Educational Measurement:
What Lies Ahead

Presented at: NCME Presidential Address
Presenters: Lauress Wise
HumRRO

February 1, 2014
The Bulletin of the Atomic Scientists has moved the big hand of its "Doomsday Clock" to three minutes to midnight, i.e. the end of the world, citing the apocalyptic threat of global warming and nuclear proliferation.
Since 1991, we’ve moved closer and closer to the end

Qualitative Scale:
For Humanity To Survive

• We must solve many pressing problems:
  – Depletion of Earth’s resources
    • Global warming, we may already have passed a tipping point
  – Overcrowding and infectious disease
  – Wars of religious intolerance

• Today’s Youth Must Learn To:
  – Reason critically
  – Tell truth from lies
  – Work collaboratively

• Educational measurement is the compass to guide us in the right direction
We Have Our Own Contentious Arguments

- 1 versus 3 parameter IRT models

- Angoff Standard Setting
  - State of Art or
  - Fundamentally flawed?

- Value-Added Using:
  - Vertical Scales
  - Regression-Based Models
  - Student Growth Percentiles
We need NCME Now More than Ever

- A place where individuals with different beliefs and perspectives (test developers, test users, and amused observers) can come together, solve problems, and move ahead
  - Building evidence to test hypotheses and evaluate assumptions about our educational measures
  - Enacting standards for test development and use
  - Supporting new and developing measurement professionals
Presentation Overview

- How policy demands for educational measurement have evolved over the past several decades (what lies behind)
- Current challenges for measurement professionals (what lies ahead):
  - Providing meaning to score scales
  - Providing accurate and complete normative data
  - Assessing higher-order skills
  - Providing more complete diagnostic information
History is Important

- Those who cannot remember the past are condemned to repeat it. – George Santayana, The Life of Reason, 1906
- To know where we are, we have to understand where we have been (I said that)
  - We need to know how the policy uses of educational testing have evolved to really understand how best to inform educational policy today
  - Not just “cool”, but also useful
Early Goals for Educational Assessments

- Test uses for elementary and secondary education were low-stakes and limited through the 1950’s
  - College admissions tests
  - Some interest and ability tests used for career guidance
  - Diagnostic information on student strengths and weaknesses
  - Occasional evaluation of impact of changes in instructional practices

- Decisions to test and selection of tests made largely by LEAs
  - Some statewide testing (selected grades; matrix sampling)

- Accountability was not an issue
  - General trust in school leaders and teachers
It Used to Be All About the Student

- Measuring Student Growth
  - In earlier years, assessments were used to measure individual student progress
And then the Russians launched Sputnik!

It became harder to assume we were first in the world in math and science!

Led to national testing projects such as:

- **Project TALENT (1960)**
  - Inventoried skills, abilities, interests and other attributes of nearly 400,000 high school youth
  - Tracked progress 1, 5, and 11 years after high school graduation

- **National Assessment of Educational Progress (NAEP, 1969)**
  - Assessed achievement in several subjects at grades 4, 8, and 12

---

**Launch date:**
4 October 1957, 19:28:34 UTC

---

Evolving Policy Demands

- A Nation at Risk: The Imperative for Educational Reform, Commission on Educational Excellence, 1983
  - Trust in teachers and school leaders was no longer sufficient.

- National Assessment Governing Board created in 1989
  - Introduced proficiency standards
    - How much should students know and be able to do?
  - Launched state-level comparisons (Grades 4 and 8)

- Voluntary National Test (late 1990’s)
  - Purpose was never fully clear: compare students across state lines?

- No Child Left Behind (signed 2001)
  - Standards-based assessments (Grades 3-8, HS) required for all states
  - Participation in State NAEP (grades 4 and 8) required for all states
  - Consequences for low-scoring schools introduced
  - States adopted their own content and performance standards
Under NCLB

- Assessments Became a Lever for Reform

*If you test it, they will teach.*
*(School of Dreams)*
Figure D-3. NAEP score equivalents of states' proficiency standards for mathematics, grade 4: 2003

Fallout of NAEP State Mapping Study

- The NAEP State Mapping Study provided a graphic illustration that expectations for student achievement varied considerably by state
  - Raising basic fairness issues

- Several efforts were launched to do something about this inequity
  - Hunt Commission
  - National Governor’s Association alliance with the Council of Chief State Schools Officers
    - Led to drafting of the Common Core State Standards
  - National Research Council workshops on common standards
    - Funded by the Hunt Commission
    - Brought together former governors Hunt, Romer, and Wise
Continuing Evolution of Policy Uses

- **Common Core State (Content) Standards**
  - More rigorous content standards (some math moved to earlier grades)
  - Higher order content (e.g., integration and evaluation)
  - Proficiency levels (performance standards) not yet established

- **Race-to-the-Top (2009 - …) [marathon?]**
  - States encouraged to adopt Common Core State Standards
  - Two consortia funded to develop “Common Core” assessments
  - States encouraged to use assessment data in teacher evaluation

- **International Benchmarking**
  - Increasing concern about global competitiveness of US workforce
  - NCES released a report on linkage of NAEP and TIMSS scales and standards for Grade 8 mathematics and science (Oct. 2013)

- **Next Generation Science Standards**
  - Currently under review by many states; not yet federally mandated
Teacher Evaluation

- With Race to the Top States Were Encouraged to Use Tests for Teacher Evaluation
  - Hold them accountable and they will teach better (or else!)

Achievement Standards

- Novice
- Basic
- Proficient
- Advanced

Teachers

Policy-makers
Challenging Goals for Next Generation Testing

The next generation of assessments must address several challenging policy demands:

1. Attaching clearer meaning to reported scores
2. Providing complete and accurate national and international norms
3. Creating valid, fair, and reliable assessments of higher-order skills
4. Providing better instructionally relevant diagnostic information
Providing Meaning to Test Scores

Good enough for what?

Test Result: Good Enough
1. Providing Score Meaning Through

- **Predictive Statements**
  - College and career ready
    - Whatever that really means
  - Or prepared for the curriculum at the next grade level

- **Criterion Referencing**
  - Explicit descriptions of what students at a given score level know and can do
  - Often accomplished through item mapping
Predicting College Readiness is Also Not New

- The SAT, developed by the College Board was first administered on June 17, 1901.  
  - Developed by Carl Brigham (who also worked on the Army Alpha)  
  - Designed to predict success in college
- The ACT was developed in 1959 by E. F. Lindquist  
  - Also designed to predict success in college
- What is new is trying to assess mastery of the K-12 curriculum and readiness for what comes next at the same time
Predicting Work Readiness is Not New

- **Chinese Imperial Examination**
  - Established in 605 during the Sui Dynasty,
  - Candidates who did well on these essay tests were invited to work as civil servants in the Imperial Palace

- **Army Alpha – Robert Yerkes (and others), 1917**
  - Group administered test to determine readiness for military service during World War I
  - We won that war!

- **Army Aviation Psychology Program, Led by John C. Flanagan**
  - Developed Tests of readiness for specific Air Corps positions during World War II
  - We won that one too!
Predictive Interpretations of Test Scores

- NAGB has been working for nearly a decade to define preparedness for college, career, and military service

College Readiness
- College Board and ACT have identified college readiness standards
- Confusion among admissions, placement, grade point, and graduation rates
- Differential readiness for community colleges versus high selective colleges

Work Readiness
- Focus on academic skills needed for success in training for jobs that offer opportunities for career advancement without requiring a college degree
- May be different from college readiness

More research is needed
- To establish more empirical linkages to outcomes of interest
- Are measures of the CCSS really good indicators of readiness?
- How does readiness relate to graduation requirements/exit exams?
Criterion Referencing

- Expand performance level descriptors
  - Consider multiple tracks within a broad subject area

- Build into content frameworks
  - Through performance level descriptors?
  - Theory indicating stages of learning
    - Better than after-the-fact labels based on empirical item mapping.

- Some aspects of item difficulty are difficult to explain
  - Text complexity differentiates reading score levels
  - May need cognitive labs to understand difficulties of items mapped onto different score levels
2. Accurate Normative Information

- Providing complete and accurate norms

Test Result:
Better than Others

What others?
2. Improved Normative Interpretations

- We think we know how to create accurate norms
- Several tests have long been used to provide national comparisons for student achievement
  - Iowa Test of Basic Skills
    - Introduced in 1935 by E.F. Lindquist (and others)
    - Provided percentile scores in a variety of subject areas
  - Differential Aptitude Test
    - Introduced in 1947 by the Psychological Corporation
    - Designed to measure an individual’s ability to learn through measures of reasoning (verbal, mechanical, etc.)
- And then states all started using different tests
Issues in Building Better Norms

- Definition of the Target Population
  - Age or grade? (And who is still in school?)
  - ALL students? (Or only those who can be tested?)

- Definition of the Construct being Assessed
  - Controlling context and cues and modes of testing
  - Removing irrelevant social/cultural confusion

- Benchmarking to other assessments
  - With somewhat different content and administration and reporting procedures
Improved National and International Norms

- NAEP provides national norms for selected grade levels and subjects, but no individual student scores
- Neither PARCC nor SBAC will be a national test!
- PARCC and SBAC have developed different definitions of proficiency and will likely use different reporting scales
  - E.g. 4 versus 5 performance levels
- Much more research is needed:
  - Check approaches to international benchmarking
  - Assess comparability of PARCC and SBAC score levels
  - Comparability of paper and computer based test forms
  - Estimate accuracy at all score levels and for disaggregated groups of interest
  - Check stability over time
Norms May Not Be Forever

Trends in Mean Percentile Rank
(Based on Linn, Graue, & Sanders, 1991)
3. Measuring Higher-Order Skills

- Reasoning and Critical Thinking

Critical of Whom?

Test Result: Critical Thinker
Use of Performance Tasks for Higher Order Skills

- Several states developed performance-based assessments in/around the 1990’s
  - Kentucky Instructional Results Information System (KIRIS),
  - Maryland School Performance Assessment Program (MSPAP)
  - California Learning Assessment System (CLAS)
- Fall 1993 issue of the JEM was devoted to articles on performance assessments
- Performance Tasks were largely dropped due to issues of:
  - Cost (development and scoring),
  - (lack of) Generalizability, and
  - Inefficiency (testing time requirements)
Assessing Higher Order Skills

- A number of issues were/are being addressed in the SBAC and PARCC 2014 field tests, such as:
  - Generalizability across tasks
  - Scoring consistency/accuracy for performance tasks
  - Dimensionality across content areas and item types

- Many issues will remain
  - If the idea is to test application of skills to solve novel problems, how long will the problems remain novel?
  - Stability of higher order skills over time and across domains?
  - Can we measure process as well as outcomes?
Assessing Higher-Order Skills

- To get a better outcome than last time, we need to:
  - What are they?
  - Do we know how and are we trying to teach them?

- We need better scales of performance success
  - Example of writing skill
    - Rubrics, benchmarks, training sets
  - Measure process as well as output?
4. Diagnostic Information

What’s wrong with me?
Two Approaches to Diagnostic Information

- Multiple skills – provide information on strengths and weaknesses across multiple subdomains for a subject
  - Subscores

- Learning progressions – identify current point in the progression
  - What has been mastered
  - What to work on next
Assessing Multiple Skills

- Multivariate adaptive testing (van der Linden)

- Most Current Diagnostic Testing Models
  - Build around the idea that multiple skills are needed to answer some items correctly
  - Some difficulty in unpacking those skills clearly

- Maybe the real issue is diagnosing what students do not (yet) know or are not able to do (and why)!
  - Key may be in the analysis of incorrect answer choices
  - Or deeper analysis of why performance is suboptimal
Learning Progressions for Diagnosis Are Not New

- Jean Piaget (1896-1980) described discrete stages of cognitive development
  - Sensorimotor Stage (birth through language acquisition)
  - Pre-Operational Stage (approximately age 2 through 7)
  - Concrete Operational Stage (generally age 7 through 11)
    - Includes inductive reasoning, solving concrete problems
  - Formal Operational Stage
    - Includes abstract reasoning
  - Hypothesized specific cognitive stepping stones within each stage – a model for learning progressions.

- Diagnosis: knowing where a student is in a learning progression, suggests what to work on next
Using Learning Progressions

- Micro level (within grade) may integrate progression along different paths
- Macro level (across grades) to validate assertions that the Common Core Standards reflect logical progressions toward readiness
- Several models are being developed; but none are yet widely adopted, e.g.,
Example of Learning Progressions:

**Elementary/Middle Math**

**Numbers and Operations**
- Count objects to 10
- Add/subtract 1-digit numbers
- Add/subtract 3-digit numbers
- Multiply/divide integers
- Add/subtract/multiply/divide decimals
- Add/subtract/multiply/divide fractions
- Use positive integer exponents
- Use positive and negative exponents in operations

**Measurement**
- Compare lengths
- Measure length by repeating unit
- Measure area / perimeter
- Perform simple unit conversions
- Use formulas to compute area
- Compute volume/ surface of simple 3-D figures
- Use Pythagorean Theorem to solve complex measurement problems

**Know time units**
- Know time unit conversions
- Perform simple unit conversions
Conclusion

● Next Generation Testing is a means to meeting multiple policy objectives, not an end in itself
  – More than just a pretty test

● Among the objectives we seek to meet are:
  – Providing meaning to score levels
    • Assessing readiness for what comes next
    • Criterion referencing
  – Providing national and international norms (holding all students in the world to the same high standards)
  – Assessing higher order skills (and evaluating ways of teaching them)
  – Providing useful diagnostic information
My Personal Thanks Go Out To:

– The NCME Board and Executive Committee
  • Wim for setting the example
  • Rich for sound advice and hard work

– Susan Rees, our Executive Director, for keeping things moving
  • Jane Shepard and many others at TRG for their support

– Committee chairs and members, too countless to list here for doing the real work of NCME

– To all of you working in the field, particularly those who are presenting at this meeting

With your help and hard work, the end may yet be a ways off!