

Recommendations for the Race to the Top Assessment Program

Randy Bennett
ETS
rbennett@ets.org

Presentation at the Race to the Top Assessment Program Public & Expert Input Meetings,
General and Technical Assessment, January 20, 2010, Washington, DC

Question #1

- The Department is considering requiring “a through-course summative assessment system” – that is, a system that includes components of assessments delivered periodically throughout the school year whose results are aggregated to produce summative results. If we do this, how should we ask applicants to describe their approaches and/or plans for such a system, including any special considerations related to “through-course summative assessments” on the issues outlined below? What evidence should we request if such summative results are part of an accountability system?
 - a. Validity – including construct, content, consequential, and predictive validity
 - b. External validity for postsecondary preparedness
 - c. Reliability – including inter-rater reliability if human scored
 - d. Fairness
 - e. Precision across the full performance continuum (e.g. from low to high performers)
 - f. Comparability across years
- If States administer components of the “through-course assessments” at different times or in a different sequence, but the aggregated summative results are part of an accountability system, what are the issues around validity, equating, or comparability that we should be aware of?

Q#1 Recommendation #1

- ED consider *suggesting* “through-course summative assessment” as a preferred (rather than required) model
- Why?
 - There may be other, equally promising models
 - Ideally, several consortia should be funded, each following a different model, thereby allowing a real-world trial of the viability and effectiveness of those competing models

Q#1 Recommendation #2

- At a minimum, ED request the following evidence for consortia proposing “through-course summative assessment” (or any assessment) for accountability
 - A theory of action
 - A research plan for evaluating the theory of action

A Theory of Action

- The elements of the “through-course summative assessment” system:
 - e.g.: Periodic tests, project work, portfolios
- A logical and coherent rationale for each of those elements, including backing for that rationale in research (if available), e.g., :
 - “Periodic Tests: Periodic tests are intended to provide more timely feedback on student achievement of standards”
 - “Project Work: Project work is intended to allow assessment of competencies that can’t be measured through periodic tests”

A Theory of Action (con't)

- The claims that will be made from assessment results, e.g.,:
 - “Student performance on periodic tests and project work represents achievement of common standards”
 - “Students who perform at the ‘proficient’ level are ready to proceed to the next grade’s work”
 - “Teachers of classes with ‘lower than expected’ performance should be administratively reviewed and, potentially, sanctioned because they are likely to be ineffective”

A Theory of Action (con't)

- The intended effects of the assessment system and the mechanisms thought to cause those effects, e.g.,:
 - “Project work will encourage a focus on important competencies not promoted by traditional assessments”
 - “Linking teacher sanctions to student performance will cause improved teaching practice”
 - “Focus on important competencies and improved teaching practice will lead to higher achievement”

A Research Plan

- Is the theory of action logical, coherent, and scientifically defensible?
 - TAC review
 - Early public presentation, invited critique by independent experts, rejoinder by the proponents, and publication

A Research Plan (con't)

- Are the *stated* assessment claims empirically supported?
 - “Student performance on periodic tests represents achievement of common standards”
 - Proposed alignment study and cognitive interviews
 - “Students who perform at the ‘proficient’ level are ready to proceed to the next grade’s work”
 - Proposed predictive study
 - “Teachers of classes with ‘lower than expected’ performance should be administratively reviewed and, potentially, sanctioned because they are likely to be ineffective”
 - Proposed blind observational study comparing teaching practice in “lower than expected” and “higher than expected” classes

A Research Plan (con't)

- Are the *implicit* assessment claims empirically supported?
 - “Scores aggregated across periodic assessments can be compared”
 - “Aggregated scores can be used to measure growth”
 - “Scores from constructed-response tasks, including project work, are generalizable across raters”
 - “Scores from alternate test forms can be used interchangeably”
 - “Scores from periodic assessments can be used to identify students at risk of failing to meet proficiency”
 - “Scores from periodic assessments have the same meaning across population groups”

A Research Plan (con't)

- Was the system implemented as designed, were the intended effects on individuals and institutions achieved, and did the postulated mechanisms appear to cause those effects?
 - “Project work will encourage students and teachers to focus on important competencies”
 - Study of classroom processes before and after advent of the assessment system
 - “Linking teacher sanctions to student performance will cause improved teaching practice”
 - Analysis of teacher lesson plans before and after the assessment system was introduced

Q#1 Recommendation #3

- To minimize through-course assessment timing and sequence effects, ED might strongly encourage each consortium to:
 - Agree upon an administration sequence and set of administration windows
 - Provide a plan for protecting test content (so students taking the test later in the window do not unfairly benefit)
- A single sequence prescribes only the top-level curricular order for a grade (e.g., the topics in quarter 1 vs. quarter 2)
 - Within-quarter sequences, and how to address the topics, can be left open

Question #2

- The Department is considering inviting applicants to create a “system” for developing and certifying the quality and rigor of a set of common end-of-course summative exams in multiple high school subjects. What evidence should we ask applicants to provide to ensure that, across a consortium, their proposed “system” will ensure consistent and high levels of rigor?

Q#2 Recommendation #1

- ED ask bidders to propose a method for certifying quality and rigor
 - A proposed process and the evidence to be used, e.g.:
 - Alignment of the EOC tests with common standards
 - A comparative review against other highly regarded EOC tests (e.g., International A Levels)
 - A review of each EOC test's technical characteristics
 - A qualified, independent body to refine the process, impanel experts, and conduct reviews

Question #3

- If the Department requires computer-based test administration, are there specific implementation challenges that we should ask applicants to consider and address in their proposal? In particular, what evidence or strategies should we require of applicants to ensure that the computer-based and any needed paper-and-pencil versions assess comparable levels of student knowledge and skill while preserving the full power of the computer-based item types? Are there special challenges related to computer-based testing for students with disabilities and what additional evidence or strategies should we require of applicants to ensure that computer-based tests yield valid results for this population of students?

Q#3 Recommendation #1

- ED consider *suggesting* computer-based assessment as a preferred model for a significant component of the RttT Assessment Program competition
 - Workplace and advanced academic settings routinely require individuals to do cognitive work on computer
 - To the extent that common standards reflect these requirements, paper testing may not be able to measure the standards fully
- Whatever the bidder's chosen model, ED should require the bidder to justify the fit with common standards, as well as with other goals of the RttT Assessment Program

Cross-Mode Comparability

- Comparability is important when, e.g.,:
 - Assessment results are compared over time and the delivery mode has changed from paper to computer
 - If scores are not comparable, trends may no longer be interpretable
 - Assessment results are compared across individuals and some individuals have taken the test on paper while others have taken it on computer
 - If scores are not comparable, comparisons may be unfair
 - Population groups are compared and the proportions of students taking the test on computer differ across groups
 - If scores are not comparable, group comparisons may be meaningless

The Dilemma

- Moving a large testing program to computer is likely to require a multi-year transition
- If maintaining cross-mode comparability is important over that period, innovation that threatens comparability will be difficult to implement

Q#3 Recommendation #2

- Require bidders to propose a strategy for dealing with that dilemma, e.g.:
 - The Incremental-Innovation Model
 - The Concurrent-Innovation Model

The Incremental-Innovation Model

- Approach
 - Create parallel paper and computer tests from the same content specs, collect comparability data, and equate, if possible
 - Run paper and computer programs in tandem, transitioning more students and schools to computer until paper administration becomes an exception
 - At that point, introduce innovation that takes advantage of the computer in ways that can't be duplicated on paper
- Advantage
 - Preserves the meaning and fairness of score interpretations that depend upon comparability
- Disadvantage
 - Delays innovation

The Concurrent-Innovation Model

- Approach
 - Create innovative computer-based assessments and (non-comparable) paper assessments
 - Set performance standards separately for each test
 - Have a representative sample take both tests and create a “concordance” (as for ACT/SAT), which might allow cross-test comparisons and aggregations
 - Run paper and computer programs in tandem, transitioning more students and schools to computer until paper administration becomes an exception
- Advantage
 - Advances innovation
- Disadvantages
 - May appear unfair to some as the paper tests may seem inferior
 - To the extent the tests measure considerably different constructs, cross-test comparisons and aggregations may have little meaning

Q#3 Recommendation #3

- Where comparability is important, ED should require bidders to provide evidence consistent with professional standards (e.g., *APA Guidelines for Computer-based Tests and Interpretations*)
 - Scores may be considered equivalent when, across modes:
 - The rank orders closely approximate one another
 - The score distributions are approximately the same (or have been made approximately the same through statistical adjustment)

Question #4

- The Department wants to encourage ongoing innovation and improvement of assessment design, development, administration, and use. However, given that we are proposing four-year grants, what should we ask of applicants to ensure that they have structured a process and/or approach that will lead to innovation and improvement over time?

Q#4 Recommendation #1

- Require that bidders present a long-term vision for a next-generation assessment system
 - A rationale for why that vision is meaningful
 - A set of steps to progressively move toward it
 - A clear statement of why the system developed under the RttT Program would be a significant step toward the vision
 - A plan for continuing progress toward the vision after the RttT funding ends

Q#4 Recommendation #2

- Require that bidders present a specific plan for continuous innovation *during* the RttT period
 - Include one or more existing assessment- or education-innovation centers as consortium partners
 - Closely involve students, teachers, and administrators in design, tryout, and evaluation

Continuous Assessment-Innovation Models

- School-Based Model
 - Select, by competition, a subset of schools of varying demographic characteristics to serve as assessment-innovation partners
 - Designate them for a set period (e.g., 3 years)
 - Give them a waiver from accountability requirements that would impede innovation
- Project-Based Model
 - Select participating schools on a rolling, project-by-project basis
- Hybrid Model

Question #5

- With the help of experts, we identified two issues that seem to require additional, focused research. Have we described the issues correctly? Are there other issues that need additional focused research?
 - a. Use of value-added methodology for teacher and school accountability
 - b. Comparability, generalizability, and growth modeling for assessments that include performance tasks

Q#5 Recommendation #1

- Value Added Modeling (VAM)

“[A 2008 BOTA-NAE workshop on VAM concluded that] ... there is little scientific consensus about the many technical issues that have been raised about [VAM] techniques and their use.”

“BOTA agrees with other experts who have urged the need for caution and for further research prior to any large-scale, high-stakes reliance on these approaches.”

Letter from NRC BOTA to Arne Duncan, October 9, 2009, pp. 8-9

–Fund Focused Research

Q#5 Recommendation #2

- Performance assessment
 - We know a lot about the problems but, for use in accountability, there's a lot more we need to know about workable solutions, e.g., how to:
 - Aggregate scores over through-course assessments comprised of performance tasks
 - Create meaningful scores from unstandardized projects and portfolios, especially if scored locally
 - Score computer-based performance assessments in which every mouse-click, keystroke, and resulting event are recordable
 - Fairly assess students with disabilities and ELLs with any of these methods
 - Should be funded as Focused Research, in addition to main-competition research for consortia using such approaches

Q#5 Recommendation #3

- Learning progressions
 - Have considerable potential for guiding assessment design within and across grades, and for providing (tentative) formative feedback from summative assessment
 - There are relatively few, well-researched progressions and no examples of their use in design or reporting for large-scale assessment
 - A significant Focused Research program is needed to:
 - Generate and empirically support progressions in ELA and math
 - Incorporate them into large-scale assessment design and reporting
 - Create related classroom assessments that can point towards appropriate instructional materials
 - Evaluate impact of the above

Summary

1. Through-Course Summative Assessment

- Suggest it as a preferred model
- Require a theory of action and an associated research plan as an integral part of all bids
- Suggest that each consortium agree upon a single sequence and set of administration windows (or require evidence to support the meaning of scores from different sequences)
- Require a plan for protecting the security of test content

2. System for Developing and Certifying EOCA Quality and Rigor

- Require bidders propose a certification method, including (a) a process and evidence to be used and (b) a qualified independent body to refine the process, impanel experts, and conduct reviews

Summary

3. Computer-Based Assessment

- Suggest it as a preferred model for at least a significant portion of the RttT Assessment Program competition
- Require bidders to:
 - Justify the fit of their chosen assessment mode with common standards and RttT Assessment Program goals
 - Propose a strategy for dealing with the “comparability dilemma”
 - Where needed, provide comparability evidence consistent with professional standards

4. Innovation and Improvement Over Time

- Require bidders to present:
 - A long-term vision, including a plan for progressing toward it once RttT Assessment Program funding ends
 - A specific plan for continuous innovation during the RttT Assessment Program period

Summary

5. Focused Research

- Value Added Modeling
- Performance assessment
 - Aggregation methods, score meaning from projects and portfolios, computer-based performance assessment, and fairness for special populations
- Learning progressions

Recommendations for the Race to the Top Assessment Program

Randy Bennett
ETS
rbennett@ets.org

Presentation at the Race to the Top Assessment Program Public & Expert Input Meetings,
General and Technical Assessment, January 20, 2010, Washington, DC