10 Game-based assessment for education

Jack Buckley, Laura Colosimo, Rebecca Kantar, Marty McCall, and Erica Snow

Imbellus, Inc., United States

This chapter discusses how recent advancements in digital technology could lead to a new generation of game-based standardised assessments in education, providing education systems with assessments that can test more complex skills than traditional standardised tests can. After highlighting some of the advantages of game-based standardised assessment compared to traditional ones, this chapter discusses how these tests are built, how they work, but also some of their limitations. While games have strong potential to improve the quality of testing and expand assessment to complex skills in the future, they will likely supplement traditional tests, which also have their advantages. Three examples of game-based assessments integrating a range of advanced technologies illustrate this perspective.

Introduction

Rapid technological developments such as virtual/augmented reality, digital user interface and experience design, machine learning/artificial intelligence, and educational data mining have led to the improvement of simulated digital environments, and accelerated progress in the quality and design of digital simulations and video games. While this has led to the development of a range of "e-learning" applications to be used both inside and outside of the classroom (from virtual labs to medical e-learning tools with simulations), this technological advancement has also opened avenues for a new generation of standardised assessments. Such game-based assessments allow for the assessment of a broader range of skills (e.g. creativity, collaboration or socioemotional skills), as well as better measurement of some aspects of the "thinking" of respondents, including in traditional domains like science and mathematics. Moreover, the use of simulated environments enables assessing knowledge and skills in settings that are more authentic to "real life" applications of those skills.

While promising, this new generation of assessments brings its own challenges. For examples, they are more costly and difficult to develop than traditional standardised tests based on a simple succession of discrete questions or small tasks. Nevertheless, some game-based standardised assessments have already been successfully developed and will likely be one part of how the learners of tomorrow will be assessed. The chapter is organised as follows: we first argue that game-based assessments address many of the critiques of traditional assessment and have the potential of being aligned more closely to teaching and learning in the classroom; we then explain how these assessments work, what kind of technology they use, what kind of data they draw on, and highlight the challenges in building them; we provide some examples of game-based standardised assessments, before reflecting on the role they could have in the future, and what national infrastructure may be required to deliver them at scale.

Why game- or simulation-based assessment in education?

The use of standardised assessment in education – increasingly coupled with well-defined standards for academic content – is far from a new idea, dating back some four decades in some high income countries and at least 20 years internationally (Braun and Kanjee, 2006[1]). More recently, leaders in education policy, teaching and learning, and cognitive theory have come together to call for greater coherence among instruction, curriculum, and assessment and for a comprehensive assessment system that informs decisions "from the statehouse to the schoolhouse" (Gong, 2010[2]). As part of this improvement initiative there has been a growing movement around and interest in new assessment technologies and approaches, including immersive, game- or simulation-based assessments (GBAs) (DiCerbo, 2014[3]; Shaffer et al., 2009[4]; Shute, 2011[5]). As we discuss below, these new approaches take advantage of the increasing prevalence of educational technology – primarily personal computers and high-speed connectivity – in schools, as well as advances in psychometrics, computerised assessment design, educational data mining, and machine learning/artificial intelligence available to test makers.

In education, traditional standardised assessments have long been dominated by a model centred on collections of discrete questions (or "items") designed to cover content in an assessment framework by addressing parts of the domain to be measured (Mislevy et al., 2012[6]). GBAs, on the other hand, aim to blur the line between traditional assessment and more engaging learning activities through the use of games and simulations designed to measure constructs in an environment that maximises "flow" and rewards students for demonstrating their cognitive processes in more engaging and authentic situations, not just their ability to memorise key facts (Shute et al., 2009[7]).

While traditional educational assessments are designed to generally meet standards of technical quality in areas like *validity* (does the assessment measure what it is supposed to measure?), *reliability* (does it do this consistently and with minimal error?), and *fairness* (is the assessment culturally sensitive, accessible, and free of bias against any groups of test-takers?), they have nevertheless been criticised from a range of perspectives. We briefly review each of the following specific criticisms of traditional standardised assessment (e.g. Sanders and Horn, 1995[8]) and how game-based assessment may ameliorate them:

- the need to apply modern psychological theory to assessment;
- insufficient alignment of assessment with curriculum and instruction (Duncan and Hmelo-Silver, 2009[9]);
- lack of integration of assessments for different purposes, including formative, interim, and summative (Perie, Marion and Gong, 2009[10]);
- inability of traditional assessment to measure some important and increasingly policy-relevant constructs (Darling-Hammond, 2006[11]), and;
- declines in student engagement and motivation (Nichols and Dawson, 2012[12])

Application of modern psychological theory to assessment

The seminal volume *Knowing What Students Know* (National Research Council, 2001[13]) brought cognitive theory into the assessment realm using a framework accessible to teachers and policy makers. It called for an examination of mental functions involved in deep understanding, concepts that are difficult to assess with the sort of short, disconnected questions typical of standardised tests (Darling-Hammond et al., 2013[14]). New task types (or assessment items frequently used in classrooms but not on standardised tests) requiring complex performance on more realistic tasks were called for, including essays, projects, portfolios, and observation of classroom performance. Games and simulations have become more central due to the potential for eliciting evidence of deeper understanding and cognitive processes. Interpretation of streaming data from gameplay or interaction with a carefully-designed digital user interface allows researchers to evaluate how people go about solving problems and can lead to more targeted feedback (Chung, 2014[15]). For example, modern academic content standards in science increasingly require students to learn and demonstrate scientific practices (i.e. that they can think and reason like a scientist) as well as science facts. Accordingly, game-based-assessment allows the test maker to build scenarios and simulations where the students' reasoning and process can be observed through their complex interactions with elements in the game or simulation.

The need for assessment to be better aligned with curriculum and instruction

In keeping with the push to improve education, curriculum has seen a shift, incorporating learning theory and evidence-centred design approaches that provide students with grounding phenomena and hands on examples (Mislevy et al., 2012[6]; Arieli-Attali et al., 2019[16]). However, traditional standardised assessments have remained relatively stagnant, providing only limited information for teachers and learners, and furthering the divide between what is learned (content of curriculum) and what is tested (content of assessments) (Martone and Sireci, 2009[17]). As researchers and policy makers continue to call for new assessment frameworks that incorporate theories of learning and foundational transferrable skills consistent with classroom activity (National Research Council, 2012[18]; Darling-Hammond et al., 2013[14]; Conley, 2018[19]), this has led to increased interest in the development of games, simulations, and intelligent tutoring systems designed around learning progressions or specific instructional units.

Increasing assessment coherence

Assessments are broadly categorised by their purpose: how are scores used and interpreted? *Summative* tests are given at the end of instruction to evaluate what has been learned. Examples of summative assessment applications in education include annual large-scale accountability tests and college entrance exams, but also "drop from the sky" monitoring tests like PISA, TIMSS, and various national assessments (Oranje et al., 2019[20]). Summative assessments may be high-stakes for students (college entrance exams or graduation tests) but often are low-stakes for students but higher-stakes for other actors in the education system. Interim tests are given during the instructional period to evaluate progress toward summative goals and suggest instructional changes. Formative tests are also given during instruction but are closely linked to specific instruction and individual performance. Unlike interim assessments which can be aggregated at various education levels and are related to broad summative goals, formative assessments are adjusted to individual needs and to immediate teaching strategy (Shepard, Penuel and Pellegrino, 2018[21]). Each of these levels of educational assessment has different purposes and often requires appropriate measurement models and validation methods (Ferrara et al., 2017[22]).

The amalgamation of all these various types of assessment can create confusion for educators and parents and often comes at the expense of students' instructional time. Accordingly, there is growing interest in rationalising this confusing and fractured system. In the United States, for example, as theoretically-based, instructionally-relevant assessments have become more prominent, there have been widespread calls for "coherence" in the assessment enterprise across all levels (Gong, 2010[2]; Marion et al., 2019[23]). That is, policy makers and educators increasingly want all assessments that students encounter throughout the school year to function together as a single, coherent system.

While games and simulations in assessment have been most often targeted at the formative level, recent advances in development and scoring have made their use in large-scale summative tests in national accountability systems and international comparisons more feasible (Verger, Parcerisa and Fontdevila, 2019[24]; Klieme, 2020[25]). For example, in a coherent system, an immersive GBA could be used in a variety of ways. Formatively, the GBA could provide continuous feedback and personalised suggestions in the course of instruction. As an interim measure, the student can be assessed under more standardised simulation conditions to gauge progress toward summative goals. In summative mode, the student could be presented with a novel but related GBA scenario to solve without formative supports, allowing for a better understanding of what students have learned and are able to do. Box 10.1 highlights an example for assessing vocational skills in Germany.

Measuring different constructs - "hard-to-measure" skills

Another critique of traditional standardised assessments is that they are ineffective in measuring knowledge, skills, and abilities beyond very simple content in very circumscribed domains (Madaus and Russell, 2010[27]). For example, while a traditional standardised test may be a valid, reliable, fair, and efficient way to measure algebra, it may not be a modality suitable for measuring constructs like creative thinking or collaborative problem solving. This is an especially relevant critique in education for two reasons. First, modern curricular frameworks around the world increasingly are multidimensional, including cross-cutting skills as well as more traditional academic content. For example, the Next Generation Science Standards in the United States (www.nextgenscience.org/) include not only disciplinary core concepts, but also cross-cutting ideas in science, and scientific and engineering practices. Second, there is growing realisation among international policy makers of the importance of so-called

"21st Century Skills" or skills associated with "deeper learning," such as critical thinking, communication, collaboration, and creativity (Trilling and Fadel, 2009[28]; Vincent-Lancrin et al., 2019[29]; Fadel, Bialik and Trilling, 2015[30]). The use of games or simulations is a very promising way to assess these complex constructs either as part of a revised curricular framework or as a novel addition to the content covered by the usual standardised tests (Stecher and Hamilton, 2014[31]; Seelow, 2019[32]).

Box 10.1 Using simulations to train and assess vocational skills in Germany

With the ASCOT+ projects, Germany's Federal Ministry for Education and Research is supporting the development of digital training and assessment for vocational skills in different domains (car mechatronics, problem solving for technical systems, commercial problem solving, inter-professional and socio-emotional skills in nursing). In addition to digital training units using videos and simulations, the project is developing assessments that will be used as exams to certify apprentices' skills. For example, in the domain of commercial professions, a competency-oriented assessment task creator is being developed to allow assessors to design exams that certify students' and workers' competences, leading to a shift from knowledge-based to competence-based examination. An assessment bank of digital examination tasks that can be slightly modified or combined is proposed for assessors' customisation. It will be launched (and legally recognised for exams in Germany) in 2022. In the domain of car mechatronics, examination tasks are also developed to test trainees' competences in a simulated environment – and also to develop their skills.

Source: Bundesministerium für Bildung und Forschung (n.d.[26])

Measuring constructs differently - interaction and engagement

It goes without saying that most test-takers do not enjoy the traditional assessment experience (Nichols and Dawson, 2012[12]; Madaus and Russell, 2010[27]). One of the attractions of game-based assessment – beyond those noted above – is the promise of delivering valid and reliable measurement of complex constructs while bringing some of the engagement and immersive properties of modern video games. While there is growing evidence supporting this benefit across a broad range of operational game-based assessments (Hamari et al., 2016[33]), it is important to remember the inherent difference in purpose between games played for enjoyment versus those used for measurement (particularly, but not limited to, those used in high-stakes contexts). In addition, the need for GBA to meet assessment's more stringent scientific criteria of validity, reliability, and fairness, means that the transferability of engagement and immersion may be somewhat limited or at least different in nature (Oranje et al., 2019[20]). Simply put, game-based assessments might not be as fun as "real" games.

We now turn to a closer examination of the features of GBA and a brief discussion of how to design game-based tests.

How do we build game-based tests?

Designing from the ground up

Using games and game-based features as a means to increase engagement and capture hard-to-measure constructs is not a new idea (Cordova and Lepper, 1996[34]). However, the assessment field's knowledge and understanding of how best to implement this type of assessment and how to best use the data it provides continues to grow and mature. There are many ways to incorporate games and game-based features into a system or assessment that have varying impact on the learner. Thus, building a GBA requires forethought about the exact types of features and their potential impact on the learner and data collection (Shute and Ventura, 2013[35]).

The assessment designer must determine, a priori, exactly what the game is attempting to measure and how each game-based element provides evidence that allows such measurement. This includes storyboarding out the measures of interest, determining the evidence needed to capture them, and the exact quantification of that evidence. As Mislevy (2018[36]) notes, *"the worst way to design a game- or simulation-based assessment is to design what seems to be a great game or simulation, collect some observations to obtain whatever information the performances provide, then hand the data off to a psychometrician or a data scientist to 'figure out how to score it.'"* While there are benefits to post hoc exploratory analyses, they should not be the driving mechanism for how one scores the assessment. Before the assessment design team starts to develop the game specifications, they must first outline what they intend to measure and how this will be accomplished. This includes the quantification of evidence and scales that will be used.

This important foundational work cannot be done post hoc as this will often result in poor psychometric performance or lack of interpretability. For example, while adapting an existing game for use as an assessment may, at first glance, appear to generate a large amount of data for each test-taker, it is often the case that such data may yield items or measurement opportunities that are poorly-aligned to the desired content domain, exhibit high intercorrelation (rendering many of them useless), or are at the wrong level of difficulty (i.e. too easy or too hard for the target population). Therefore, the item design process should take place nearer to the beginning of the entire project, as designing a GBA takes a significant amount of forethought and discipline – and mistakes can be very costly.

However, this is not to say that analysis of actual test-taker data in the GBA development process is not important. Not only should the designers conduct traditional empirical psychometric analyses necessary to create valid and reliable assessments, they should also take advantage of the wealth of additional data generated by GBA to apply novel methods from domains like machine learning to extract more useable information about test-takers' ability or other constructs where possible (e.g. Gobert, Baker and Wixon, 2015[37]).

Games for assessment vs. "gamification"

We draw an important distinction here between designing games or simulations explicitly for measurement purposes and "gamification" or the addition of game-like elements to existing tasks or activities to increase engagement, flow, or motivation (Deterding et al., 2011[38]). An example of gamification would be adding a leader board, badges, personalised avatars, or progress bars to classroom activity and, while this may be useful in improving student engagement or motivation, it is not the sort of designed game-based assessment we discuss here. It is important to note, however, that the distinction between GBA and gamification is somewhat more grey in the assessment of social-emotional learning in education and "non-cognitive" skills in education and workplace selection. Nevertheless, it is still possible to assess these other types of skills and dispositions via designed games (Yang et al., 2019[39]) and not merely the addition of game-like elements to traditional tests.

Telemetry and the question of "stealth"

Game-based or simulation-centred assessments collect a wealth of data that is often missed or unable to be captured by traditional tests – sometimes "stealthily" or unbeknownst to the test-taker (Shute and Ventura, 2013[35]). This includes patterns of choices, search behaviours, time-on-task behaviours, and, in some cases, eye movement or other biometric information. These rich data sources can be used to help illustrate the cognitive processes that a student engages in as they complete a task (Sabourin et al., 2011[40]; Snow et al., 2015[41]), rather than just focusing on the end product of their performance. However, in order to collect and quantify this information, GBA developers need to carefully prescribe the data that the system collects, often referred to as "telemetry." This process involves mapping out every action a user can take during the design phase and assigning that action a value or name in the data infrastructure. This can most often be accomplished using data collection or measurement frameworks such as Evidence-Centred Design (ECD) (Mislevy et al., 2012[6]). Successful mapping of telemetry to measurement objectives requires a concentrated effort between designers, software engineers, and measurement scientists. As with any assessment, stakeholders should feel confident in what is being measured and how. To use telemetry in educational assessment – particularly in high-stakes and summative applications – we need to be very clear about the actions we are capturing, their interpretation, and how they should be stored and quantified.

How hard is this to do? How do costs compare to more traditional approaches?

Our experience suggests that building valid, reliable, and fair game-based assessments is considerably more complex and challenging than traditional test development. Success requires an interdisciplinary team with a broad range of skills, including game designers, software engineers ideally with a background in game, and cognitive scientists, as well as the test designers, content experts, educational researchers, and psychometricians usually needed to develop an assessment. For this reason, building GBAs is relatively expensive and is thus not always an efficient way to measure simple constructs. For example, while the benefits of GBA have led several operational assessment programs, such as PISA and the U.S. National Assessment of Educational Progress (NAEP), to add game or simulation components, due to cost they have done so in a limited fashion as part of a hybrid approach combined with more traditional item types and assessment strategies (Bergner and von Davier, 2018[42]).

An additional challenge to consider with GBAs is the need to make them accessible for students with disabilities. While the field of educational measurement has made significant progress in this area in recent decades, extending frameworks like Universal Design for Learning (Rose, 2000[43]) to game-based assessment requires careful design, extensive testing, and, in some cases, the invention of new approaches and novel technologies such as haptic feedback technologies that enable the implementation of touch-based user interfaces allowing the assessment of visually-impaired students (Darrah, 2013[44]).

New psychometric methods – and challenges

In addition to requiring a broader range of technical expertise, GBAs can also require innovation in technologies or statistical approaches to measurement. For example, psychometricians have suggested new measurement models reflecting task complexity (Mislevy et al., 2000[45]; Bradshaw, 2016[46]; de la Torre and Douglas, 2004[47]). These new models and others in development are intended to better address theories of cognition and learning and to capture complex latent ability structures. They provide measurement models appropriate to the new data streams generated by games and simulations.

Game-based assessment in education also brings new fairness and equity concerns. For example, differential access to computers in the home or school environment as well as (possibly gendered) differences in familiarity with video game mechanics or user interface components could exacerbate existing achievement gaps or create new ones. Part of the responsible development of GBA is to monitor these gaps and also to minimise differential item functioning (DIF – defined as when items don't behave as expected for test-takers of the same ability but different backgrounds) for both the usual subgroups (gender, ethnicity, language status) but potentially new ones like gaming experience (see Box 10.2). One key design element that reduces risk of differential item functioning in game-based assessment is the design of effective tutorials within each game or simulation that quickly teach the necessary game mechanics to those test-takers possibly less familiar with the user interface.

The promise of AI and machine learning

Beyond psychometric innovation, game- and simulation-based assessment also poses new opportunities for technical innovation based on recent developments in machine learning and artificial intelligence (Ciolacu et al., 2018[49]). For example, in high-stakes applications requiring multiple forms of an assessment to ensure security, computationally-intensive artificial intelligence algorithms (AI) enable the calibration of difficulty of variants of a game as part of the equating process required to ensure fairness for all test-takers. In other words, the AI can be used to "play" all of the proposed variations of the GBA as means of increasing the likelihood that they are all comparable in difficulty before moving to expensive and time-consuming pilot testing with human test-takers.

More broadly, similar AI play, as well as the application of machine learning techniques to large pilot datasets of game performance, can be used as part of the process of deriving meaningful information from telemetry data logs (i.e. the data collected during the assessment game/simulation process). That is, a key part of the development of game-based assessment should include a stage where item scores are refined and improved as via exploratory data analysis and educational data mining as larger amounts of test-taker data become available. Although this data mining should not replace the design process described above, experience suggests that computer-aided iteration here can improve the reliability and efficiency of game-based assessment by increasing the amount of useful information on test-taker performance available (Mislevy et al., 2014[50]).

Box 10.2 Choice of technology matters for gender equity: Insights from an experiment in Chile

Evidence from an experiment in a public school in Santiago suggests that gender differences in learning in educational games may depend on the technological platform used (Echeverría et al., 2012[48]). This may also apply to performance in game-based assessment. In the experiment, 11th grade students played First Colony, an educational game that requires students to apply concepts from electrostatics. Players assume the role of astronauts sent on a mission to bring back a precious crystal. Since the crystal is fragile, the astronauts can only move it using electrical force. In the version of the game implemented on a platform with multiple computer mice, students play in groups of three with each student controlling one mouse. With the mouse, students can move their astronaut, change their charge value and polarity and activate their charge to interact with the crystal. In the augmented reality version, students can perform the same actions using a tablet. Here, the classroom blends with the game world: each desk is covered with a set of markers that allow the augmented reality system to place virtual objects over the desks. Using the webcam at the top of the screen, the system determines the location of each student's astronaut by detecting the relative position of each student to the paper markers. While no gender differences in performance were observed when students played using the multiple-mice platform, boys outperformed girls when playing the same game using an augmented reality platform, with a statistically significant difference.

The results from the experiment revealed statistically significant differences in performance between boys and girls after they used the augmented reality platform. Given that there was no gender difference following the use of the multiple-mice version of the game, this suggests that the choice of platform can create a gender gap in learning that is unrelated to the game. Since girls seem to struggle more to use the augmented reality platform, it is possible that using the technology for GBA would put them at a disadvantage. Educators should select the technology to be used in GBA with care, as their choices may have unintended effects.

Some examples of game-based assessment in education

SimCityEDU: Pollution Challenge (GlassLab)

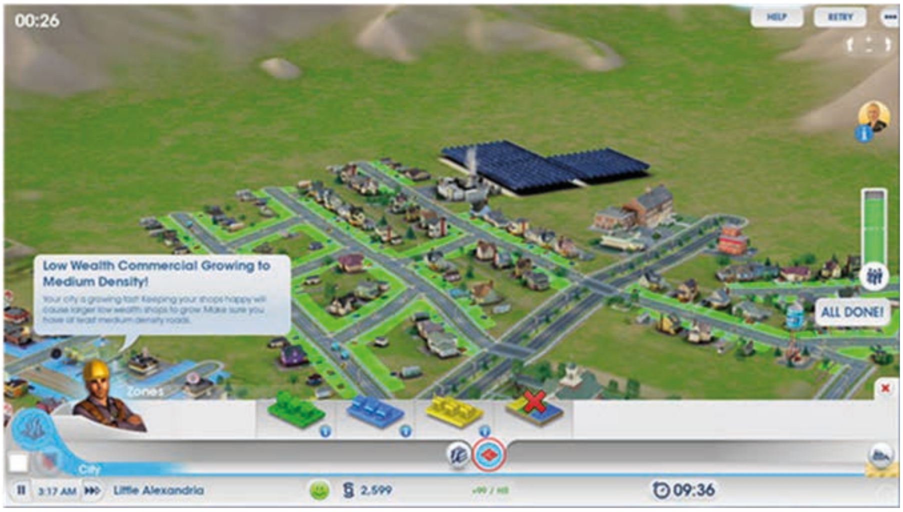
SimCityEDU: Pollution Challenge was a GBA released in 2014 by GlassLab, a collaborative development initiative funded by the John D. and Catherine T. MacArthur and Bill and Melinda Gates Foundations. It built on the design of the popular SimCity game series and put the test-taker in the role of a virtual city's mayor, tasked with balancing economic growth and environmental protection over a series of four levels of increasing complexity. SimCityEDU was designed as a formative assessment of problem solving, systems thinking, and causality in systems for students at approximately the ISCED-2 level.

The assessment content was explicitly aligned to the Framework for 21st Century Learning and Council for Economic Education standards as well as to aspects of the United States' Next Generation Science Standards and Common Core State Standards in Mathematics. Much as in the source game, problem-solving tasks were very engaging and largely spatial and economic in nature. GlassLab also devoted considerable resources to solving issues such as "tutorialisation" and telemetry processing to create useful assessment items as well as pioneering new psychometric models to support inference and reporting (Mislevy et al., 2014[50]; Mislevy, 2018[36]).

Crisis in Space (ACTNext)

In Crisis in Space, ACTNext developed a pilot version of a GBA designed to assess the collaborative problem solving and related socioemotional skills of middle school (ISCED-2) students. In this game, a pair of two test-takers is tasked with working together to troubleshoot a series of problems on a space station, with one of them in the role of an astronaut on the station and the other in mission control on the ground. By having the students actually collaborate in a cooperative game, Crisis in Space delivers an authentic and engaging experience and improves upon earlier attempts to measure collaboration via student-"agent" (chatbot) interaction.

Figure 10.1 SImCityEDU: Pollution Challenge (GlassLab)



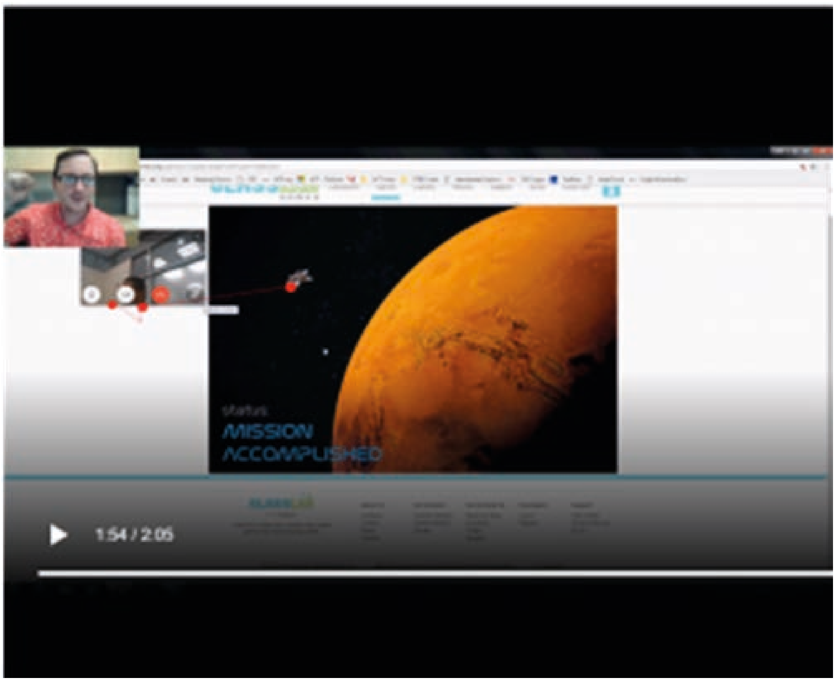
Low Wealth Commercial Growing to Medium Density!

Your city is growing fast! Keeping your shops happy will cause larger low wealth shops to grow. Make sure you have at least medium density roads.

Note: Introduced in 2013 by the now-defunct GlassLab, SimCityEDU: Pollution Challenge was a transformation of the popular SimCity videogame franchise into an assessment of middle school (primarily ISCED 2). The test-taker engages with a modified version of the SimCity interface to solve various urban issues. Telemetry data are processed via sophisticated psychometric models.
Source: Games for change (n.d.[50]); Glasslab (n.d.[51])

Crisis in Space, which won the innovation prize at the 2020 e-Assessment Awards, is particularly notable for its use of a wide range of data types, including user interface-generated telemetry, audio recordings of student conversation, and test-taker eye-tracking data. ACTNext also has implemented advanced machine learning technology such as natural language processing (NLP) to process these data and score instances of collaboration as successful or unsuccessful (Chopade et al., 2019[53]).

Figure 10.2 Crisis in Space (ACTNext)



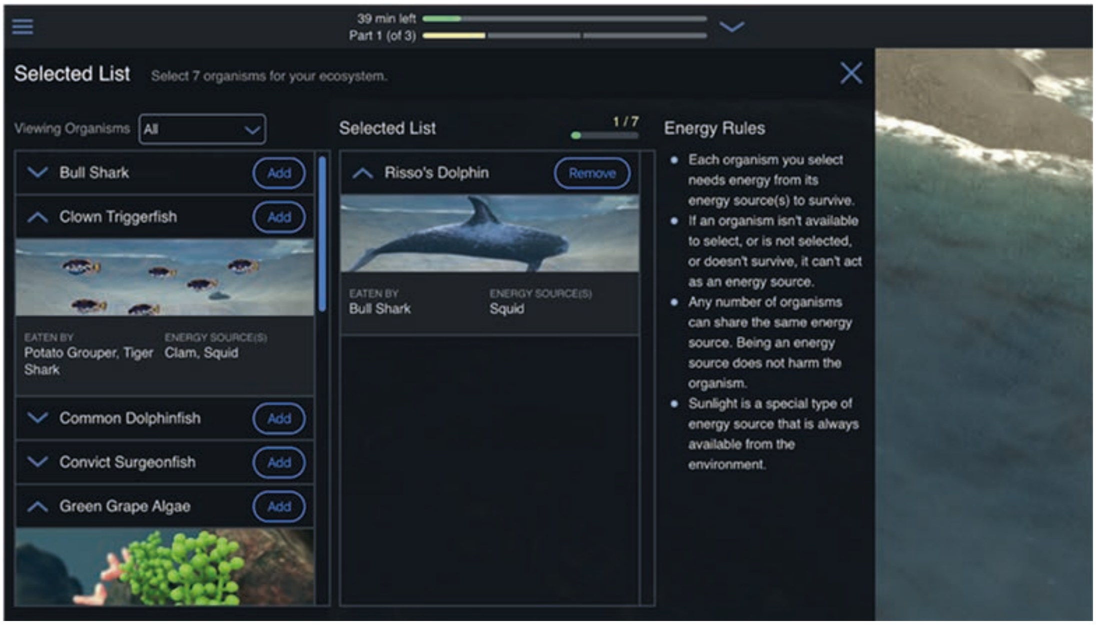
Note: Crisis in Space is a pilot game-based-assessment under development by ACT, Inc. as part of an ongoing program of research and development in collaborative problem-solving assessment by their research arm, ACTNext. In the scenario, two players work together to troubleshoot a space station. Technologies used for measurement include eye-tracking and natural language processing.
Source: https://actnext.org/collaboration-assessment-online-games/ ; https://actnext.org/research-and-projects/cps-x-crisis-in-space/ (reproduced with permission)

PEEP – Project Education Ecosystem Placement (Imbellus)

The third example game-based assessment, Project Education Ecosystem Placement (PEEP) is also in the pilot phase and is intended to measure problem solving in the ISCED-2 and 3 population via a game where test-takers are required to construct a viable food web or ecosystem and place it in the natural environment where it can thrive. PEEP, funded by the Walton Family Foundation, is an adaptation of a game-based assessment originally designed for employment selection that is currently used by the global consultancy, McKinsey and Company, to select new business analysts.

The education version has been adapted to reflect more accurate life sciences content as well as made developmentally appropriate for students. PEEP is designed to be eventually used in high-stakes, summative assessment and supports the creation of many parallel forms or versions to improve test security. To develop these, PEEP uses an algorithm to create viable ecosystem solutions of approximately equivalent difficulty based on a large library of organisms. PEEP can also be delivered as a "stage adaptive" assessment task where test-takers are presented with a series of problems to solve whose difficulty varies algorithmically depending on prior performance.

Figure 10.3 PEEP - Project Education Ecosystem Placement (Imbellus)



39 min left
Part 1 (of 3)

Selected List Select 7 organisms for your ecosystem.

Viewing Organisms: All

Clown Triggerfish
EATEN BY: Potato Grouper, Tiger Shark
ENERGY SOURCE(S): Clam, Squid

Selected List 1 / 7

Risso's Dolphin
EATEN BY: Bull Shark
ENERGY SOURCE(S): Squid

Energy Rules

- Each organism you select needs energy from its energy source(s) to survive.
- If an organism isn't available to select, or is not selected, or doesn't survive, it can't act as an energy source.
- Any number of organisms can share the same energy source. Being an energy source does not harm the organism.
- Sunlight is a special type of energy source that is always available from the environment.

Note: Designed to be used as part of a longer assessment of problem solving, the PEEP task challenges students to build a viable ecosystem and place it in a natural environment where it can thrive. The assessment task was adapted from an assessment created by Imbellus for use in employment selection with support from the Walton Family Foundation and in partnership with Summit Public Schools and several other school systems. It uses student telemetry to create a range of both process and product measurement opportunities suitable for scoring via item-response theory.
Source: Courtesy of Imbellus.

What is the long-term promise of this approach and what is necessary to get us there?

Educators, administrators, and policymakers should consider integrating GBA in their educational assessment systems, as it offers unique advantages over more traditional approaches. Game-based assessments are special because they can mirror the dynamic interaction, structural complexity, and feedback loops of real-world situations. In the long term, integrated assessment systems should rely on game- and simulation-based scenarios to evaluate how students integrate and apply knowledge, skills, and abilities. Robust scenarios can involve a subset of content from an academic domain of interest, but perhaps their greatest advantage lies in facilitating the measurement of 21st Century skills like problem solving and collaboration.

The advantages of GBA, including the ability to assess historically hard-to-measure cognitive processes, better alignment with modern curricula, and increased student engagement in the measurement process, make it an important part of the future of all educational assessment systems. However, game-based approaches often do not produce as many useable item scores as we might hope given their relatively high development cost when compared to more traditional, discrete items. Thus a cost-effective and robust system of assessments might use game-based scenarios in combination with traditional and less costly discrete technology-enhanced assessment items. A good design principle for such an assessment system would be to use relatively inexpensive, traditional assessments where feasible (e.g. measuring proficiency with academic content) and reserve the more expensive GBA scenarios and simulations for the measurement of more complex cognitive constructs. Moreover the use of GBA should not be limited to summative assessment alone but should instead be part of a coherent system of assessment throughout the academic year. Such an efficient, hybrid system of assessment could theoretically be designed for many uses, including accountability reporting, driving local instruction, and individual student growth modelling.

In order to realise the promise of game- and simulation-based assessment at the national level, education ministries need to invest in the infrastructure needed to design, implement, and operationally deliver such tests. While some of this capacity can be contracted out to private-sector vendors, successful implementation will require public capabilities as well. These include sufficient computer hardware in schools (although there is a growing trend to consider "bring your own device" policies) and a networking backbone capable of acceptable data transfer speeds.

References

Arieli-Attali, M., S. Ward, J. Thomas, B. Deonovic and A. von Davier (2019), "The Expanded Evidence-Centered Design (e-ECD) for Learning and Assessment Systems: A Framework for Incorporating Learning Goals and Processes Within Assessment Design", *Frontiers in Psychology*, Vol. 10, http://dx.doi.org/10.3389/fpsyg.2019.00853. [16]

Bergner, Y. and A. von Davier (2018), "Process Data in NAEP: Past, Present, and Future", *Journal of Educational and Behavioral Statistics*, Vol. 44/6, pp. 706-732, http://dx.doi.org/10.3102/1076998618784700. [42]

Bradshaw, L. (2016), "Diagnostic Classification Models", in *The Handbook of Cognition and Assessment*, John Wiley & Sons, Inc., Hoboken, NJ, USA, http://dx.doi.org/10.1002/9781118956588.ch13. [46]

Braun, H. and A. Kanjee (2006), "Using assessment to improve education in developing nations.", in Braun, H. et al. (eds.), *Improving Education Through Assessment, Innovation, and Evaluation.*, Cambridge, Mass.: American Academy of Arts and Sciences. [1]

Bundesministerium für Bildung und Forschung (n.d.), *ASCOT+*, https://www.ascot-vet.net (accessed on 30 May 2021). [26]

Chopade, P., D. Edwards, S. Khan, A. Andrade and S. Pu (2019), "CPSX: Using AI-Machine Learning for Mapping Human-Human Interaction and Measurement of CPS Teamwork Skills", *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, http://dx.doi.org/10.1109/hst47167.2019.9032906. [53]

Chung, G. (2014), "Toward the Relational Management of Educational Measurement Data.", *Teachers College Record*, Vol. 116/11. [15]

Ciolacu, M., A. Fallah Tehrani, L. Binder and P. Mugur Svasta (2018), "Education 4.0 - Artificial Intelligence Assisted Higher Education: Early recognition System with Machine Learning to support Students' Success", *2018 IEEE 24th International Symposium for Design and Technology in Electronic Packaging (SIITME)*, http://dx.doi.org/10.1109/siitme.2018.8599203. [49]

Conley, D. (2018), *The Promise and Practice of Next Generation Assessment.*, Harvard University Press, Cambridge, MA. [19]

Cordova, D. and M. Lepper (1996), "Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice.", *Journal of Educational Psychology*, Vol. 88/4, pp. 715-730, http://dx.doi.org/10.1037/0022-0663.88.4.715. [34]

Darling-Hammond, L. (2006), "Constructing 21st-Century Teacher Education", *Journal of Teacher Education*, Vol. 57/3, pp. 300-314, http://dx.doi.org/10.1177/0022487105285962. [11]

Darling-Hammond, L., J. Herman, J. Pellegrino, J. Abedi, J. Lawrence Aber, E. Baker, R. Bennett, E. Gordon, E. Haertel, K. Hakuta, A. Ho, R. Lee Linn, P.D. Pearson, J. Popham, L. Resnik, A. Schoenfeld, R. Shalveson, L. Shepard, L. Shulman and C. Steele (2013), "Criteria for High-quality Assessment.", *Stanford Center for Opportunity Policy in Education.* [14]

Darrah, M. (2013), "Computer Haptics: A New Way of Increasing Access and Understanding of Math and Science for Students Who are Blind and Visually Impaired", *Journal of Blindness Innovation and Research*, Vol. 3/2, http://dx.doi.org/10.5241/3-47. [44]

de la Torre, J. and J. Douglas (2004), "Higher-order latent trait models for cognitive diagnosis", *Psychometrika*, Vol. 69/3, pp. 333-353, http://dx.doi.org/10.1007/bf02295640. [47]

Deterding, S., D. Dixon, R. Khaled and L. Nacke (2011), "From game design elements to gamefulness", *Proceedings of the 15th International Academic MindTrek Conference on Envisioning Future Media Environments - MindTrek '11*, http://dx.doi.org/10.1145/2181037.2181040. [38]

DiCerbo, K. (2014), "Game-Based Assessment of Persistence.", *Educational Technology & Society*, Vol. 17/1, pp. 17-28. [3]

Duncan, R. and C. Hmelo-Silver (2009), "Learning progressions: Aligning curriculum, instruction, and assessment", *Journal of Research in Science Teaching*, Vol. 46/6, pp. 606-609, http://dx.doi.org/10.1002/tea.20316. [9]

Echeverría, A., M. Améstica, F. Gil, M. Nussbaum, E. Barrios and S. Leclerc (2012), "Exploring different technological platforms for supporting co-located collaborative games in the classroom", *Computers in Human Behavior*, Vol. 28/4, pp. 1170-1177. [48]

Fadel, C., M. Bialik and B. Trilling (2015), *Four-dimensional Education: the Competencies Learners Need to Succeed.*, Center for Curriculum Redesign, Cambridge, MA. [30]

Ferrara, S., E. Lai, A. Reilly and P. Nichols (2017), "Principled Approaches to Assessment Design, Development, and Implementation", in The *Handbook of Cognition and Assessment*, John Wiley & Sons, Inc., Hoboken, NJ, USA, http://dx.doi.org/10.1002/9781118956588.ch3. [22]

Games for change (n.d.), Games for change, http://www.gamesforchange.org/game/simcityedu-pollution-challenge/ (accessed on 30 April 2021). [51]

Glasslab (n.d.), *SIMCITY edu: pollution challenge*, https://s3-us-west-1.amazonaws.com/playfully-games/SC/brochures/SIMCITYbrochure_v3small.pdf (accessed on 30 April 2021). [52]

Gobert, J., R. Baker and M. Wixon (2015), "Operationalizing and Detecting Disengagement Within Online Science Microworlds", *Educational Psychologist*, Vol. 50/1, pp. 43-57, http://dx.doi.org/10.1080/00461520.2014.999919. [37]

Gong, B. (2010), *Using balanced assessment systems to improve student learning and school capacity: An introduction.*, Council of Chief State School Officers, Washington, DC. [2]

Hamari, J., D. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke and T. Edwards (2016), "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning", *Computers in Human Behavior*, Vol. 54, pp. 170-179, http://dx.doi.org/10.1016/j.chb.2015.07.045. [33]

Klieme, E. (2020), "Policies and Practices of Assessment: A Showcase for the Use (and Misuse) of International Large Scale Assessments in Educational Effectiveness Research", in *International Perspectives in Educational Effectiveness Research*, Springer International Publishing, Cham, http://dx.doi.org/10.1007/978-3-030-44810-3_7. [25]

Madaus, G. and M. Russell (2010), "Paradoxes of High-Stakes Testing", *Journal of Education*, Vol. 190/1-2, pp. 21-30, http://dx.doi.org/10.1177/0022057410190001-205. [27]

Marion, S., J. Thompson, C. Evans, J. Martineau and N. Dadey (2019), "A Tricky Balance: The Challenges and Opportunities of Balanced Systems of Assessment.", in *Paper Presented at the Annual Meeting of the National Council on Measurement in Education Toronto, Ontario April 6, 2019.*, National Center for the Improvement of Educational Assessment, https://www.nciea.org/sites/default/files/inline-files/Marion%20et%20al_A%20Tricky%20Balance_031319.pdf (accessed on 2 January 2020). [23]

Martone, A. and S. Sireci (2009), "Evaluating Alignment Between Curriculum, Assessment, and Instruction", *0*Vol. 79/4, pp. 1332-1361, http://dx.doi.org/10.3102/0034654309341375. [17]

Mislevy, R. (2018), *Sociocognitive Foundations of Educational Measurement.*, Routledge, New York. [36]

Mislevy, R., R. Almond, D. Yan and L. Steinberg (2000), "Bayes nets in educational assessment: Where do the numbers come from?", *(CSE Report 518). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).* [45]

Mislevy, R., J. Behrens, K. Dicerbo and R. Levy (2012), "Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining.", *Journal of educational data mining*, Vol. 4/1, pp. 11-48. [6]

Mislevy, R., A. Oranje, M. Bauer, A. von Davier, J. Hao, S. Corrigan, E. Hoffman, K. DiCerbo and M. John (2014), *Psychometric Considerations in Game-Based Assessment.*, Glasslab Games, Redwood City, CA. [50]

National Research Council (2001), *What Students Know: The Science and Design of Educational Assessment.*, National Academies Press, Washington, D.C., http://dx.doi.org/10.17226/10019. [13]

Nichols, S. and H. Dawson (2012), "Assessment as a Context for Student Engagement", in *Handbook of Research on Student Engagement*, Springer US, Boston, MA, http://dx.doi.org/10.1007/978-1-4614-2018-7_22. [12]

Oranje, A., B. Mislevy, M. Bauer and G. Tanner Jackson (2019), "Summative Game-based Assessment.", in Ifenthaler, D. and Y. Kim (eds.), Ga*me-based Assessment Revisited*, Springer. [20]

Pellegrino, J. and M. Hilton (eds.) (2012), *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. Committee on Defining Deeper Learning and 21st Century Skills.*, National Academies Press, Washington, D.C., http://dx.doi.org/10.17226/13398. [18]

Perie, M., S. Marion and B. Gong (2009), "Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments", *Educational Measurement: Issues and Practice*, Vol. 28/3, pp. 5-13, http://dx.doi.org/10.1111/j.1745-3992.2009.00149.x. [10]

Rose, D. (2000), "Universal Design for Learning", *Journal of Special Education Technology*, Vol. 15/3, pp. 45-49, http://dx.doi.org/10.1177/016264340001500307. [43]

Sabourin, J., J. Rowe, B. Mott and J. Lester (2011), "When Off-Task is On-Task: The Affective Role of Off-Task Behavior in Narrative-Centered Learning Environments", in *Lecture Notes in Computer Science, Artificial Intelligence in Education*, Springer Berlin Heidelberg, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-642-21869-9_93. [40]

Sanders, W. and S. Horn (1995), "Educational Assessment Reassessed", *education policy analysis archives,* Vol. 3, p. 6, http://dx.doi.org/10.14507/epaa.v3n6.1995. [8]

Seelow, D. (2019), "The Art of Assessment: Using Game Based Assessments to Disrupt, Innovate, Reform and Transform Testing.", *Journal of Applied Testing Technology,* Vol. 20/S1, pp. 1-16. [32]

Shaffer, D., D. Hatfield, G. Navoa Svarovsky, P. Nash, A. Nulty, E. Bagley, K. Frank, A. Rupp and R. Mislevy (2009), "Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning", *International Journal of Learning and Media,* Vol. 1/2, pp. 33-53, http://dx.doi.org/10.1162/ijlm.2009.0013. [4]

Shepard, L., W. Penuel and J. Pellegrino (2018), "Using Learning and Motivation Theories to Coherently Link Formative Assessment, Grading Practices, and Large-Scale Assessment", *Educational Measurement: Issues and Practice,* Vol. 37/1, pp. 21-34, http://dx.doi.org/10.1111/emip.12189. [21]

Shute, V. (2011), *Stealth assessment in computer-based games to support learning. Computer games and instruction.,* Information Age Publishers, Charlotte, NC, http://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf. [5]

Shute, V. and M. Ventura (2013), "Stealth Assessment: Measuring and Supporting Learning in Video Games.", in John, D. and C. MacArthur (eds.), *Foundation Reports on Digital Media and Learning.,* The MIT Press, Cambride, MA, http://dx.doi.org/10.7551/mitpress/9589.001.0001. [35]

Shute, V., M. Ventura, M. Bauer and D. Zapata-Riviera (2009), "Melding the power of serious games and embedded assessment to monitor and foster learning.", *Serious games: Mechanisms and effects,* Vol. 2, pp. 295-321. [7]

Snow, E., L. Allen, M. Jacovina and D. McNamara (2015), "Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment", *Computers & Education,* Vol. 82, pp. 378-392, http://dx.doi.org/10.1016/j.compedu.2014.12.011. [41]

Stecher, M. and L. Hamilton (2014), Measuring hard-to-measure student competencies: A research and development plan., RAND Corporation, Santa Monica, CA, https://www.rand.org/pubs/research_reports/RR863.html. [31]

Trilling, B. and C. Fadel (2009), *21st century skills: Learning for Life in Our Times.*, Jossey-Bass. [28]

Verger, A., L. Parcerisa and C. Fontdevila (2019), "The growth and spread of large-scale assessments and test-based accountabilities: a political sociology of global education reforms", *Educational Review,* Vol. 71/1, pp. 5-30, http://dx.doi.org/10.1080/00131911.2019.1522045. [24]

Vincent-Lancrin, S., C. Gonzalez-Sancho, M. Bouckaert, F. de Luca, M. Fernandez-Barrerra, G. Jacotin, J. Urgel and Q. Vidal (2019), *Fostering Students' Creativity and Critical Thinking: What it Means in School*, Educational Research and Innovation, OECD Publishing, Paris, https://dx.doi.org/10.1787/62212c37-en. [29]

Yang, F., L. Leqi, Y. Wu, Z. Lipton, P. Ravilkimar, W. Cohen and T. Mitchel (2019), "Game Design for Eliciting Distinguishable Behavior.", *Paper prepared for the 33rd Conference on Neural Information Processing Systems.,* https://papers.nips.cc/paper/8716-game-design-for-eliciting-distinguishable-behavior.pdf (accessed on 2 January 2020). [39]

From:

OECD Digital Education Outlook 2021

Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots

Access the complete publication at:

https://doi.org/10.1787/589b283f-en

Please cite this chapter as:

Buckley, Jack*, et al.* (2021), “Game-based assessment for education”, in OECD, *OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9289cbfd-en

This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.

The use of this work, whether digital or print, is governed by the Terms and Conditions to be found at http://www.oecd.org/termsandconditions.