Psychometric Issues, Opportunities, and Challenges for Next Generation Assessments

Richard J. Patz
Robert L. Linn

In the invitational webinar series
Performance Assessment for the Next Generation of State Assessment Programs

Hosted by CTB/McGraw-Hill

October 28, 2010

One hour session

- 45 minutes presentation and commentary
- 15 minutes for questions, comments, and responses

- Please use the chat functions in WebEx to send questions and comments to me
- I'll use them for the Q&A session at the end

Introductions

- Rich Patz, VP for Research and Product Development, CTB/McGraw-Hill
  - Publications on statistical methods, vertical scaling of educational tests, standard setting, and assessment design
  - Research on educational measurement, statistics, and large scale assessment

- Bob Linn, Distinguished Professor Emeritus of Education Research, University of Colorado, Boulder
  - More than 250 journal articles and book chapters: wide range of theoretical and applied issues in educational measurement
  - Research explores the uses and interpretations of educational assessments; emphasis on educational accountability systems
  - National Academy of Education, VP of AERA Division of Measurement and Research Methodology, Past President of the National Council on Measurement in Education, past editor of the *Journal of Educational Measurement*, editor of the *Educational Measurement* (3rd ed.)

Psychometric Issues, Opportunities, and Challenges for Next Generation Assessments

Outline of Topics

- Psychometrics: What do we mean?
- Next Generation Assessments
  - Why a new generation?
  - Comprehensive Assessment Systems under RTTT
  - Other innovative assessments
- Psychometrics for Next Generation
  - Particular challenges
  - Opportunities & frontiers

Psychometrics

CTB McGraw-Hill

- Psychometrics: the design, administration, and interpretation of quantitative tests for the measurement of psychological variables such as intelligence, achievement, aptitude, and personality traits
- Focus here: Tests for measuring achievement of educational objectives, to inform education policy and practice
- The basics
  - Validity
  - Reliability
  - Item response theory, scaling & equating
  - Classical test theory, diagnostic, other models

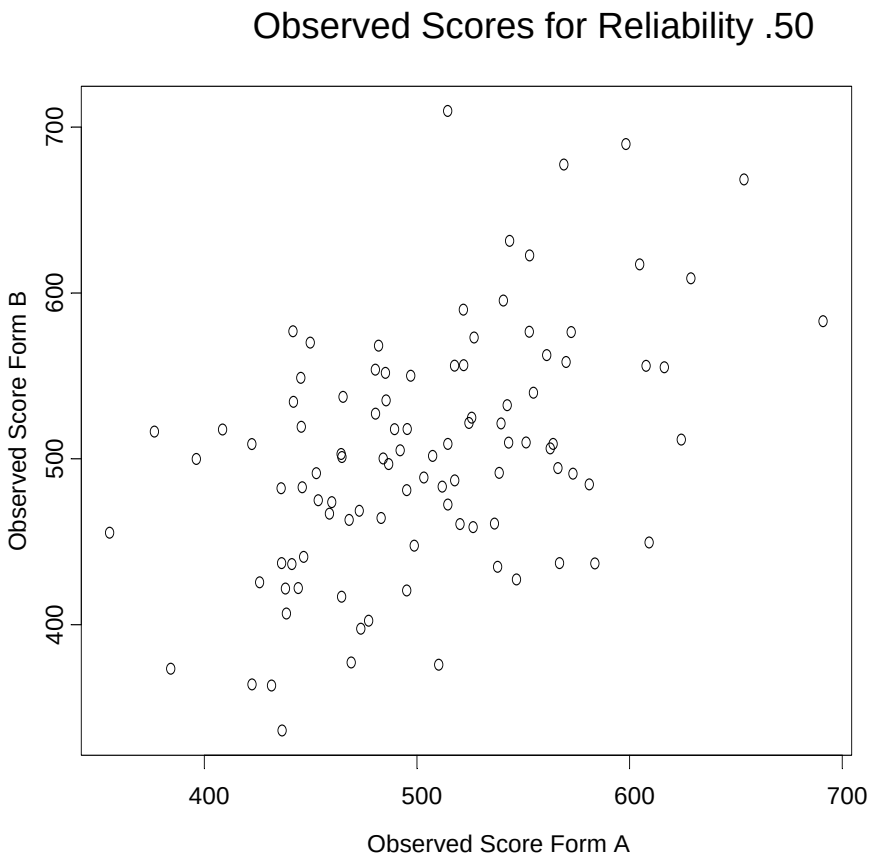
Validity

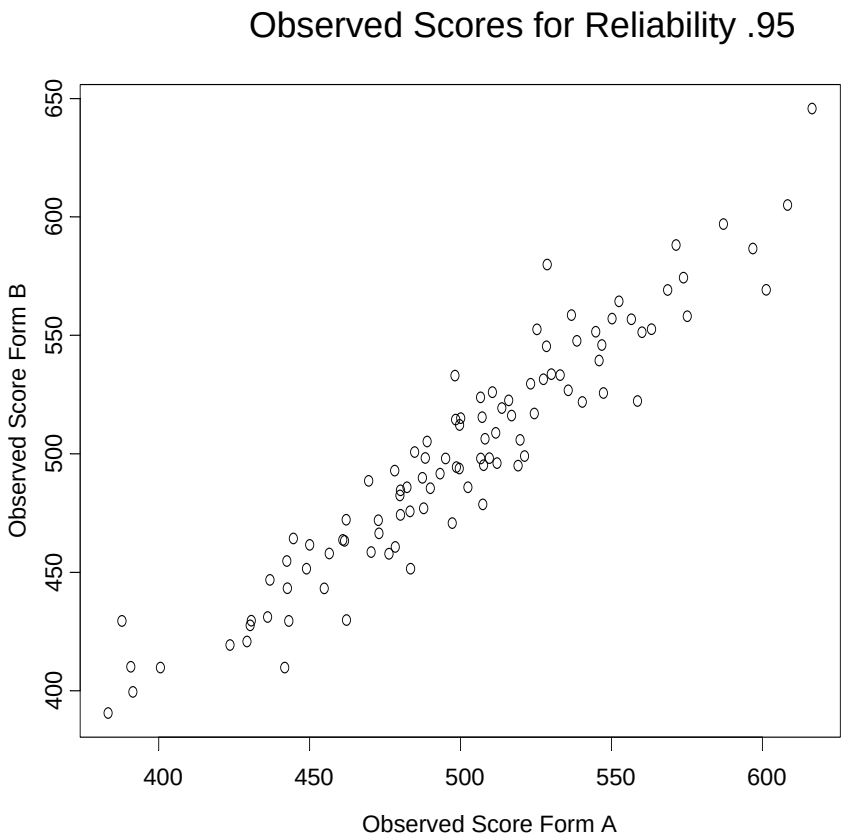
- The most fundamental consideration in developing and evaluating tests
- The degree to which evidence and theory support interpretations of test scores for proposed uses of tests
- The degree to which the test measures what it is intended to measure
- Test validation requires collection of evidence and presentation of an interpretive argument (Kane 2007)
- Important types of validity: Content, construct, criterion, consequential

Reliability

- *Consistency* of measurements when repeated on a population
- Quantifies *measurement error*
- Required for validity (but not sufficient alone)
- Some types of reliability statistics:
  - Correlations between repeated measures
  - Internal consistency reliability indices (KR-20)
  - Standard error of measurement (SEM)
  - Inter-rater agreement rates
  - Classification consistency rates

High and Low Reliability (test re-test)





Low Reliability

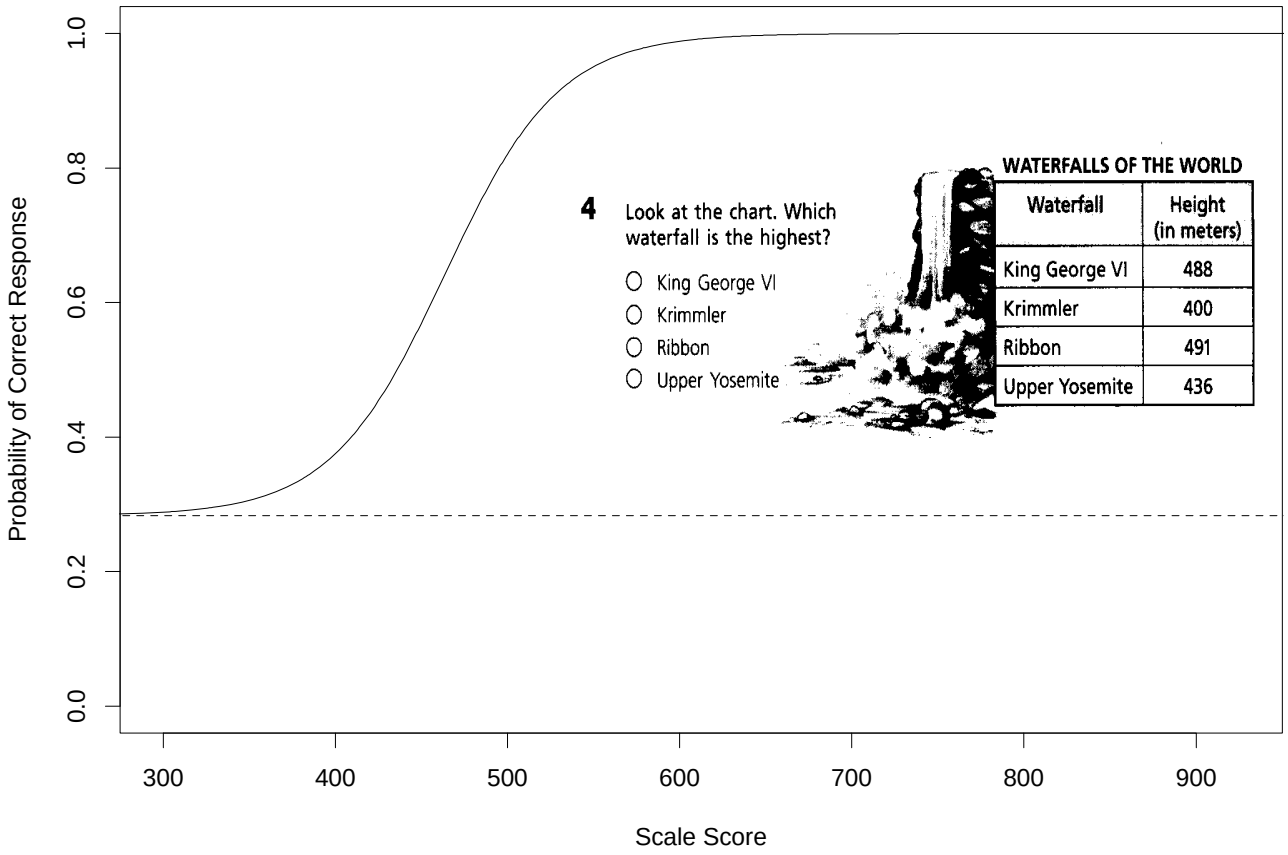
High Reliability

Item Response Theory (IRT)

- Places test items and examinees on a common scale
- Characterizes how easy or difficult particular items will be for particular examinees
- IRT models have strong assumptions, must be fit to data statistically, validated
- IRT has dramatically changed psychometric practice in past 50 years, and continues to evolve

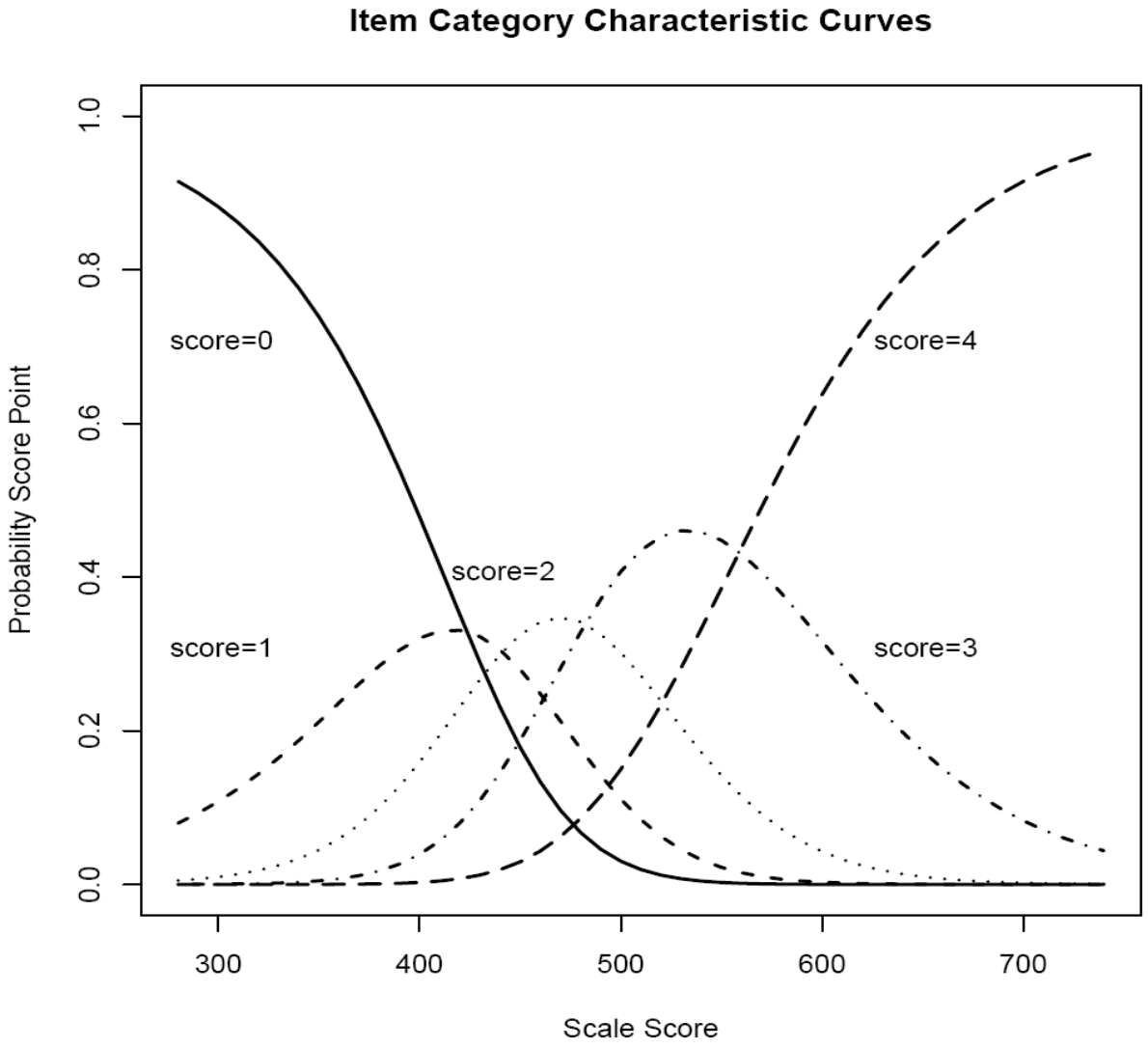
IRT Item Characteristic Curve (ICC)

Item Characteristic Curve (ICC)



3-parameter logistic (3PL) model

ICC for Performance Tasks/CR items



Generalized Partial Credit Model (GPC or 2PPC)

An IRT scale for mathematics:

300 400 500 600 700 800 900

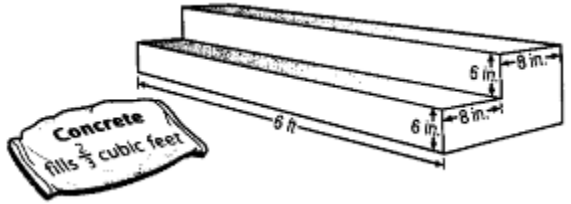
Scale Score

4 Look at the chart. Which waterfall is the highest?

- ○ King George VI
- ○ Krimmler
- ○ Ribbon
- ○ Upper Yosemite

WATERFALLS OF THE WORLD

Waterfall	Height (in meters)
King George VI	488
Krimmler	400
Ribbon	491
Upper Yosemite	436



3 How many bags of concrete mix will be needed to build this set of stairs?

- ✓ **A** 9 bags
- **B** 6 bags
- **C** 4 bags
- **D** 13 bags

Item Response Theory

- Assumptions
  - Underlying unidimensional trait (ability, knowledge, etc.)
  - Responses to items are independent given the trait
  - Probability of correct response increases with trait in a particular way
- Different IRT models specify different forms of this relationship
  - One-parameter or Rasch model
  - Two-parameter model
  - Three-parameter model
  - Multidimensional item response theory models

IRT is critical tool in modern psychometrics

IRT enables:

- Reporting scores along a scale
- Vertically scaling forms of intentionally different difficulty (e.g., across grade levels)
- Scaling multiple-choice and constructed response items together
- Equating forms using non-equivalent anchor test (NEAT) and other designs
- Assembly of equivalent forms from item banks
- Adaptive testing
- Content-focused standard setting (i.e., Bookmark)
- Item-pattern test scoring

- Note: Other approaches exist for many of these purposes, but IRT has provided a coherent framework for addressing them all

IRT has limitations: simplistic student model, doesn't handle complex content structure well, can provide accurate but limited information

Other Psychometric Models & Tools

CTB McGraw-Hill

- Classical test theory
- Generalizability theory (Cronbach et al)
- Cognitively diagnostic models
  - Bayes inference networks (Mislevy)
  - Rule space models (Tsutakawa)
  - DINA, NIDA models (Junker & Sijtsma)
  - Knowledge Space (Falmagne)
- Multidimensional IRT (Reckase)
- Tools: classical stats (p-values, pt. biserials), differential item functioning (DIF), more

Next Generation Assessments

What's Wrong with the Current Generation?

- High variation in quality across states
  - Standards: incoherent, too numerous, too narrow, not grounded in learning science
  - Technical quality: weak equating designs, field test practices, departures from known best practices
- Accountability framework
  - Status focused
  - Creates incentives for low standards
  - Penalizes diversity
- Fifty different state programs inherently inefficient
- No comparability to external indicators (state to state, state to other nations)
- US falling behind on international assessments as education reform accelerates overseas
- Weakens US economic competitiveness

Common Core State Standards

CTB McGraw-Hill

- Initiative led by states through CCSSO and NGA
- Fewer, deeper, and higher standards
- English Language Arts and Mathematics
- Prepare students for college and careers
- Progress across K-12 grades in coherent fashion
- Reinforced through federal initiatives (RTTT)
- Very different than most current state standards!

Adoption of Common Core State Standards

- The District of Columbia and 39 states have adopted the CCSS, including:
  - Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Illinois, Iowa, Indiana, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Mississippi, Missouri, Nevada, New Hampshire, New Jersey, New York, North Carolina, Ohio, Oklahoma, Pennsylvania, Rhode Island, South Carolina, Tennessee, Utah, Vermont, Washington, West Virginia, Wisconsin, and Wyoming.
- Common curriculum frameworks and revised state curricula under development
- Significant leadership changes of key players at large districts and states

Race to the Top (RTTT)

- Competitive grants from USDOE
- States compete for special funds
- Adoption of CCSS increases RTTT competitiveness
- Includes $350M fund for state consortia to build common assessments of CCSS

SMARTER Balanced Assessment Consortium (SBAC)

- Coalition of 31 states including AL, CO, CT, DE, GA, HI, IA, ID, KS, KY, ME, MI, MO, MT, NC, ND, NH, NJ, NM, NV, OH, OK, OR, PA, SC, SD, UT, VT, WA, WI, and WV
- Strong focus on computer adaptive technology
- One test at the end of the year for accountability purposes
- Heavy emphasis on extended performance tasks/events and innovative item types
- A series of interim tests to track progress towards the standards
- Portal for classroom formative materials

Partnership for Assessment of Readiness for College and Career (PARCC)

- Coalition of 26 states including AL, AR, AZ, CA, CO, DC, DE, FL, GA, IL, IN, KY, LA, MA, MD, MS, ND, NH, NJ, NY, OH, OK, PA, RI, SC, and TN
- Replaces once a year summative with multiple through-year summative assessments
- Includes complex text, research projects, classroom speaking and listening assignments, and work with digital media
- Computer-administered vs. computer-adaptive
- Formative/classroom assessments handled at the district level

Assessment Design Elements: SBAC & PARCC

- Theory of action: System of assessments support preparation of students for college/career
- Rich performance tasks
- Computer administered (PARCC), adaptive (SBAC)
- Greater depths of knowledge
- Interim/benchmark assessments (SBAC), support for formative assessment practices
- Through-course assessments (PARCC)
- Artificial intelligence (AI) scoring to degree possible

Psychometric Approaches in SBAC & PARCC

CTB McGraw-Hill

- Committed to industry best practices, adherence to technical standards
  - Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999)
  - 3PL IRT (SBAC), but look at multidimensional models
- Validating cognitive model (learning progressions) esp. important
- Validating alignment of content to standards
- Validity, fairness, accessibility of paramount importance
- Strong technical advisory committees proposed
- No radical changes/departures in technical procedures for PARCC and SBAC summative assessments (the seismic changes are other areas: content, delivery, technology, logistics)

Psychometric Challenges: SBAC & PARCC

CTB McGraw-Hill

- Establishing and maintaining comparability
  - Throughout year
  - Across grades
  - Across states
- Scoring performance events reliably
- Validating AI-based scoring algorithms
- Adaptive testing with performance events (SBAC)
- Validating tests that are not curriculum agnostic
- These are manageable (esp. compared to other aspects: content/curriculum changes, logistics)

Psychometric Issues in Performance Assessment

- Reliability
  - Raters within administration
  - Raters across administrations
  - Test reliability & test length
  - Information from multiple raters
  - AI-based ratings vs. human raters
- Other Validity Concerns
  - Coverage of Content Standards
  - “Score-able Rubrics”
  - AI algorithm changing construct, being coachable

Psychometric Issues in Performance Assessment

- Solutions exist to many challenges
  - E.g., Maryland School Performance Assessment Program
  - Yen & Ferrara (1997)
- Rater quality monitored
  - See also Hoskens & Wilson (2001)
- Cross-administration rater drift adjustments
  - See also Tate (2003), MSPAP technical reports
- Matrix item sampling covers content domain
  - See also NAEP
- Multiple Ratings
  - Wilson & Hoskens (2001)
  - Bock, Brennan, & Muraki (2002)
  - Patz et al (2002)
- AI for summative assessment
  - West Virginia WESTEST 2 Writing Assessment (Technical Report online)

Uniform application of best practices will constitute a big improvement

See Operational Best Practices for Statewide Large-Scale Assessment Programs. (CCSSO, 2010)

Adaptive Testing & Automated Test Assembly

CTB McGraw-Hill

- Automated test assembly (ATA) provides robust optimization approach to meet psychometric and content constraints when assembling test forms (van der Linden, 2009)
- Computer adaptive testing (CAT) performs optimal assembly in real time as examinee proceeds through test, enabling shorter tests (e.g., Wainer, 2000)
- SBAC will employ CAT
- SBAC and PARCC could employ ATA fruitfully

Other Assessments in Next Generation

- Look for innovation also outside SBAC & PARCC
  - States managing transition to CCSS (anchor/audits)
  - Assessment informing optimal instructional delivery and resource assignment
  - Assessments employing natural language recognition (speech, handwriting)
  - Assessments embedded in games, virtual reality
  - Better intelligent tutoring systems
  - More interactive assessment in classrooms
- Low stakes environments conducive to innovation in technology and psychometrics

Summary: Psychometrics for Next Generation

CTB McGraw-Hill

- Old/Enduring
  - Validity evidence
  - Reliability standards
  - Item response theory scaling and equating
  - Standard setting
  - Application of best practices from across states
- New/Expanding
  - More use of evidence-centered design principles
  - More adaptive testing (SBAC), automated assembly
  - Artificial intelligence (AI) scoring
  - Greater focus on growth
  - Curriculum-embedded and through-course assessments
  - More reasonable accountability framework?

Psychometric Considerations for the Next Generation of State Assessments

Robert L. Linn

CRESST, University of Colorado at Boulder

Comments in response to presentation by Rich Patz as part of the webinar series "Performance Assessment for the Next Generation of State Assessments, October 28, 2010

Next Generation

- Expected to be aligned with Common Core State Standards.
- Both the Smarter Balanced Assessment Consortium (SBAC) and the Partnership for the Assessment of Readiness for College and Careers (PARCC) plan to make heavy use of computer administered tests.
- Both SBAC and PARCC plan to increase the use of performance assessments.

Computer Administration

- Simply administering tests by computer as is planned by PARCC raises few psychometric issues.
- More issues are raised by adaptive testing which SBAC plans to use.
  - Content coverage must be assured along with psychometric information in selecting items.
  - Out-of-level test items may be used in adaptive test but were not allowed under NCLB.

Focus on Performance Assessments

- Called for by PARCC and SBAC, but are rare in current state assessments.
- Performance assessments lead to some psychometric challenges.

Why the Renewed Emphasis on Performance Assessments?

- Other countries with higher achievement in international comparisons make much greater use of performance assessments.
- Belief that Performance assessments encourage teaching and learning of important educational outcomes.
- **PARCC** plans to include performance tasks as part of their “through-course assessments”
- **SBAC** plans to include 1 performance task in reading, 1 in mathematics, and 2 in mathematics as separate events.

Beliefs About Performance Assessments

- Some things measured by performance assessments cannot be measured by multiple-choice or short constructed response questions.
- Tradeoff between validity and reliability
  - Higher validity comes at price of lower reliability

Why Performance Assessments?

- Validity is the most important psychometric consideration
- Belief that performance assessments can enhance validity through increased fidelity to the real-world task of interest
  - Problem solving
  - Conceptual understanding
  - Inquiry skills in science
  - Trouble shooting

Reliability

- Next to validity, reliability is a major psychometric concern
- Increasing number of tasks - a primary way of enhancing reliability.
- Performance assessment literature
  - 6 to 12 tasks may be needed to reach desired level of reliability.

Number of Performance Tasks

- The needed number of tasks for obtaining reliability higher than is likely to be practical for state assessments.
- Alternative approach is to combine one or two performance tasks with other modes of assessment (e.g., multiple-choice) to achieve reliable composite score.
- The SAT Writing Test uses multiple-choice questions to bolster the reliability of the Writing Score.

Comparability

- Scores of examinees taking different tasks, or the same task at different times should be comparable
- Comparability central to ensuring validity and fairness
- Same task given at different times may not be comparable due to communication between examinees

Recent History in K-12 Education

- Performance assessments were popular in a number of state programs in the 1990s
- Have largely disappeared from state assessments for two reasons
  - Reliability
  - Cost

Missing Evidence from Previous Performance Assessment Efforts in K-12

- Evidence that higher-order skills, depth of understanding, and problem solving are better measured by performance assessments than by multiple-choice and short-answer tests
- Evidence that use of performance assessments enhances teaching and learning

Conclusion: Psychometric Values

1. Validity
   - Is it enhanced by performance assessments?
   - Evidence is needed to support claim.
2. Reliability (Generalizability)
   - How many tasks are needed?
   - How many are feasible?
3. Comparability
   - Different performance assessments.
   - Same performance assessment on different occasions.

Q&A