Perception-Based Evidence of Validity

Tzur M. Karelitz
National Institute for Testing & Evaluation (NITE), Israel

Charles Secolsky
Measurement and Evaluation Consultant

Can public opinion threaten validity?

- According to the 2014 PDK/Gallup survey:
  - 68% of public school parents do not think standardized tests help teachers know what to teach.
  - 61% of Americans oppose using standardized tests to evaluate teachers.
- The Opt-Out movement advocates refusal to take state-mandated tests, to protest the overuse and misuse of tests in school.
- In NY, 20% of the students opted-out of the 2015 Reading & Math tests. Students were:
  - more likely to be white and native English speakers
  - less likely to be economically disadvantaged
  - more likely not to have achieved proficiency last year

Face Invalidity

Face invalidity occurs when stakeholders do not perceive score-based inferences and actions to be appropriate.

- Face invalidity can negatively influence:
  - examinees' motivation to prepare and perform well
  - their willingness to take the test
  - the opinions of policy makers, public, media, judicial system...
    - Nevo (1985) and Messick (1989)

- Public opinion can influence decision makers who determine whether the test will continue as is, adapt to accommodate current concerns, or cease to exist.
- **So why do we dismiss Face Validity?**

Face Validity (FV)

A test is face-valid if it looks valid, particularly to layman. (Cureton, 1951)

- Validity by assumption: claiming a test is valid without statistical evidence, merely because it seems to relate to its purpose.
  - this practice "totally unscientific and indefensible" (Mosier, 1947)
- Appearance of validity: a test should not only be valid, it should also appear valid to stakeholders.
  - this is desirable from a practical sense, but it is not validity. (Mosier, 1947)
- Nevo's (1985) operational definition for FV:
  - A rater rates items or tests using relative or absolute judgments, as suitable or relevant for their intended use.

Criticism of Face Validity

- A test can seem valid without actually being valid. Therefore, **by itself**, FV shows no real evidence of validity.
- FV is regarded as the simplest and least scientific form of validity.
- In the 1974 edition of the Standards, FV was referred to as a "non acceptable basis for interpretive inferences from test scores."
  - The term is missing from all recent Standards and major text books.

Validity and Validation

Standards for Educational and Psychological Testing (AERA, APA, NCME, 2015)

- **Validity** is the degree to which evidence and theory support the interpretations of test scores for proposed uses of the test.
- **Validation** involves gathering evidence to:
  A. Support claims about particular interpretations of test scores
  B. Demonstrate that the proposed uses of test scores are appropriate
- **Evidence for validation** can originate from five sources:
  a. The test content
  b. The internal structure of the test
  c. The underlying response processes
  d. Relations to other variables
  e. The consequences of testing

Why the term *Face Validity* must die

- The definition of FV is simplistic, misleading and outdated:
  - It considers validity as a property of the test, and not as the appropriateness of test score interpretation and use.
  - It represents a perception about the test, whereas validity arguments are a complex logical chain of assumptions and inferences.
- The term has strong negative connotations, and should not be used to describe tests.

Why the concepts behind FV must live on

- FV stems from **stakeholders' perceptions of**:
  - The purpose of the test, and its ability to achieve it
  - The quality of the test
  - The way scores are interpreted and used
  - The consequences of using the test
- **Perception** is an interpretive process influenced by a variety of factors: past experiences, knowledge, beliefs, attitudes, etc.
  - Perceptions can influence subjective judgments and actions.
- Uncovering stakeholders' perceptions about the test can be useful for test development, validation and public relations.

Collecting data about perceptions

- We can collect perceptions about various aspects of the test as seen through the eyes of different stakeholders.
  - These can be collected via surveys, interviews, focus groups, feedback questionnaires, etc.
  - The usefulness of the data depends on the stakeholder level of experience and familiarity with the test.

General public

Examinees

Novice users

Perceptions about:
Test purpose, content coverage, item quality, measured attributes, score meaning and usage

Policy makers

Public figures

Test validators

Expert users

Subject matter experts

Test developers

Perception-Based Evidence (PBE)

- PBE can be useful for **identifying validity threats** and evaluating the sustainability of the test.
- PBE can be used for **generating alternative claims** about the interpretation and use of test scores.
- PBE can help **gain insights for interpreting validity evidence** collected from other sources (test content, response processes, etc.)
- PBE can be used to **evaluate the clarity and plausibility** of validity arguments.

Collecting PBE to generate alternative claims

- To evaluate the plausibility of a proposed argument, validity claims need be to juxtaposed against alternative claims (Kane, 2006, 2013).
- Perceptions of stakeholders are a good source for alternative claims.
  - Test developers have limited ability to generate real alternatives.
  - Non-experts can provide insights regarding construct deficiency or construct-irrelevant variance.
  - Researchers can **identify popular beliefs** about the test, **design studies** to compare these beliefs against the proposed claims, and use the results to **build a compelling validity argument**, and to **improve public relations** for the test.

Collecting PBE during the inception of a test

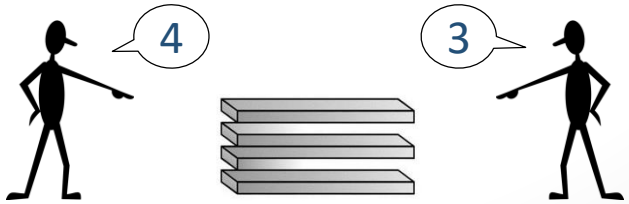
- Different stakeholders may have different perceptions about the purpose of the test, its design, or the constructs to be measured.
- Validation means ensuring that test scores are interpreted and used appropriately for their intended purposes.
- We need to evaluate evidence not only with respect to how test developers perceived these purposes to be but also with respect to how society did.

Perceptions & the five sources of validity evidence

Source of Evidence	Perceptions of:	
	Experts	Non-experts
Test content	Evaluate alignment to test specifications	Evaluate curricular alignment
Response processes	Design studies, evaluate performance & interpret results	Describe the testing experience
Internal structure	Design and interpret factorial analyses	Provide insights about underlying constructs
Relations to other variables	Design and interpret correlational analyses	Provide alternatives about the purpose of the test and its relation to other variables
Consequences of testing	Evaluate the effect of negative consequences on validity	Provide evidence of positive & negative impacts, unintended consequences, (mis)use & (mis)interpretation

Collecting PBE to evaluate the clarity and plausibility of the interpretive argument

- The validators' task is to evaluate the extent to which the interpretive argument is sufficiently clear, plausible, and coherent. (Kane, 2006, 2013)
  - The argument should be clear and plausible to everyone, not just the test developers!
- Validators could compare expert and non-expert perceptions regarding specific claims to identify points of agreement and disagreement.
  - Issues where everyone agrees show support for a strong argument.
  - Issues where perceptions differ are indicative of lines of argument where the claims are unclear or the inferences are not very plausible.



Concluding remarks

- The term *Face validity* should not be used. Still, perceptions are important because they influence many aspects of educational measurement.
- Perception-based evidence is useful for test development and validation, especially for constructing and evaluating validity arguments.
- Test developers should routinely collect, analyze, and report evidence based on the perception of various stakeholders about aspects of the testing system.