Test and Sampling Designs in Support of Validity

Alina A. von Davier
Educational Testing Service

Educational Assessment, Accountability, and Equity: Conversations on Validity Around the World

March 28—29, 2012
Teachers College, Columbia University, New York
Goal

Traditionally, when speaking about minorities and subgroups, the focus is on content, test development and post-administration psychometrics (Kane, 2012; Chalhoub-Deville & Deville, 2008; Sireci, Han, & Wells, 2008; Geisinger, 2004; Cronbach & Meehl, 1955).

Here I propose the use of test and sampling designs to strengthen preemptively the validity argument for a heterogeneous population.

This approach addresses both the “measurement” and the “contest” aspects of test validity (Kane, 2012; Holland, 2008).
Outline

• Defining a target population (TP) for a test over time
• Validity challenges for a test with a shifting TP
• View an assessment as a sampling & matching matter: a sample of items & people
  • Choices for the test and sampling design for heterogeneous populations
  • Usefulness of a multi-stage adaptive test design
• Example from a global language test
A Target Population

• The TP is the population of test takers for whom the test results should be valid (Kane, 2012; Zumbo, 2007).

• TP is straight-forward for most tests.

• The distribution of the measured skill is unimodal (usually) for the test takers in the TP.
Validity Challenges

• The population of test takers might change over time and differ from the initial TP.
  - For example, this could happen due to an influx of linguistic minorities or when the test use is expanded to other groups.

• When subgroups of test takers have different (language) skill levels
  - the accuracy of the scores might decrease for the added group
  - the equating function might become group-dependent
  - the consequences of test use might not be equally fair for all
New Use for an Old Achievement Test: A Placement Test

• A global English linear test was administered to a polarized population of test takers for placement purposes:
  – to those who speak English to some degree
  – to those who do not speak any or speak very little English.

• The test is long and very reliable, so it does contain some very easy items

• Test takers with very low levels of English skills tend to perform less predictably on the common items and guess more often on all test items, thus endangering the quality of equating.
Challenges for Test Validity

- There are not enough discriminating items at the lower end of ability scale.

- The new subgroup is almost as large as the old subgroup, although it has a variable samples across administrations.

- The initial target population consisted of those who had (some/good) English skills.

- Test equating is group dependent.
Bimodal Test Score Distribution:
Observed, IRT –fitted (r), Log-linear pooled (b) and log-linear MDSG (g)-- (Duong & von Davier, 2011)
The Difference between Equating Functions for the Most Able Group and All (Duong & von Davier, 2011)
Consequences

• The placement of test takers into learning groups might not be accurate at all cut-score points of interest (especially for those with scores between 40 and 64)

• The opportunities for efficient learning as well as those for career advancement might be in jeopardy.
Dilemma: What to Do?

• To use an adaptive test, a CAT (too difficult to implement for the test users of this test)

• To use a multi-stage adaptive test (MST)
Potential Solutions: Test Design

- Choose an MST that is targeted to the bimodal population
- Investigate several MST designs with two levels
  - von Davier & Duong (in preparation)
Routing Module
Medium Difficulty
Large Variance

Easy Difficulty
Medium Difficulty
High Difficulty

X in T

Score

Frequency

200
150
100
50
0

0 10 20 30 40 50 60 70 80 90 100
Potential Solutions: Sampling Design

• Consider designing a sample scheme for analyses, such as on optimal sample

• Consider assigning weights to various subgroups for measurement, calibration, and equating
  • Qian, von Davier, Jiang (submitted).
Related Research

• Social impact moderation using item-test design and automated test assembly in CAT
  – Stocking, Jirele, Lewis, Swanson (1998)
  – Stocking, Lawrence, Feigenbaum, Jirele, Lewis, van Essen (2002)
  – Luecht (1998)

• Strengthening construct validity and validation of test use using cognitive models (Mislevy, 2012; Embretson & Gorin, 2001; Graf & Fife, in press).
Thank you!

Contact Information:

avondavier@ets.org