alpha and assume that reliability/generalizability is addressed, be explicit in the use case!

Tailoring Validation to Proposed Interpretations

* Main points
  + Test scores can have multiple interpretations/uses – however… it is the proposed interpretation/use that is validated (not the test/scores itself)
  + Validity of proposed interpretation/uses depends on how well evidence supports the proposed interpretation/use.
  + More ambitious interpretations require more backing than less ambitious ones.
* Some attributes are defined as tendencies to perform/behave in some way, which are OBSERVABLE.
  + Observable s.t. they are defined by observable performances, not making any big claims about underlying traits or constructs.
    - They make the WEAK claim that performances of interest reflect some characteristic/cluster of characteristics by the test taker, but do NOT make any claims about what these characteristics are or how they function.

Assessment of Observable Attributes:

* Interpretation of observed scores on tests in terms of performance requires 3 main inferences
  + Scoring inference
  + Generalization from the observed score to ‘expected score over a universe of generalizations’
  + Extrapolation from the ‘universe score’ to the ‘target score’
* Note that the target domain is meant to reflect the attribute of interest and is not restricted to test items or test-like tasks!
* Ask how the test scores will be interpreted and used, what’s the population the test takers will come from, and the contexts in which this all will happen

Validating Interpretation w.r.t. Observable Attributes

* Scoring: Warrant for scoring inference is a rule/rubric that attaches scores to observed performance on test (e.g., mean of several items). The performance is the data, the score is the claim!
* Generalization: This inference treats the observed score as an estimate of the universe score (e.g., the average over all realistic generalizations of the test taker). The score itself doesn’t necessarily change… but interpretation is extended from claims about observed performances, to claims about expected performances in many circumstances. Takes the observed score as the data, and the universal score as the claim.
  + You have to make the case that this sample (which is not random) is representative enough in some sense that the generalization inference will work as intended.
    - You do this by making the sampling of tasks/conditions as representative of the ‘universe of generalization’ as possible.
    - Then, identify and eliminate any effects that would make the sample substantially unrepresentative of the ‘universe of generalization’
  + E.g Large standard errors or broad confidence intervals imply weak conclusions about the universe score!
* Extrapolation to the target domain:
  + If processes involved in the test tasks and other tasks are understood (no, in our case), and if the test tasks are the same process as most tasks in the target domain (absolutely not). Extrapolation is reasonable!

Additional Claims:

* IUA’s for observable attributes involve scoring, generalization, and extrapolation, but we can have other inferences!
  + Test performance in one time and context may want to predict performance on a different variable, in a different context, in the future.
  + The warrant for this – a regression equation that yields an expected score on a criterion for the target domain, based on an observed score on the test.
  + We use a ‘predictive validity study’ in which the relationship between test scores and the variable of interest is investigated for some sample of individuals.
    - The standard error of estimate for the regression equation provides a qualifier for this predictive warrant.
* Scale Based Inferences – Observed scores may be given a ‘norm-referenced’ interpretation w.r.t. the percentages of some population (e.g. 3rd graders in US public schools, Freshmen engineering majors at a given college, etc.)
  + The warrant for this inference would associate each test score w/ a percentage, and the backing for the warrant would be derived from data on the score distribution in the population.
* “Trait Hypothesis”: target domains are NOT defined arbitrarily, but can reflect the content of instruction, skills assumed to be needed in a context, or assumptions about underlying traits.
  + Those who perform in a certain way on tasks get the given scores.
  + E.g., those that can generally solve algebra, get high scores in algebra tests, and those who cannot, get low scores.

The Evidence Needed for Validation

* The validation begins by evaluating the IUA, then evaluating the plausibility of inferences/assumptions.
  + A test is valid for measuring an attribute if variation in the attribute causes variation in the test scores!
* We can design a test to reflect some underlying attribute and interpret scores in terms of that underlying attribute,!
  + Observable attributes describe consistencies in performance… but they do NOT explain these consistencies!
  + E.g., How do we determine whether a test is measuring what it’s supposed to measure, if we do not know what it is supposed to measure?

Theory Based Interpretation:

* Developing theories that account for test scores can enhance how useful the scores are. If the theory makes predictions about phenomena, these scores can be used, in conjunction with theory, to make predictions about phenomena!
  + E.g., blood pressure readings are useful b/c they are directly related to proper function in the heart and circulatory system.
* IUA’s for indicators of theoretical constructs:
  + IUA’s here can be developed backwards! The theory indicates that observations are appropriate for a particular construct, and the scoring rule that is appropriate.
    - E.g. individuals with HIGH values on a construct are likely to perform ONE WAY on some tasks, and those with LOW values on the construct are likely to perform ANOTHER WAY on the tasks, performance on these tasks can be used to develop an indicator for the construct.
  + Basically, indicators are designed to provide plausible estimates of the construct and to be relatively free of systematic and random error.
* The construct is defined in terms of roles in theory, the observable attributes are defined in terms of target domains of observations.
  + These attributes can be indicators, but their use does not eliminate any prior interpretation of the observable attribute in and of itself, it adds another layer of interpretation to the attribute.
  + However… a test may require some competencies (e.g., computer skills) that are NOT associated with the construct being estimated.

Causal Interpretation:

* Causal explanations can be justified in at least 2 ways:
  + Empirically – Introduction of the ‘cause’ consistently leads to the effect, or removal of the cause removes the effect
  + Theoretically – Establishing a well-confirmed theory that implies that the cause generally will yield the effect.

Score Uses

* Consequences for evaluating test-score uses
  + Evaluation of test score uses require evaluating the consequences of the uses, negative consequences can render a score use unacceptable
  + However… rejection of a score use does NOT necessarily invalidate prior underlying score interpretations.
  + Validation of score interpretation on which a score use is based, does not, in itself, validate the score use.
* Decision rules are evaluated in terms of expected success in achieving goals at a reasonable cost and acceptable consequences.
  + Test scores indicate something about the test takers, and decisions are made in accordance to those rules. Decision inference takes an interpreted score as it’s datum and yields a decision as it’s claim.
  + The decision rule is the warrant for the decisions, backing for this warrant consists of analysis indicating the appropriateness of the decision-rule given the purpose of the program.
* Decision Rules and Consequences
  + Trading off the benefits against the harms, if a satisfactory solution can be found while trading off against loss, then that is worthwhile.

Cook 2015: A contemporary approach to validity arguments: a practical guide to Kane’s framework

* How to practically apply the guidelines developed by M.T. Kane
  + Rigorous validation = articulating the claims and assumptions attached to the proposed decision, testing these assumptions, and organizing evidence into a coherent validity argument
  + Scoring: Translating an observation into one or more scores
  + Generalization: Using the scores as a reflection of performance in a test setting
  + Extrapolation: Using the scores as a reflection of ‘real world’ performance
  + Implications: Applying the scores to inform a decision or action