Research & Development

Listening. Learning. Leading.®

Validity, Fairness, and Testing

Michael Kane
Educational Testing Service

Conference on Conversations on Validity Around the World
Teachers College, New York
March 2012

Outline

- Validity
- Perspectives on Testing
  - Measurement perspective
  - Contest perspective
  - Policy/bureaucratic perspective
- Fairness
  - Equitable treatment
  - Consistency in score meanings
  - Fairness in test use

Validity

Validation

- To validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the proposed conclusions and decisions ... Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified. (Kane, 2006, p. 17)

The Goal is:

"To make clear, and to the extent possible, persuasive the construction of reality and the value weightings implicit in a test and its application."

Cronbach (1988, p.5)

The Argument-based Approach Employs Two Kinds of Argument.

- An ***interpretive argument*** specifies the proposed interpretations and uses of scores by laying out a chain or network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the scores.

- The ***validity argument*** provides an evaluation of the interpretive argument's coherence and of the plausibility of its inferences and assumptions.

Two-step Process

1. Specify the proposed interpretation and uses in some detail (e.g., as an interpretive argument).

2. Evaluate the overall plausibility of the proposed interpretations and uses (i.e, the validity argument).

Presumptive Arguments

Presumptive arguments draw conclusions about real-world claims, based on reasonable assumptions and warranted inferences.

They are more akin to general scientific reasoning than to logic.

The do not prove the conclusion in any absolute sense, but they can provide a strong presumption in favor of the claims (e.g., about a scientific theory) being made.

Toulmin's Model
Basic Version

- Toulmin (1958) developed a very general model presumptive inferences and arguments.
- According to Toulmin, an inference takes us from a starting point (a **datum**) to a conclusion (a **claim**).
- The inference is supported by a "rule of inference" (or **warrant**), which is supported by evidence (or **backing).**
- The model is quite general, but the data, claims, warrants, backing, etc. can be different for different fields and for different inferences within a field.

Toulmin's Model: Refinements

- Our level of confidence in a presumptive inference is often specified in terms of some **qualifier** (e.g., "usually", "generally").
- For example, in interpretive arguments, the qualifiers are often specified in terms of confidence intervals of some kind.
- Presumptive arguments can be challenged in specific cases, even if the general form of the argument is well established.
- So, they allow for exceptions.

Toulmin's Model of Inference

Datum ⇒ [*warrant*] ⇒ so {Qualifier} Claim

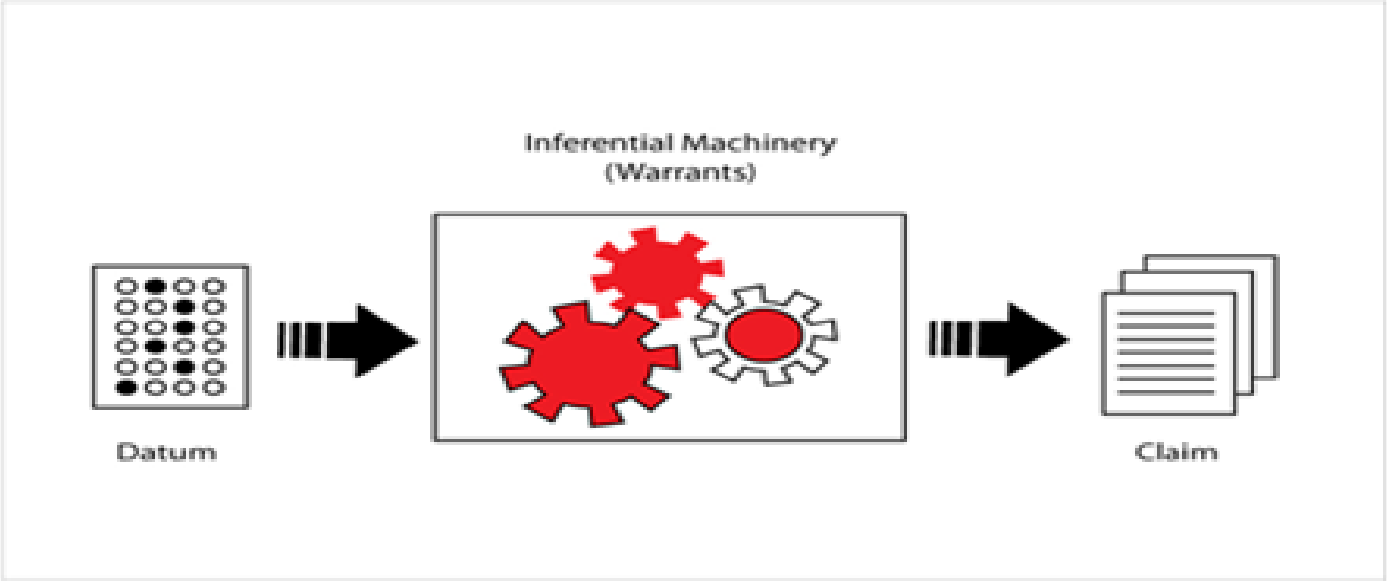
⇑ ⇑

Backing **Exceptions**

Toulmin's Model
A Quantitative Example

- Regression equations are commonly used to predict criterion scores from test scores.
- The test scores are the data and the predicted criterion scores are the claims.
- The regression equation is the warrant, and the empirical study used to derive the regression equation provides the backing for this warrant.
- The standard error of estimate is a qualifier.
- If the regression equation were developed in one context, its applicability to students in another context (e.g., in a different grade) might be questioned.

Warrants as Generic Inferences



7

Warrants and Backing

- The warrants are general if-then rules, and the backing provides support for the rule.
- Once the warrant is developed and supported by adequate evidence, it can be applied to a large number of specific cases.
- *Interpretive arguments* will generally involve a chain or network of inferences, with the claims in early inferences serving as the data in later inferences.
- The *validity argument* provides a critical appraisal of the coherence of the interpretive argument and of the warrants/backing for the inferences in this argument.

Interpretive Argument for a Math Achievement Test

- The *interpretive argument* specifies the interpretation and use of scores in terms of a network of inferences leading from observed performances to conclusions and decisions based on the performances.
- Scoring: from responses to score.
- Generalization: from observed score to universe score or "true" score.
- Extrapolation: from universe score to domain score.
- Abstraction: from domain score to construct/trait value.
- Decision: from domain score or construct value to decision.

Decisions

- Decisions are inferences that take us from an interpreted score to a decision.
- The warrant is a decision rule.
- The backing for a decision warrant consists of evidence indicating that the consequences of using the decision warrant are generally positive, and any negative consequences are tolerable.
- The decision warrant is likely to be the last span in the interpretive argument for a testing program.

Begging the Question

- Taking a part of the interpretive argument (particularly a questionable part) for granted
  - E.g., using content-based evidence to justify conclusions about aptitudes
- Begging the question of consequences
  - Assuming that a test that has been validated as a measure of some attribute is appropriate for some use.

Summary of the Argument-based Approach

- It is not tests that are validated, but interpretations and uses of test scores.
- The interpretive argument provides a starting point and a framework for validation by specifying what is being claimed and what needs to be evaluated.
- The validity argument provides an evaluation of the reasonableness or plausibility of the proposed interpretation and use of scores, by critically evalustaing the interpretive argument.

Three Perspectives on Testing

The Measurement Perspective

- Within the measurement perspective, the main concern is the accuracy of the test scores as estimates of the “real” or “true” value of the attribute/trait being measured.
- A second and equally salient concern is the precision of the scores, which is evaluated in terms of the variability in an individual’s scores over replications of the testing procedure.

The Contest Perspective

- The basic idea behind the measurement view is that a test *measures* something about the test taker. The basic idea behind the contest view is that a test that matters *is a contest* with winners and losers. …
- These two views can emphasize different goals.
- *Measurement view*: make a test 'reliable' and 'valid.'
- *Contest view*: make a test 'fair' and 'understandable.'

(Holland, 1994, p.28)

Policy/bureaucratic Perspective

- Decision makers in licensure and certification, college admissions, personnel management, and other contexts make high-stakes decisions for many candidates.
- Decisions have to be made efficiently, reasonably, and fairly.
- Such decision makers are likely to employ well-defined, systematic procedures to promote consistency and objectivity, where objectivity is defined in the negative sense of not being subjective, personal, or capricious.

T. Porter on Objectivity

- Scientific objectivity thus provides an answer to a moral demand for impartiality and fairness. Quantification is a way of making decisions without seeming to decide. (Porter, 1995, p.8).

- Mechanical objectivity … implies personal restraint. It means following the rules. Rules are a check on subjectivity: they should make it impossible for personal biases or preferences to affect the outcome of an investigation (Porter, 1995, p.4).

Fairness

Definitions of Fairness

- Equitable treatment
- Consistency in score meaning
  - Lack of bias is estimates of attribute value
    - Lack of Construct irrelevant variance
    - Lack of Construct under-representation
- Avoidance of unnecessary *adverse impact*
  - Adverse impact must be outweighed by positive benefits
  - Adverse impact should be kept to a minimum

Fairness and Perspectives on Testing

- All three perspectives support equitable treatment of test takers and consistency in score meanings.
  - The measurement perspective favors standardization as a way of controlling random and systematic error
  - The contestants want a “level playing field”
  - The policy/bureaucratic perspective favors objective procedures as a way of avoiding subjectivity and bias.
- The policy/bureaucratic perspective has the strongest interest in avoiding/minimizing adverse impact.
  - The measurement perspective sees this as a policy issue.
  - The contestant perspective has mixed views on adverse impact.

Cronbach vs. Messick 1

- Messick highlighted social consequences as an aspect of construct validity, but argued that:
  - If the adverse social consequences are empirically traceable to sources of test invalidity, then the validity of the test use is jeopardized. If the social consequences cannot be so traced ... then the validity of the test use is not overturned.
    (Messick, 1989, p.88)
- In this framework, adverse impact serves mainly as a red flag suggesting that bias may be a problem.
- The emphasis is on the technical properties of scores from a measurement perspective.

Cronbach vs. Messick 2

- In contrast, Cronbach (1988), maintained that consequences could invalidate test use, even if they could not be traced to any flaw in the test, because:
  - 'tests that impinge on the rights and life chances of individuals are inherently disputable" (p. 6).
  - 'Validators have an obligation to review whether a practice has appropriate consequences for individuals and institutions, and especially to argue against adverse consequences.' (p. 6)
- The emphasis is on how the program works in practice the perspective is more focused on policy.

Griggs v. Duke Power Co. (1971)

- Before 1965, all of the black employees at the Duke Power Company were in the labor department (the lowest salaried of its five departments).
- In 1965, as the Civil rights Act of 1964 was about to take effect, the Duke Power introduced two aptitude tests for employment in the upper level jobs.
- The testing procedure tended to exclude minorities, and no effort had been made to demonstrate any countervailing benefits.
- The Supreme Court found that the new requirements were “adopted ‘without meaningful study’ relating them to performance” (Guion, 1998, p. 190).

Need to Justify Adverse Impact

- The Duke Power decision established the principle that, if the decision procedure has adverse impact (i.e., a substantially lower "pass" rate for protected groups), it has to be counterbalanced by positive consequences (e, g., enhanced performance on the job) if the program is to be considered legally acceptable
  - (Pyburn, Ployhart, and Kravitz, 2008; Guion, 1998).

- As a corollary, subsequent decisions and the EEOC guidelines indicated the need to minimize adverse impact, when some adverse impact is justified.

Conclusions?

- The perspectives are not people. People can adopt multiple perspectives.
- It is not the case that any one perspective is better than the others.
- The problems tend to occur when an analysis based on one perspective is used to justify claims that fit a different perspective, thereby begging important questions.
- In particular, the fact that a test is psychometrically sound does not imply that any particular use of the scores is justified. Uses are evaluated in terms of consequences!
- In order to address fairness effectively, we need both the policy and measurement perspectives.

Thank You.

A Tale of Two Perspectives
Act 1

- MP: “This candidate’s score is only one point above the passing score, a small difference compared to the standard error of ten points, it is hard to say whether he should have passed.”
- PP: “What do you mean? His score is above the passing score, so he passed. Did you make a mistake in scoring his paper, or did something unusual happen during testing?”

Act 2

- MP: “There was no mistake, but the score is right next to the passing score, so if we repeated the testing, his chances of failing would be almost 50/50.”
- PP: “But, why should we repeat the testing? He passed and has no reason to test again! As a matter of fact, it would be against the rules for him to take the test again.”

Act 3

- MP: “But we don’t know his ‘true score’, and if he had taken a different form of the test or took it on a different day, he very well might have failed.”
- PP: “He took the form of the test that was administered on the most recent test date, and he passed. We have no reason to reconsider this decision.”