Perspectives on the Evolution and Future of Educational Measurement

Robert L. Brennan
Center for Advanced Studies in Measurement and Assessment (CASMA)
University of Iowa
Outline

• Measurement Theory
  Validity
  Reliability
  Item Response Theory
  Equating, Linking, and Scaling
  Need for Integration of Models
• Measurement Practice
  Validity and Consequences
  Reliability and Standardization
  Accountability in K-12
  Technology and Computerization
  Legislation, Litigation, and Notions of Fairness
Measurement Theory

• *Educational Measurement* (4 editions)
  – Lindquist (1951)
  – Thorndike (1971)
  – Linn (1989)
  – Brennan (2006)

• *Standards for Educational and Psychological Testing* (five editions)
Validity

• Cureton (1951): “Validity is therefore defined in terms of the correlation between the actual test scores and the “true” criterion scores.” (pp. 622–623)

• The Trinitarian Model (content, criterion, construct)
• Centrality of Construct Validity (Cronbach, 1971)
• Validity as an Integrated Evaluation (Messick, 1989)
• Validity as Argument---Validation (Kane, 2006)
Reliability

Constant Definition: reliability refers to consistency of scores across replications.

• Early 1950s:
  – Thorndike’s (1951) influence on Lindquist and Cronbach?
  – Coefficient Alpha.

• Mid 1950s to mid 1970s
  – Development of Generalizability Theory (Cronbach et al.)

• Mid 1970s to Present
  – Feldt and Brennan (1989)
  – Haertel (2006)
Item Response Theory

- Nothing in first two editions of *Educational Measurement*
- First referenced in *Standards* (1985)
- Only two standards refer explicitly to IRT in *Standards* (1999)
Equating, Linking, and Scaling

- Flanagan (1951, pp. 750–760) provides an integrated discussion of equating and linking
- By the 1980's, scaling, equating, and linking became much more public matters
- Petersen, Kolen, and Hoover (1989, 3rd ed.)
- Kolen (2006, 4th ed.)
- Holland and Dorans (2006, 4th ed.)
Need for Integration of Models

• Examples:
  – Many equating methods do not even recognize true scores
  – G theory (macro); IRT (micro)

• Issues
  
  What Constitutes a Replication?  Fixed/random
  What are True Scores?
    Expected/Platonic; 1PL vs 3PL
  What is Error?
    IRT has no error term; multiple sources of error
  No “Right” Model
Measurement Practice
(Trends or “Boundary” Issues)

• Validity and Consequences
• Reliability and Standardization
• Accountability in K-12
• Technology and Computerization
• Legislation, Litigation, and Notions of Fairness
Validity and Consequences

• Validity has to do with the proposed interpretations and uses of test scores
• Intended positive consequences:
  first call on test-developer resources
• Intended negative consequences
• Unintended positive consequences
• Unintended negative consequences:
  Who is responsible???
E.G. Use of ACT and SAT to rank order states
Reliability and Standardization

- Standardization: fixed conditions over replications
- More fixed facets yields higher reliability
- Tailoring the testing experience, or the decision about a tested examinee, to personal characteristics of the examinee--- Inconsistent with standardization
Accountability in K-12 (Dramatic change)

• Some History:
  – NCLB: Chapter/Title I of ESEA
  – NAEP: group-level reporting (BB/B/P/A)

• Testing Stakes: Low (past) to high (now)
  – High stakes usually drives curricula
  – High stakes can have a corrupting influence
  – High stakes can change measurement char’s

• Currently, standards are ambiguous
Technology and Computerization (Considered in all editions of EM)

• Test Administration: CBT and CAT
• Item and Test Scoring
  – Optical Scanning (Lindquist)
  – Computerized grading of essays
  – Simulations
• Score Reporting
  – Too late
  – One size fits all
  – Generally poor diagnostic information
• Unintended Consequences
  – Technology dictates what is measured?
  – State of the art vs. excellent measurement
Legislation, Litigation, and Notions of Fairness

• An example: Flagging
  – Breimhorst vs. Educational Testing Service

• Role of the Standards in Legislation and Litigation
  – Not as much as profession might desire
  – No enforcement mechanism
  – Tension between individual rights and professional responsibility

• “Unintended” Consequences of Increased Litigation
  – Increased cost
  – Decrease in number and/or quality of validity studies

• No universally accepted “gold standard” for issues of fairness
Concluding Comments

• Two trends
  – Increasingly sophisticated measurement models operationalized in “black boxes”
  – Social issues are so influential in testing that the boundaries between the two are often blurred

• Measurement professionals should serve as “honest brokers” in supporting the aggressive pursuit of validation efforts.