Rethinking Face Validity: Using Stakeholders' Perceptions for Validation

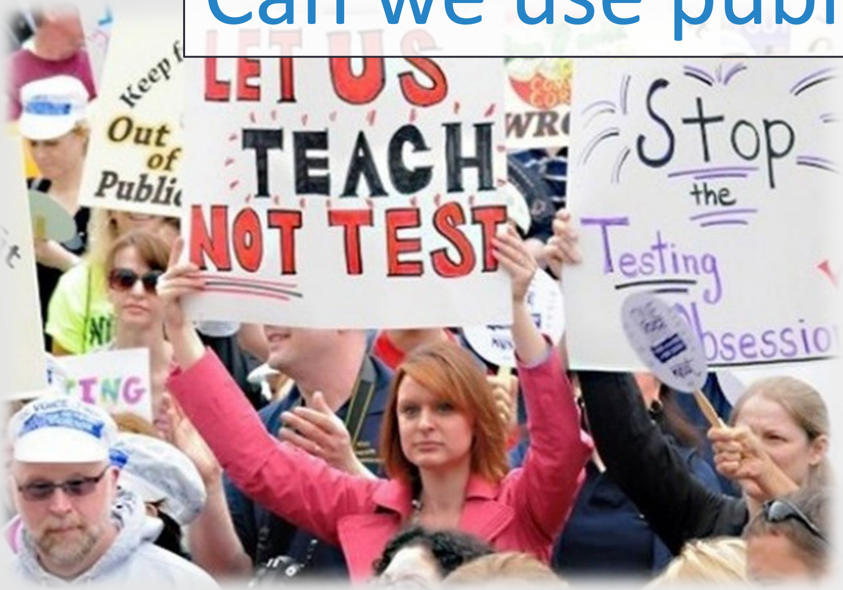
Tzur M. Karelitz

National Institute for Testing & Evaluation (NITE), Israel

6th IAEA Webinar, July 2021

Can public opinion threaten validity?

Can we use public opinion to improve testing?



YOU CANT TEST ME INTO COLLEGE
Testing ≠ LEARNING!
YOU CANT TEST ME IN TO COLLEGE!

Organization of this talk

- Two examples
- Face (in)Validity
- Validity, validation & justification
- What are stakeholders’ perceptions?
- Using stakeholders’ perceptions for test development & validation
- Concluding remarks

The Israeli MEITZAV assessment system

- MEITZAV Achievement Tests (5th & 8th grades):
  - Annual tests in 4 core subjects: First language (Hebrew/Arab), Math, English, Science & Technology
- In 2012, the Supreme Court ordered the MoE to make public all school results.
  - Test developers warned this will negatively impact the test's credibility & usefulness:
    - The ranking of schools based on test performance will put pressure on principals to quickly improve test results using inappropriate actions.
- ...which is exactly what happened in subsequent years
  - Media reports on cases of bad testing practices in schools caused a heated public debate:
    - Examples: Curriculum shrinkage, Massive drill-and-practice before the test, Removal of weak students on test day, False reporting of students' learning disabilities, Teachers dictating answers to students...
  - Principals and teachers felt hurt by the accusations of misconduct. The Teachers Association and organized parents groups called to boycott the test.
  - **The MEITZAV was terminated in 2018 and a new system is planned for 2022.**

U.S. College admission tests and the pandemic

- Before COVID, about 1,000 of the U.S. higher education institutions were **test-optional** or **test-blind**
  - Test optional- test scores (SAT or ACT) are not required for admissions but can be submitted
  - Test blind- test score are ignored even if submitted
- During COVID, many testing dates were canceled. In response, another 600 institutions dropped the testing requirements for 2021.
  These policies are in effect for ~**65% of B.A. institutions for fall 2022** (FairTest*).
  - Will cohorts become more diverse without losing academic quality?
  - Institutions begin to question the necessity of standardized test scores for their admission process.

*https://www.fairtest.org/1500-us-fouryear-colleges-and-universities-will-no

U.S. College admission tests and the pandemic

- In 2020, the University of California school system announced it **will not use test scores** for admissions until 2025, when a new testing system will be implemented.
  - This is the result of a lawsuit brought by students with disabilities and minority students.
  - The decision was made in spite of the recommendation of an expert committee, and a unanimous decision of the academic senate, to reinstate the tests after COVID.

Koljatic, Silva & Sireci (EM:IP, July 2021):

"We believe the legitimacy of admission tests will continue to be challenged until the testing industry adopts a new way of conducting their business to regain the goodwill of relevant stakeholders in society that so far have been largely ignored."

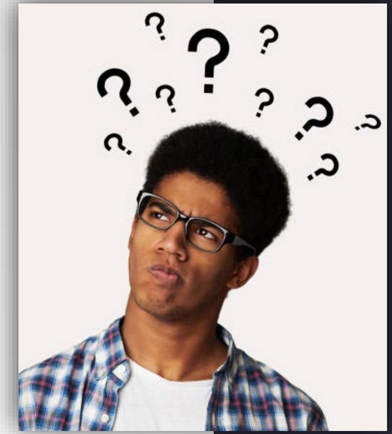
Face Invalidity

- Face invalidity occurs when stakeholders do not perceive test scores' interpretation and use to be appropriate.
- Nevo (1985) & Messick (1989)- Face Invalidity can **negatively influence**:
  - examinees' motivation to prepare and perform well on the test
  - their willingness to take the test
  - the opinions of policy makers, the public, the media, and the judicial system.
- The public can influence decision makers who determine whether the test will continue as is, adapt to accommodate criticism, or cease to exist.
- **So why do we dismiss Face Validity?**

Face Validity (FV)

- FV is a subjective judgment about whether the test seems to measure what it aims to measure.
- Rulon (1946) - Some tests are **obviously valid** because they cover all the relevant content or skills. In these cases, the test's FV is all the evidence we need.
- Cureton (1951) - "*A test is face-valid if it looks valid, particularly* ***if it looks valid to laymen***."
- Turner (1979) - FV is a more fundamental concept than construct validity. "*Some measures must be face valid in order for any measure to be construct valid.*"
- Nevo (1985) - an operational definition for FV:
  - A rater rates items or tests using relative or absolute judgments, as suitable or relevant for their intended use.
    Raters are non-experts: examinee, novice user, interested individual

Confusion regarding Face Validity



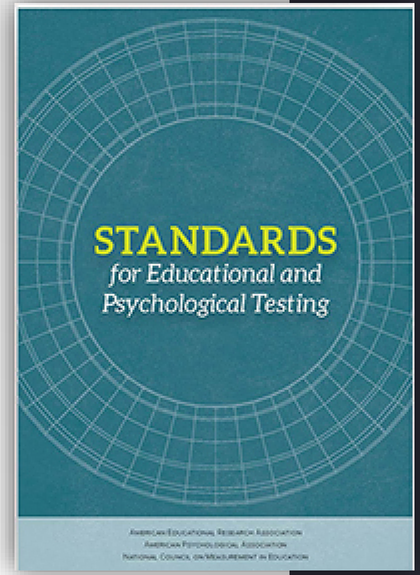
- Mosier (1947) identified that FV is used to mean different things:
  - Validity by assumption: claiming a test is valid without statistical evidence, merely because it seems to relate to its purpose.
    - this practice "*totally unscientific and indefensible*"
  - Validity by definition: when the test has complete content coverage.
    - this legitimate usage was probably the original intent of FV (i.e., obviously valid tests).
  - Validity by hypothesis: when a test is expected to be valid because it is similar to other tests that have been proven valid for the same purpose.
    - Used when there is an immediate need for a test, and validation will occur later.
  - 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.

Criticism of Face Validity

- FV is regarded as the simplest and least scientific form of validity.
  - Cureton (1951) : *“Face validity is often important in the public relations aspects… but as a validity concept it merely reflects* ***inadequate or superficial*** *analysis.”*
  - The Standards (1974): *“a non acceptable basis for interpretive inferences from test scores.”*
- FV is separated from other types of validity and cannot replace them.
  - A test can seem valid without actually being valid. Therefore, **by itself**, FV shows no real evidence of validity. The term is misleading because FV is not validity.
- The definition of FV is simplistic and outdated:
  - It refers to a notion of validity that is no longer supported.
  - Validity arguments comprise a complex logical chain of assumptions and inferences.
- The term has strong negative connotations, it **should not be used to describe tests**.

Validity and Validation

Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014)


STANDARDS
for Educational and
Psychological Testing
American Educational Research Association
American Psychological Association
National Council on Measurement in Education

- **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 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

Validation and Justification

- Cizek (2012, 2016) argues that modern validity confuses two separate and equally important endeavors that need different types of evidence:
  - **Validation** of test scores' meaning, interpretation, or inference
    - Evidence sources: content, structure, processes and relations
  - **Justification** of test use or actions based on test scores
    - Evidence of consequences of testing
- Evidence for validation and justification "*are not compensatory in any logical sense and cannot be combined into a coherent, integrated evaluation*."
- Test scores must be valid before we can use them.
  - Validation is a necessary but not sufficient condition for test use.
  - "*The validity of the test scores is typically unaffected by actions based on the test scores, the uses of the test results or the consequences of those uses*."

What are perceptions of stakeholders?

- **Perception** is an interpretive process that can influence subjective judgments and actions. It is influenced by past experiences, knowledge, beliefs, attitudes, etc.
- Relevant **Stakeholders** vary in their expertise: test developers, policy makers, content experts, test users, examinees and their families, the general public, etc.
- Stakeholders hold perceptions about different aspects of the test:
  - The necessity of a test for a specific purpose
  - The purpose of the test and its ability to achieve it
  - The coverage of content and desired attributes
  - The quality of test items
  - The way scores are interpreted and used
  - The consequences of using the test, etc.
- Perceptions can be collected using surveys, interviews and focus groups.

Why we must collect stakeholders’ perceptions

- Psychometricians do not pay enough attention to the societal implications of testing.
  - Sireci (2021) - *“We can talk about a lack of differential predictive validity and differential item functioning (DIF) ad infinitum, but if the adverse impact is so consequential it prohibits educational opportunities for a whole community of children, how can we justify use of the test for this purpose?”*
- Studying stakeholders’ perceptions can be useful for test development, validation and justification.
  - A vital piece of evidence for evaluating the consequences of testing.
  - Negative perceptions might damage validity, by influencing examinees behavior or affecting the way users and policy makers interpret and use the test.

Using stakeholders’ perceptions during the inception of a test

- We design tests based on what is perceived to be important or relevant.
  - Stakeholders may have different perceptions about the necessity of a test, its purpose, its design, or the constructs that need to be measured.
  - These perceptions should be documented in the conceptual assessment framework or rationale.
  - Discrepancies between stakeholders indicate possible vulnerabilities and caveats.
- Validation means studying whether test scores are interpreted and used appropriately for their <u>intended purposes</u>.
  - Test developers also need to consider how the intended purposes are perceived by other stakeholders: policy makers, test users and examinees.

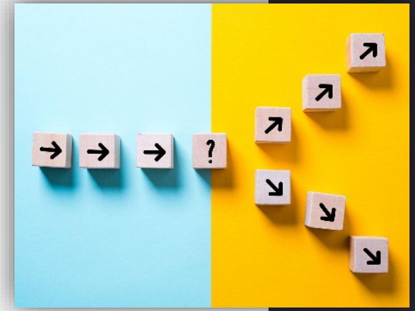
Using stakeholders’ perceptions for test development

- Subject matter experts review items for inclusion on a test, evaluate their alignment with content standards, or comment on their technical adequacy.
  - This input is based on their perceptions.
- During pilot studies, examinees provide input about the perceived difficulty, adequacy, clarity, and other properties of items or testing conditions.
  - This input is crucial for making the test appropriate, sensible and relevant, while maintaining the desired psychometric properties.

Using stakeholders’ perceptions for validation and justification

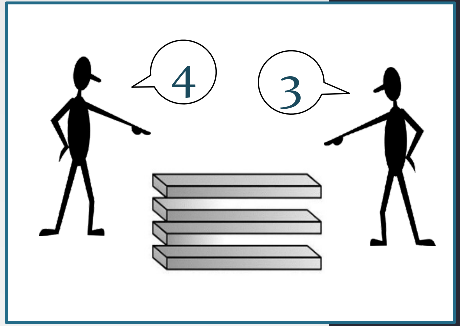
- Stakeholders’ perceptions can help identify gaps between intended and actual (perhaps unintended) interpretations, uses and consequences.
  - This input is useful for identifying validity threats, gauging the test’s positive and negative impacts and evaluating its sustainability.
- Stakeholders’ perceptions can be used to generate alternative claims about the interpretation and use of test scores.
  - This input might also help gain insights when interpreting validity evidence collected from other sources (test content, response processes, etc.)
- Stakeholders’ perceptions can be used to evaluate the clarity and plausibility of validity arguments.

Collecting stakeholders’ perceptions to generate alternative claims



- To evaluate the plausibility of a proposed argument, validity claims need be to juxtaposed against alternative claims (Kane, 2006, 2013).
  - *“The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it”* (Cronbach, 1980)
- Stakeholders’ perceptions are a good source for alternative claims.
  - They can provide insights about 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 more compelling validity argument.

Stakeholders’ perceptions about the clarity and plausibility of an 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 interpretive argument is the network of assumptions and inferences that underlie the proposed interpretations and uses of test scores.
  - The argument should be clear and plausible to stakeholders, 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.
- Stakeholders’ perceptions are important because they influence many practical aspects of educational testing.
- Stakeholders’ perceptions should be used for test development and improvement, validation of test scores, and justification of test use.
- Test developers should routinely collect, analyze, and report evidence based on the perception of various stakeholders about different aspects of the testing system.