



Listening. Learning. Leading.®

Technical Challenges With Innovative Item Types

Stephen Lazer
ETS
December 10, 2009

What I will cover today

- Some disclaimers
- What is old is new
- Defining “innovative item types” and the problem
- Discussion of issues
- Suggestions for paths forward



Disclaimers

- I like performance and technology-enabled item types and have extensive experience with them (performance assessments in the arts, group assessments in history, problem solving using technology, observational teacher measures, situational judgment tests).
- I say this because much of what I will say will be cautionary.
- Not to discourage the use of these exercises, but to help ensure success.



There is no new thing under the sun

- What are the emerging trends?
 - Greater use of performance testing to measure an increasing array of skills and get a more nuanced picture of students
 - Rejection of over-reliance on multiple choice because of measurement limitations and perceived impact on instruction
 - Increased use of technology to allow measurement of content and skills not possible on paper and to tailor assessments to individuals
 - Wish to use assessment tasks that are authentic and that are worthwhile learning activities
- If these are all true, it must be 1990.
- This is, of course, an overstatement; we may be closer to tipping points now, and many of the earlier discussions did not hit K12.
- But we need to consider what we should have learned in the last 20 years.



What do we mean by “innovative item types”

- Some mean anything beyond multiple choice.
 - Some types of CR items are not particularly innovative, even in large-scale assessments.
 - However, in sectors of the assessment world they have played a limited role.
- Other item types are quite innovative in display terms, but don't seem to change the constructs being measured.
 - One can argue they increase engagement, but no particular evidence.
 - One can also argue they reduce the specific test-taking skill associated with MC.
 - At worst they can be “eye-candy.”



For purposes of this talk, “innovative” means

- Exercise types that allow us to expand measurement beyond constructs possible with largely MC testing
 - These include open-ended and performance testing
 - They also include using technology to expand constructs
 - To test things we wished we could test on paper
 - To test skills and knowledge that do not exist absent the technology itself
- I’m talking today about use of these items in high-stakes tests
 - Other uses are possible and carry less baggage (and maybe less promise).



What do we hope to get from the “innovative items?” (1)

- From increased use of performance testing:
 - Possibility of greater content relevance
 - Possibility of instructional feedback (responses have a footprint)
 - Certainly a broadening of the skills and content that can be measured; offers opportunity for greater integration of skills and understanding
 - Should have other positive consequences for instruction



What do we hope to get from the “innovative items?” (2)

- From increased use of technology:
 - Will allow for maintained or enhanced validity
 - Some skills, like writing, will only be logically measurable in a computer environment because that will be how kids write.
 - Computers could allow us to measure an array of skills that are not possible or practical in large-scale pencil-and-paper situations (lab simulations, use of reference materials, use of technology itself).
 - Could allow for tailoring of test difficulty to candidate’s level and skills (may limit innovation at the item level)
 - Could more effectively connect summative and formative assessments
 - Makes possible use of electronic scoring methods that may in turn expand new item use in an economically defensible manner and could speed score turnaround



Issues we may face with these items: Intro

- These issues do not uniformly face all types of these items
- They represent the sorts of issues one faces, but one can find counter examples



Issues we may face with these items: Cost

- Some of these sorts of items can be expensive and time-consuming to develop
 - Particularly a problem for exercises for which we lack a set of “operational norms and practices” (like simulations, in which most work has been on learning systems and not assessments)
- Many of these exercises require human scoring, which adds major cost
 - Unlike development or analysis costs, these go up on a per-student basis
 - Also add schedule time
 - There are electronic scoring options, although these carry limitations as well
- Computer-based testing can add hidden costs (machines, development of larger item pools), even if tests use traditional items



Issues we may face with these items: Test development know-how

- For better or worse, we have a good understanding of how to produce large numbers of MC items with known performance characteristics
- This knowledge also exists for some types of CR items, although the cost of a mistake in development is far higher
- For some item types, little “operational knowledge” or templates for development
 - This again creates both cost and the chance for failures, usually in items that count quite a lot
- Need improved cognitive models (lack brute force of MC tests)
- There are plenty of bad open-ended and computer items



Issues we may face with these items: Generalizability

- These items may allow for measurement of a broader construct, but may lead to weaker measurement of it.
 - Items can be time consuming, leading to fewer questions
 - Probability of strong task-by-person interaction (particularly if highly contextualized)
 - Where human scoring is involved, another source of variation
- Not a reason to avoid these items (generalizability does not trump all)
- But low enough generalizability would, of course, belie any validity argument
- Clearly an issue with long performance tasks, but also likely an issue with some item types (like simulations) where interdependence among observations may effectively render these non-independent observations
 - Content relevance may strongly argue against local independence
- There may be ways around this problem by using periodic assessment to lengthen testing time, or constraining question length.



Issues we may face with these items: Validity

- How to make sure items measure what we really intend; looking authentic does not equal valid
- Validity depends on the claim one wants to make, and construct definition is thus essential



Issues we may face with these items: Scoring

- Well-known issues regarding scoring quality
- Can conflict with pressure for rapid results
- Automated scoring is becoming possible, but these systems come with their own challenges
- If we wish to make “statistical comparisons” over time, need to control scoring from a trend perspective
 - Trend comparisons may not be the only “good,” but we must be aware of trade-offs



Issues we may face with these items: Analysis

- For some emerging item types, lack of shared, well understood, and effective QA and QC tools
 - Standard item analysis and DIF approaches do not work well with traditional CR items, and not at all with other forms of performance tasks
 - This is important if these items are to become part of the mainstream and not “boutique jobs”
- Lack of model for getting maximum information out of some emerging item types; complex items can yield little data
- Some current psychometric approaches can be antithetical to some new types of items (e.g., local independence)
- These items may create problems given some emerging system requirements (for example, vertical scaling/growth modeling)



Issues we may face with these items: Information

- One of the real promises of performance and technological items is more and more nuanced information about test takers
 - Early attempts have usually not lived up to this promise
 - We need to try to be sure that we have a plan for scoring and analysis that yields this
 - Current thinking is “snail trail” information on computer tests but no systematic understanding of how to analyze and report these data



Issues we may face with these items: Equating/Trend

- Use of some sorts of extended performance tasks can create a “double whammy” for equating.
 - If tasks can be reused, one must control human scoring carefully.
 - The fact that tasks are long and memorable may mean they cannot be reused
 - Careful development may not obviate the need for equating
- Current notions of equating/trend may make little content sense in technology-rich assessments
 - We would never continue to use 1996 word processors to keep trends intact
- This is a very big deal if we are trying to track either system or individual progress, or to measure teacher value added



Issues we may face with these items: Technology

- Some types of items need computers and possibly longer windows
- We need to ensure that interfaces and activities are not so complex that it takes weeks for someone to learn to take the test
- Advanced item types may yield different group patterns, which may be an equity issue depending on construct definition
- We must make sure we remember it is an assessment



Issues we may face with these items: Field Testing

- Length and complexity of tasks means they probably cannot be tried out embedded in operational tests
- Stand-alone field testing is increasingly difficult
- This places more pressure on equating



Issues we may face with these items: Conflicting Imperatives

- Some of the wish to use innovative items can contrast with expressed desires for:
 - Faster scores
 - Cheaper tests
 - Shorter tests
 - Adaptive tests
 - Accommodations for special-needs test takers
- None of these are insurmountable conflicts but need to be considered carefully



Does this mean we should stick with what we've got?

- No
- But it does mean we've got to think about implementing these items in high-stakes tests carefully.



What to do (1)

- I do not know, but I have a little advice.
- If ever we needed to learn from the past, now is the time.
- We need to be conscious of the limits of what any single test can do – we need to talk more about assessment systems, where different components serve different needs but reinforce each other.
- We have amazing opportunities; we need to have a frank discussion about how to avoid blowing them.



What to do (2)

- Realize that fundamental change may take some time and research
- Make sure designers think at the item, test, and construct level, and don't just create "clever tasks"
- Realize that traditional psychometric concerns don't necessarily get to win out, but that ignoring them is a mistake
- Realize that innovations can come in stages and in different parts of the system



What to do (3)

- Consider approaches (such as periodic assessment) that get around some of the generalizability issues (as long as interim assessments do not have to be that reliable)
- Pay special attention to the psychometrics of emerging item types
- Beware of gathering evidence without an analysis plan





- Questions?
- Comments?