

Bias in Standardized Testing

WERA Conference | December 2018

Jack B. Huber, Ph.D.
Director of Data Management



Why do we have
standardized
tests?

Where did they
come from?

A brief history

Mid-19th century

- Written tests replace oral exams in schools

Turn of the century

- Call for common college entrance testing
- First intelligence testing
- Development of technology for mass scoring of tests
- World War I - testing in the military

Since 1965 - ESEA

- Tests for program evaluation

**What we value
enough to
measure with
standardized tests**

Achievement

Competency

Aptitude

Cognitive ability

Personality

Health (screening)

Career interests

Readiness for college

How we typically interpret test results in educational settings

What makes a score meaningful

1. “It’s higher than the rest of the group (i.e., 97th percentile).”

Referenced to norms: How a score compares to other scores in a larger distribution.

2. “It’s above/below the standard.”

Criterion-referenced: How a score compares to a single predefined cut score

How do we use tests in classrooms, schools and districts?

Formative uses

- Classroom data to inform next teaching steps
- Common planning next units to teach
- Screening students into tiers for intervention
- Diagnostic data to isolate individual student needs

Summative uses

- Declaring mastery of content (standards) for course grade
- Documenting mastery of knowledge, skills and abilities for certification

How do we evaluate the quality of tests?

Validity

What is the evidence to support how we're using a test?

Different kinds of validity evidence:

- Logical analysis
- Reliability
- Criterion-related validity
- Predictive validity
- Concurrent validity
- Construct validity
- Studies of scores and their consequences

Validity questions in practice...

How do we identify students for gifted programs?

Combination of three measures:

- Achievement: Level 4 SBA
- Cognitive ability: 97th+ percentile
CogAT V/Q/NV
- Creativity

What do we infer from these scores about the
mind and correct placement?

How do we go about this testing?

What tend to be the outcomes of this testing?

Why this is a validity question

Validity questions in practice...

How do we know when a student is ready for college?

The slate of evidence

- “College-ready” PSAT, SAT or ACT, etc.
- Grade point average

What do scores on these measures declare? Predict?

Why this is a validity question

Why we're here today...

Bias

- What is bias -- in a test/item, in scoring/grading?
- Is it possible to be free of bias?
- What can we do to minimize or manage bias in the development process?
- Is bias always undesirable?
- If we write better items over time, does that render older scores incomparable to new ones?
- How do we identify and deal with bias in practice?
- What evidence will satisfy the challenge?