Psychometric Considerations In Game-Based Assessment

Executive Summary

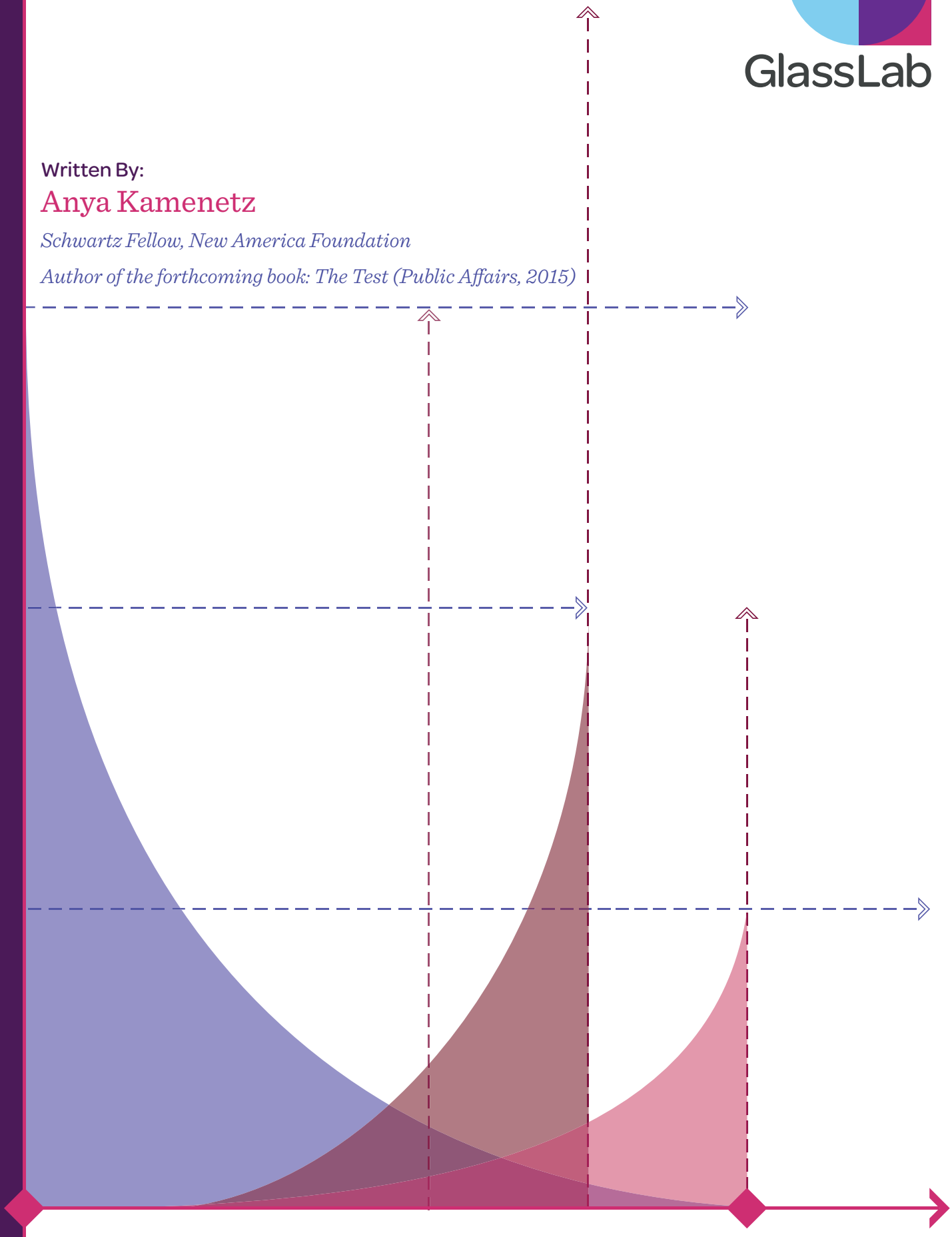
GlassLab

Written By:

Anya Kamenetz

Schwartz Fellow, New America Foundation

Author of the forthcoming book: The Test (Public Affairs, 2015)



Executive Summary

"Psychometric Considerations in Game-Based Assessment" is a work of multidisciplinary scholarship. It's also the first publication of an unprecedented multidisciplinary collaboration. GlassLab brings together game designers, learning scientists, and psychometricians from Institute of Play, the Entertainment Software Association, Electronic Arts, Educational Testing Service, Pearson's Center for Digital Data, Analytics & Adaptive Learning, and other organizations. Their goal is to build and emphasize the deep interconnections across the often quite separate areas of learning, testing, and game-making.

The game-based assessments this team is developing, starting with SimCityEDU: Pollution Challenge!, which debuted in November 2013, are something new under the sun. They are not "gamified" educational software. Nor are they games "thrown over the wall" to learning scientists and psychometricians to add scoring features after design is complete. These are real games that seek to evoke and measure real learning in real ways. The authors of this paper want to see GlassLab's work replicated and extended. They "seek to contribute to an integrated framework for designing, implementing, and using game-based assessments; one which builds on current best practices in learning, game design, and assessment design."

This paper goes into detail about how that can be done.

It also clarifies, as this summary will, why it should be done.

Testing for Growth

There are two familiar types of assessments in broad use in schools today: summative tests and formative tests. The first are typically multiple-choice and short answer standardized tests. These tests are optimized to assess "crystallized" intelligence, that is, accumulated knowledge and skills. They are often high-stakes, given at the end of courses: for example, a state accountability test that determines the rating of a school, district or teacher, or a high school graduation exam. A less common, but highly valuable form of summative assessment is the performance-based assessment, which calls on students or groups to complete an authentic task, research project, performance or presentation, to be judged according to a rubric.

The main purpose of a formative test, by contrast, is to give feedback to students and teachers to guide learning along the way. In order to provide the best instruction, teachers need more fine-grained, timelier information about students' states of knowledge than is provided by summative tests. They

often try to collect this information through pop quizzes, an instant poll or calling on a student. What's largely missing from the traditional assessment palette of formative and summative tests, of high stakes and pop quiz, is a way to get at the process of thought and at "fluid" intelligence, the capacity to think logically and apply reasoning in novel situations. The concept of "dynamic testing" goes back to pioneering educational psychologist Lev Vygotsky, who posited a test—teach—retest model. The idea is to get a sense of the student's best possible effort given appropriate support. This could be thought of as testing for teachability, integrating teaching into the assessment process.

As the authors will explain, to inform their work at the intersection of game design, instructional design, simulation design, and assessment design they assume a common, socio-cognitive perspective on learning. Learning in this view is not just the response of a disembodied neuron to an electrical signal. It happens in a context marked by particular patterns of activity, whether school, family, workplace, or community. Learning involves not only skills and knowledge, but also identities, values and epistemologies, ways of representing meaning with particular uses of language, common to particular groups—such as people who play SimCity, study Danish together, or troubleshoot computer networks. "Drop from the sky" summative tests, in particular, do a poor job of representing or evoking the complex overlapping domains, both macro and micro, in which learning takes place.

This view takes "intelligence" not as a fixed, hereditary trait, but as it actually functions in the real world: as an application of knowledge and skill, alone or with others, to solve problems and create new knowledge in areas like math, science, history, economics, literature, design, technology, and so on. You might call this fluid (vs. crystallized) intelligence, practical intelligence, or intelligence in action.

When you think about intelligence this way, it is clear that you can become more intelligent by continuously tackling new challenges in new ways.

For educational purposes it appears both more helpful and more hopeful to understand a student's capacity for growth, her particular learning strengths and weaknesses, and how they interact with her environment, rather than merely take a snapshot of whether she is above proficient or below proficient in a specific subject at a given point in time. But various models of dynamic testing have been tried over the years without achieving strong predictive power or wide use. Meanwhile, assessing fluid intelligence, particularly from a socio-cognitive perspective, seems to require complex performance tasks that are not easily machine-scored. The ideal would be to eavesdrop on the process of thinking in action, something that sounds highly infeasible.

Until now.

Games for Learning and Assessment

Over the last few decades, learning scientists have become quite interested in electronic games for several reasons. Unlike some other learning activities, they seem to spark players' intrinsic motivation to continue, that is, they are fun and engaging in themselves. They are capable of simulating complex situations, from troubleshooting computer networks to flying jet planes, to a degree of detail that becomes easily translatable to the real world. And importantly, they are adaptive, meaning the level of play adjusts as the player becomes more adept. Another way of expressing this is that games teach you how to play them. Another way is that they are designed to keep players in the "proximal zone of development" or the "state of flow" at which learning and engagement takes place most optimally. Yet another way is that games provide immediate and implicit formative feedback, with the processes of learning, performance, and assessment unified into a single cycle. GlassLab aims to do just this, with games that "winning" requires learning to, for example, model systems or construct arguments.

Games can stand alone as a learning activity, providing valuable information to the player and her teacher, who can intuitively characterize the evidence provided by the game. In low-stakes situations, this is perfectly satisfactory. The quality of evidence about learning gathered is more important than the analysis of that evidence.

But another reason learning scientists are interested in games is because of their potential to provide a new window into the process of learning. Familiar assessments produce a one-dimensional stream of data: a right or wrong answer on a multiple-choice question, or a rater's score on an essay. By contrast, playing video games, as with other kinds of educational or interactive software, generates a "digital ocean" of data that provides many more potential sources of evidence of students' thinking and learning in action: not only their meaningful actions within a game's logic, such as bulldozing a power plant or calling up a voter report, but their random mouse movements or the rate of their keyboard clicks. In some systems, the telemetry can extend to logging players' breathing, heart rate and facial expressions.

The authors of this paper are interested in designing games with an assessment architecture that can analyze all of this data and explicitly characterize and communicate its findings outside the local setting, as for moderate and higher stakes purposes. They want to gather "evidence about evidence," or support for the reliability, comparability, and validity of these findings. And they want to take advantage of all the rich data provided by games to produce a dynamic portrait not just of crystallized knowledge in a given content domain, but of higher-order constructs such as, in the case of SimCityEDU, systems thinking.

This isn't just a new kind of test. It's testing for a new kind of learning.

It is here that psychometrics comes in.

Psychometrics for Games

Psychometrics, the measure of the mind, uses statistical tools to reason from limited evidence to underlying constructs. Psychometrics allows designers of game-based assessments to create probability models that connect students' performance in particular game situations to their skills, knowledge, identities, and values, both at a moment in time and as they change over time. Through the use of statistical tools such as Bayes theorem, designers can update beliefs about these variables as new evidence arrives during gameplay. Psychometrics provides a way of characterizing the quality of the evidence thus accumulated. And it supplies a framework for sorting out evidence provided through game play in order to make better design decisions and in turn, collect better evidence.

In Chapter 12, "Psychometric Properties," the authors discuss ways of thinking about the quality of inferences and decisions based on the fallible and finite information provided by assessments. They reference four core psychometric properties of reliability, generalizability, comparability, and validity. Colloquially, reliability refers to the chances that a student tested at a level 4 really is a level 4. Generalizability asks whether the student would demonstrate the same capability given a different kind of assessment, or another game-based assessment of the same skill with different content. Comparability concerns how fair it is to compare or rank students based on their results.

Validity, a "paramount" value in psychometrics, concerns decisions made and actions taken as a result of the information provided by the tests. For example, do students learn from the formative feedback provided by a game-based assessment? Will the game-based assessment indeed tell us something meaningful about the student's understanding of systems thinking in other situations? Do teachers get actionable information to improve teaching? Does the feedback give designers the information they need to make future versions of the games better? Of course, answering these questions requires new assessments or other measures to validate the previous assessments – to see how the information from the assessment actually plays out beyond the game context.

Games for Psychometrics

So that's why game-based assessment designers need psychometricians. Why do psychometricians need GBA designers? Because collaborating in this way promises to expand the frontiers of psychometrics.

While the discipline of psychometrics was developed for reasoning from the "digital desert" of relatively sparse data provided by traditional assessments to posited fixed traits such as Spearman's g factor or Binet's IQ, GBAs require taking the "digital ocean" of data produced in play and using it to make inferences about dynamic, shifting, socio-cognitive capabilities For this reason, the authors argue, applying psychometrics to game-based assessments advances the science of psychometrics.

There are several unique features of game-based assessments that push the frontiers of psychometrics, which the authors call "interesting factors." These arise in traditional tests but come up more often and more obviously in game-based assessments.

First, like performance assessments and other "authentic" assessments, GBAs tend to be more open-ended and complicated challenges. Therefore, there is a high chance of misinterpreting students' responses due to variations in their background knowledge and skill outside the framework of what the test is testing. In truth, "what else students know" is always a factor in performance on any kind of test, but the current conventions for analyzing results of high-stakes multiple choice tests largely ignore this situative and socio-cognitive reality. Designers of game-based assessments ought to be more careful. For example, in SimCityEDU: Pollution Challenge! students are supposed to be demonstrating their understanding of systems thinking with respect to a system that includes jobs, pollution, and several other factors. But their performance may also be influenced by a lack of familiarity with mouse clicking, "hovering," and "plopping," by not knowing the real-world facts about the relative pollution produced by coal vs. solar plants, by problems reading graphs and charts, or by gaps in English vocabulary, to give a few examples. The authors suggest surrounding GBAs with both teacher guidance and in-game guidance to support both optimal interpretation of game play and optimal learning, building from what students already know and can do.

A second and related factor is the stability of complex constructs such as systems thinking from one context to another. The authors urge caution in interpretation based on previous experience with performance assessments. Much more empirical work is needed in order to establish generalizability. Creating a variety of game-based assessments for systems thinking featuring different systems, but with a common set of concepts and representations, could assist students in activating these concepts in different contexts.

Thirdly, GBAs pose complex reporting requirements to multiple users—student, teacher, outside observer—at various times and various levels. For example, task-level feedback is provided directly to the student during game play; an in-class diagnostic test is concerned with giving the finest possible grain of detail on an individual, perhaps at the end of a challenge; a designers' feedback survey might want to capture trends at the end of a game or a series of games completed by a large group of players. GBAs may have multiple evidence accumulation processes running simultaneously.

Fourthly, there is the question of adaptivity. Games at their best are highly adaptive, following an individualized path for each user and, as observed earlier, keeping the player within her zone of proximal development, balanced between anxiety on the one hand and boredom on the other. Values as measured by the assessment change over the course of play, which is another way of saying that students learn while completing the assessment, an experience not common with traditional non-performance assessments. Three lines of research converge around this dynamic aspect of

GBAs: experiences around the leading edge of one's capabilities optimize learning, assessment, and engagement at the same time. Therefore, design principles from learning, psychometrics, and instruction all reinforce each other.

Fifth, also unlike most tests, students play games again and again, getting better each time—how can the evidence model update to accommodate these changes? For example, rather than scoring a particular variable as "succeeded" or "failed," the values may need to expand to include cases like "successful on the first try," "successful on 3rd try," or "gave up after first try."

And finally, students may play games together. This requires the psychometric modeling to either take place at the level of a team rather than an individual, or even more complex, to model the participation of each individual within a collaboration.

Design Guidelines for Evidence-Centered Game Design

In the last chapter of the paper, "Implications for Design," the authors describe how to build the "hybrid creature" that is a GBA. Hopefully, the result is more powerful chimera than a clumsy Frankenstein's Monster.

Their design approach is Evidence-Centered Game Design, ECgD, a name that echoes the Evidence-Centered Design approach to assessments pioneered by coauthor Bob Mislevy. Game elements include domain analysis, domain modeling, identification of work products, types of evidence, scoring models, and student-model variables.

ECgD is an attempt at integrating the processes of game design and assessment design with four stages:

Definition of competencies from a non-game realm.

- A strategy for integrating externally-defined competency with gameplay competency.
- A system for creating formative feedback that is integrated with the game experience.
- A method for iteration of the game design for fun, engagement, and deep learning, simultaneous with iteration of the assessment model for meaning and accuracy.

In this integration, psychometrics and assessment design feed back into game design to help promote learning. The assessment design helps game designers see how theory and experience in some learning domain can be applied to build the situations, the challenges, and the actions that players interact with. Psychometric models provide metrics for information about players' capabilities, so designers can test alternative scenarios for evidence just as they do now for engagement. These engineering tools build on designers' skills and educators' understanding of learning in the domain, to improve the combined impact of play and learning.

The authors suggest the development of modular elements that can be repeated to make it easier to create new GBAs. The idea is to combine ideas like familiar game mechanics from the game world in natural ways of capturing and making sense of players' actions from the assessment world, so the designers new GBAs don't have to rediscover everything for each new game. Here are two examples:

- A generalized table format that needs to be filled out with drag-and-drop elements, which can be used in a variety of GBAs for a student to express a provisional hypothesis and unlock laboratory tools to carry out experiments.
- A system diagramming tool, such as STELLA (Richmond & Peterson, 2001), which allows students to model a system with a palette of objects and connections that represent stocks, flows, feedback loops, etc., then run the model. The same underlying code can be used to present information as stimulus material, serve as a tool in investigations, and create work products for a variety of subject domains and game contexts. It is minimally constrained, provides strong evidence about students' facility using systems concepts, and, because the work product is a file of objects and attributes, lends itself to automated scoring routines that examine its properties and check the results of its runs on standard test data.

Conclusion

Schools have a great need to assess 21st century skills such as systems thinking, collaboration, problem-solving, and communication in the context of important subject-area knowledge and concepts.

The path that the authors have outlined is not easy. They call on experts in disparate fields to collaborate to build and test more GBAs in order to bolster their empirical track record, which is necessary to see if they can fulfill their promise.

GBAs offer a new flavor on the assessment platter; they address themselves to higher-order skills and concepts in a way that is open-ended, personalized and engaging, yet with the application of psychometrics to the rich data they generate, they have the potential to demonstrate high validity and reliability.

GBAs could help students, teachers and the education system operationalize our definition of learning and intelligence in a more sophisticated way, by capturing individual patterns of thought rather than just patterns of numbers: like going from a bar code to a photograph.