NWEA
Northwest Evaluation Association

From Reliability to Validity:

Expanding Adaptive Testing Practice to Find the Most Valid Score for Each Test Taker

Steven L. Wise, NWEA

IACAT 2011

Monterey, CA

Traditional Conception of Validity

- A century of tradition in standardized testing
- Model: experimental design
- Standardization viewed as necessary to ensure validity
- Equal is equitable

CAT as the Exception

- One of the few examples of individualized testing.
- Item difficulty is tailored to each examinee.
- The intent, however, is increased efficiency.
  - Focus on reliability (reduced standard error)
  - Equivalence with paper & pencil tests is valued
  - Validity is enhanced through improved reliability

How Else Might We Individualize Testing Using CAT?

- By addressing construct-irrelevant factors influencing individual test scores (usually in negatively biased ways).
- Individual Score Validity (ISV) – how free is a particular score from construct-irrelevant factors (often called construct-irrelevant variance, or CIV).

An ISV-Based View of Validity

- **Test Event** -- An examinee encounters a series of items in a particular context.
- All 3 elements are potential sources of CIV.
- Examples:
  - Test anxiety (examinee)
  - Amount/difficulty of reading required (item)
  - Test stakes (context)
- ISV can be affected by all 3 elements.

An ISV-Based Approach to CAT

- In a given context, the most serious threats to ISV can usually be identified.
- Examples:
  - High-stakes: examinee test anxiety
  - Low-stakes: examinee test-taking motivation
- The nature and degree of validity threats will usually vary across examinees.
- **CAT Goal**: individualize testing to address CIV threats to score validity (i.e., maximize ISV).

Example: Low Test-Taking Motivation

- When given an item, a disengaged examinee will often exhibit a rapid item response. Such a response will tend to have an accuracy rate near chance (and far lower than 50%).
- A CAT can monitor—as the test is given—the speed at which examinees give responses and the accuracy of those responses.
- This is important information that a CAT can use.

Adapting a Test to Low Motivation

- **Change the types of items administered:**
  - Use items that require less reading
  - Use items that contain figures/graphs
  - Use items that don't tend to elicit rapid responses.
- **Intervene to preempt non-effortful behavior**
  - Display messages or warnings to examinee
  - Alert proctor

NWEA
Northwest Evaluation Association

Some Research Issues

- What are some innovative methods for expanding CAT that address ISV threats while preserving measurement of the target construct?
- How might CAT help address the ISV challenges posed by test anxiety?
- How should policy-makers deal with scores that have been shown to have low ISV?

Thank you for your attention.

Questions?

E-mail: steve.wise@nwea.org

NWEA
Northwest Evaluation Association