CASE STUDY

Game-Based Assessment: Practical Lessons from the Field

Jack Buckley and Erica Snow

Game-Based Assessment: Practical Lessons from the Field

Jack Buckley and Erica Snow

Abstract

In this chapter we discuss a particular application of digital games for learning: game-based assessment (GBA). This approach to assessment allows for the measurement of a broader range of skills (e.g., "durable" skills such as creative problem solving and collaboration), as well as better measurement of some aspects of the "thinking" of respondents, including in traditional domains like science and mathematics or adult learning in the workplace. While promising, GBA is not without practical challenges.

For example, game-based assessments can often be more costly and difficult to develop than traditional standardized tests based on a series of discrete questions or small "testlets" or tasks. Despite this challenge, GBA is not infeasible or impractical; in fact, we have been developing GBAs for education and workplace applications for over seven years, including in the high-stakes workforce selection context. Here we draw from our hard-earned experience in this domain and share some lessons we have learned that may be helpful for the next wave of GBA developers.

Authors Note We would like to thank our current and past colleagues at Roblox, Imbellus, and Mckinsey & Co. who contributed to the work presented in this chapter.

Introduction

Games and learning have long been intertwined. While perhaps the earliest evidence of the use of games as a teaching tool dates at least to Classical Greece, if not to the creation of African board games some 5,000 years ago (Hellerstedt & Mozelius, 2019), the advent of digital computing marked the beginning of a new era of computer games and simulations in the service of learning.

The earliest digital learning games, such as "The Sumerian Game," developed for the IBM 7090 in 1964 (Wing, 1967) allowed learners to interact with and learn the principles of complex systems in a novel and engaging way, albeit handicapped by the technological limitations. In the subsequent decades, every advance in computing technology (e.g., home microcomputers, CD-ROM drives, the Internet, high-speed broadband, machine learning, educational data mining) have been harnessed almost immediately for learning. Simultaneously, the applications of these technologies spread across many domains and populations, from preschool mathematics to computer programming in the workplace.

Although this history is fascinating and holds many lessons for the educational content developer of today, in this chapter we concern ourselves with a narrower subset of the application of digital games for learning: *game-based assessment* (GBA). This approach to assessment allows for the measurement of a broader range of skills (e.g., "durable" skills such as creative problem solving and collaboration), as well as better measurement of some aspects of the "thinking" of respondents, including in traditional domains like science and mathematics or adult learning in the workplace.

While promising, GBA is not without practical challenges. For example, game-based assessments can often be more costly and difficult to develop than traditional standardized tests based on a series of discrete questions or small "testlets" or tasks. Despite this challenge, GBA is not infeasible or impractical; in fact, we have been developing GBAs for education and workplace applications for over seven years, including in the high-stakes workforce selection context. In the pages that follow, we will draw from our hard-earned experience in this domain and hopefully share some lessons we have learned that may be helpful for the next wave of GBA developers.

The remainder of this chapter is organized as follows: after a brief discussion of some preliminaries and definitions, we turn to a description of our GBA design process. We then illustrate that process with several real examples from our work at both Imbellus, a GBA startup, and Roblox, a gaming platform technology company. We share examples (and lessons) from both the K–12 education and workforce learning contexts. We conclude with some thoughts on the future of GBA.

Preliminaries

Why Game-Based Assessment?

In our experience there are two primary reasons to consider the development of a GBA instead of taking a more traditional (and often less costly) approach. The first is that, compared to traditional assessment, GBA can allow for *measuring different constructs.* Increasingly, in both P-20 education and in workforce learning and selection, there is significant interest in measuring "durable skills" (or "soft skills" or "21st Century Skills") such as critical thinking, communication, computational thinking, collaboration, systems thinking, and creative problem solving (Trilling & Fadel, 2009). The use of games or simulations (more on the distinction below) is a promising way of measuring these constructs (Stecher & Hamilton 2014; Seelow 2019).

Aside from durable skills, curricular frameworks in P-20 education around the world are increasingly multi-dimensional and include cross-cutting skills as well as traditional academic content. For example, the Next-Generation Science Standards (NGSS Lead States, 2013) in the United States include scientific practices and cross-cutting concepts as well as traditional scientific domain knowledge. These new dimensions can be difficult to assess via traditional means (Smith et al., 2022). As global education systems increasingly expand their curricular standards to include these kinds of constructs, there will be increasing demand for formative and summative assessments to keep pace.

The other reason to consider GBA is that the use of games allows the test developer to *measure constructs differently.* Even if one's task is to assess learners' knowledge of familiar and relatively uncomplex content such as traditional mathematics, vocabulary, or factual knowledge, the use of GBA can improve engagement and immersion (Hamari et al., 2016). This increased test-taker engagement can be particularly important in applications like pre-hire workforce assessment, where candidates are not a "captive audience" and can simply choose to exit the application process.

However, regardless of the domain, it is important to remember that GBA is not a panacea for differences in opportunity-to-learn. If learners do not have equal access to instruction in the basic building blocks of a given domain, layering a game into the assessment experience will not ameliorate this (Porter 2007). It is also worth noting that games played for enjoyment do not have to meet the test-maker's criteria of validity and reliability. GBA, while more engaging and immersive than a "bubble sheet" test, it is constrained in many ways (Oranje et al., 2019).

Game-Based vs. "Gamified"

In recent years the idea of "gamification" or the layering of game-like elements (e.g., leaderboards, badges, or personalized avatars) to non-game educational and assessment content and tasks (Deterding et al., 2011) has become pervasive. This practice may, indeed, increase learner engagement, but we draw a distinction between this gamification and the development of true games for learning and assessment. Citing Salen and Zimmerman's (2004) definition of a "game" as, "a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome," Plass, Homer, and Kinzer (2015) provide an example that illustrates the distinction between games and gamification:

Consider as an example the gamification of math homework, which may involve giving learners points and stars for the completion of existing activities that they consider boring. Game-based learning of the same math topic, on the other hand, even though it may also include points and stars, would involve redesigning the homework activities, using artificial conflict and rules of play, to make them more interesting and engaging. (Plass, Homer, & Kinzer, 2015, p. 259).

We appy the same distinction for the specific case of GBA, although it is not always easy to observe in practical application.

Games vs. Simulations

Finally, it may be useful to attempt to draw a similar distinction between games and various types of "simulations." While we are not aware of any broadly-accepted definition, the typology of Narayanasamy et al. (2006) is a useful one. They distinguish between "games," "simulation games," and, "training simulations." While the three have many aspects in common, there are two important distinctions among the categories. The first is in the area of goal-orientation. Simply put, games and simulation games are centered around goal-oriented activity, while training simulators are not. Further, games have an end state, while simulation games and training simulators continue without a determined end point (i.e., one does not "win" at Microsoft Flight Simulator).

Aside from durable skills, curricular frameworks in P-20 education around the world are increasingly multidimensional and include cross-cutting skills as well as traditional academic content.

The second area of difference among the categories is the presence or absence of a gameplay "gestalt," or pattern of interaction (perception, cognition, and motor performance) that allows for successful play (Lindley 2002). Games and simulation games both have patterns that allow for the creation of gameplay gestalts; training simulations have standard operating procedures that are well-defined and generally do not change.

Our GBA work generally seems to fall in the space between games and simulation games. The GBA tasks we have developed are goal-oriented (test-takers must complete various tasks that are transparent and quantifiable, although there are other item scores generated by their interaction with the game, as we discuss below) and allow for the formation of gameplay gestalt via patterns of perception and cognition.

Designing GBAs

The Use of Evidence-Centered Design

To develop our GBAs we use a modified version of Evidence-Centered Design (ECD; Mislevy, Almond, & Lukas, 2003), a well documented and validated approach to task design that has been used across a variety of domains and media (Frezzo, Behrens, & Mislevy, 2010; Liu & Haertel, 2011; Sweet & Rupp, 2012).

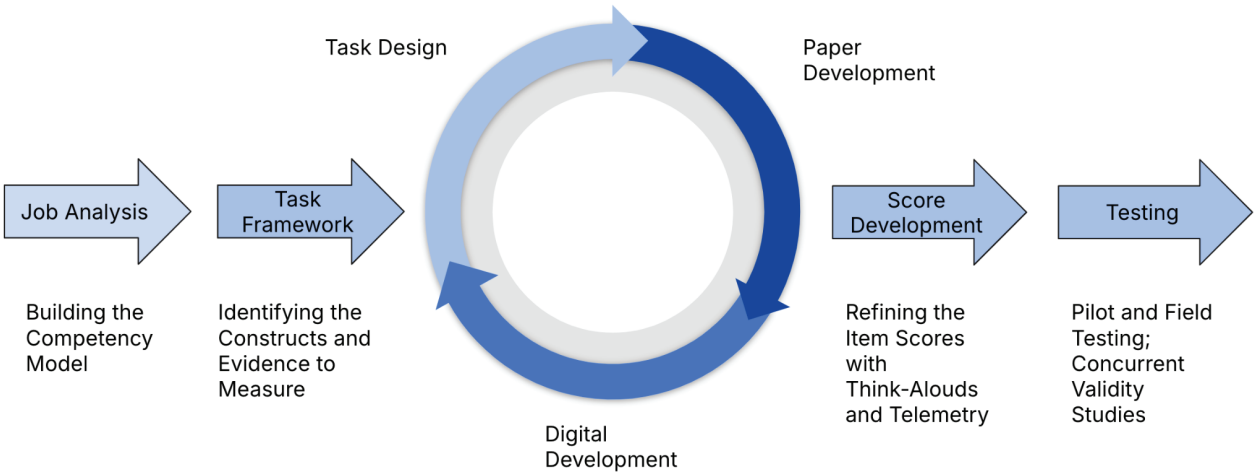
Our GBA development starts by identifying the constructs or KSAs (Knowledge, Skills, and Abilities) of interest. We identify these constructs or KSAs through cognitive task analysis or job analysis, which identifies the underlying skills, thinking, and abilities required to successfully perform a task and/or demonstrate a standard of knowledge. For hiring selection assessment these skills are often identified as key indicators of success at the company within the specific role.

Once we have conducted the job analysis and identified the target constructs/KSAs we begin to develop a task framework which will be used as a starting point for developing our GBAs. These frameworks help facilitate the collaboration between game designers, learning scientists, content experts, data scientists, and psychometricians by identifying 1) the primary KSAs that we as scientists and designers want to build the task around, 2) the specific pieces of evidence that need to be collected to capture the KSAs, and 3) the constraints and structures potential game-based tasks must include.

After our scientists and content experts develop a task framework, we bring in our game designers and UX/UI experts for iteration on creating possible GBA tasks that meet the requirements outlined in the task framework. Our scientists walk the design team through the task framework with a specific focus on the evidence we need to collect within a possible task. Then the design team begins to iterate on possible narratives/scenarios that could be used to build out the task. As we begin to map out the various task designs we start a prototyping process

that begins with paper prototypes and then shifts to digital prototypes as the work progresses. We conduct think-alouds (sometimes called cognitive labs) to gauge both usability issues with the possible tasks as well as "pressure test" the assumptions we are making about the types of thinking the task evokes and requires for successful completion.

Developing the GBA Tasks: A Modified ECD Approach



Stealth Assessment and Scoring

To score users' performance within our game-based tasks we take a stealth assessment approach to scoring (Shute, 2011). Stealth assessment provides an unobstructed view into the cognitive process of the user while they engage in the GBA. The user does not know what they are being scored on and, in most cases, it is not immediately obvious what is being measured. This allows for a more authentic view of their skills and abilities. We build our stealth assessments using the designed telemetry data generated by interaction with the task. That is, every item score is computed using test-takers' telemetry within the task. Telemetry captures the test-takers' every choice, behavior, timestamp, and click within the GBA. Every item score is pre-developed through the modified ECD process, not based on a "black box" modeling approach.

Development of item scores is a meticulous process that requires our interdisciplinary team to outline out how each potential behavior (or patterns of behaviors) maps to a specific construct and how that behavior can be transformed into an item score. Once an initial set of items is identified, we build preliminary pseudo-code for each of these items. This pseudo-code specifies algorithmically how different behaviors will be scored using the telemetry data generated by the actions players engage in the GBA. Item scores are tested throughout the prototyping process and at a full pilot stage. Data is collected and the team monitors overall item performance and construct coverage.

Evidence Centered Design (ECD) and stealth assessment provide frameworks for finding evidence of knowledge, skills or abilities in game-based assessments. This approach also can assist in combating cheating as it is not immediately clear within the game what the "right answer" is and often, there are many correct answers or ways that an item can be scored to give the test-taker full credit. This assessment approach within games allows an unobstructed look at a series of evidence identifying not only what a user knows, but the process they engaged in to get there.

Design Challenges

One of the biggest challenges in developing GBA is its interdisciplinary nature. While all cognitive assessment is (or should be) interdisciplinary to some degree (Pellegrino, Baxter, & Glaser 1999), successful development of GBA requires an exceptionally broad range of domain and disciplinary participation, including Learning Science, User Interface/User Experience (UI/UX) Design, Game Design, 3D Art, Software Engineering, Psychometrics, and Data Science (Table 1).

Table 1. *A Typical GBA Development Team*

Role	Quantity
Overall Lead	1
Project Manager	1
Learning or Cognitive Scientist	1–2
Industrial-Organizational Psychologist (workforce) or Content Expert (education)	1
Game Designer	1
3D Artist	1
Data Scientist	2
UI/UX Designer	1
Game Development Lead	1
Game Developer	1–2
Backend Engineer (if integrations required)	1
Psychometrician	1

No one discipline owns the entire process; instead there is a series of hand-offs throughout the development cycle that require high levels of attention to detail and constant communication. While our learning scientists kick the process off through construct identification and development of the design pattern, the first major handoff is to a game design team. This design team may or may not initially have experience in game-based assessments and what works in the world of game design for entertainment does not always work for assessment. As the designers build out a narrative, the data scientists and psychometricians need to have constant eyes on the design to make sure the evidence needed to develop item scores is included.

Often the design team will want to have flawless user experience in the UI/UX phases, however, that may result in poor measurement. For instance, when designing a guidebook for a task, from the UI/UX perspective it is a better user experience to have fewer clicks or choices to be able to access information, resulting in less friction for the player. However, for measurement we want to include added clicks and actions to be sure exactly what a user is looking at and how they decided to access that information. This can result in added layering or nesting of information.

These differences in philosophies often put disciplines at odds. Thus, iteration is present throughout the entire process from early design all the way to operational testing. This type of interdisciplinary work requires flexibility with everyone keeping an eye on the common goal, building a reliable and valid assessment. This goal can sometimes come in conflict with other goals such as user engagement, enjoyment, and experience.

Digital GBA at operational scale also requires an entire software engineering team, consisting of game developers and, possibly, backend engineers if the game-based task must be integrated into other reporting or analytics systems. Once again, until this team gains experience with peculiarities of GBA (compared to entertainment game development), there will likely be friction between them and the assessment science professionals.

Why Don't We Just Use Existing Games?

If designing game-based assessments is such an interdisciplinary challenge, why not simply adapt existing commercial (or academic) games for measurement in the classroom or workforce? Certainly performance on some existing games is correlated with the sorts of cognitive and durable skills we seek to measure. For example, Simons et al. (2021) show that business school students with higher scores on the award-winning commercial strategy game Civilization, "had better skills related to problem-solving and organizing and planning than the students who had low scores."

While we believe there could be some efficiencies in using existing games as assessment, we have four major concerns with this approach, especially in the high-stakes context:

1. Fairness: existing games are generally designed to be entertaining, not to ensure that all test-takers have an equal opportunity to demonstrate KSAs/competencies;
2. Content alignment: existing games are unlikely to be designed to allow evidence statements based on curriculum designers, employers' (or others') required competencies.
3. Construct-irrelevant variance: commercial games often have interlocking game systems and design elements that are uncorrelated with the constructs of interest and may be extremely distracting;
4. Time: amount of time available for selection at the top of a hiring funnel (or even in a college entrance examination) is limited compared to the time spent playing many existing games, so it can be difficult to generate item scores efficiently.

For these reasons, we generally advise teams building GBA to design their own experiences using a principled process like ECD.

We generally advise teams building game-based assessments to design their own experiences using a principled process like Evidence-Centered Design.

Fairness and GBA

One of the guiding principles for all assessment is fairness. As the sixth Principle of this Handbook states, "Assessment equity requires fairness in design of tasks and their adaptation to permit their use with respondents of different backgrounds, knowledge, and experiences." (Baker et al., 2025). The premise of testing is that tasks provide evidence of skill mastery for all examinees. If any factors unrelated to skill affect performance, assessment validity is diminished. Indeed, according to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p. 49), "fairness is a fundamental validity issue." In addition to the typical fairness areas of concern to all test makers, GBA introduces additional complexities. Chief among these is the need to ensure that background knowledge of and experience with games and gaming does not provide an unfair advantage to the test-taker.

One way to ensure that gaming experience does not create inequity is to measure test-takers' experience with games and conduct the same sorts of group difference and differential item functioning (DIF) analysis that one would usually conduct on sociodemographic categories like gender or primary language of instruction (or in the workplace). For example, in our work, we frequently capture the self-reported video game experience of our test-takers and construct a reference group of infrequent gamers (e.g., less than 10 hours played in the last 12 months) and a focal group of more frequent gamers. We then estimate quantities like item-level DIF, percent correct by group, and scale scores by group (including interactions with other sociodemographic factors) to ensure that we observe no substantively significant differences. If we detect DIF or see large group differences, we redesign item scores or even aspects of the GBA task as necessary to ameliorate.

Chief among these is the need to ensure that background knowledge of and experience with games and gaming does not provide an unfair advantage to the test-taker.

It is worth noting that game experience or familiarity does not always theoretically predict better assessment performance on GBA. One reason for this, which we have seen in practical application, can be explained by the aforementioned idea of gameplay gestalt (Lindley 2002). Simply put, very experienced gamers may develop ingrained perspectives about gameplay and possible game-states due to repeated play of other games. This can cause these test-takers to make incorrect assumptions about the GBA tasks by relying on this experience to categorize them, possibly leading to the use of suboptimal heuristics instead of appropriate cognition. If this effect is detected in testing, the GBA task may require substantial redesign.

Developers of game-based assessments must build a broader validity argument supporting particular uses of their assessments in the classroom or workplace.

Finally, another way of ensuring fairness of GBA for non-gamers is the familiar strategy of creating and disseminating test guides and practice materials—including actual playable practice GBA tasks to help familiarize non-gamers with the user interface and "feel" of game-based assessment and, as we discuss below, reduce test anxiety.

Validity and GBA

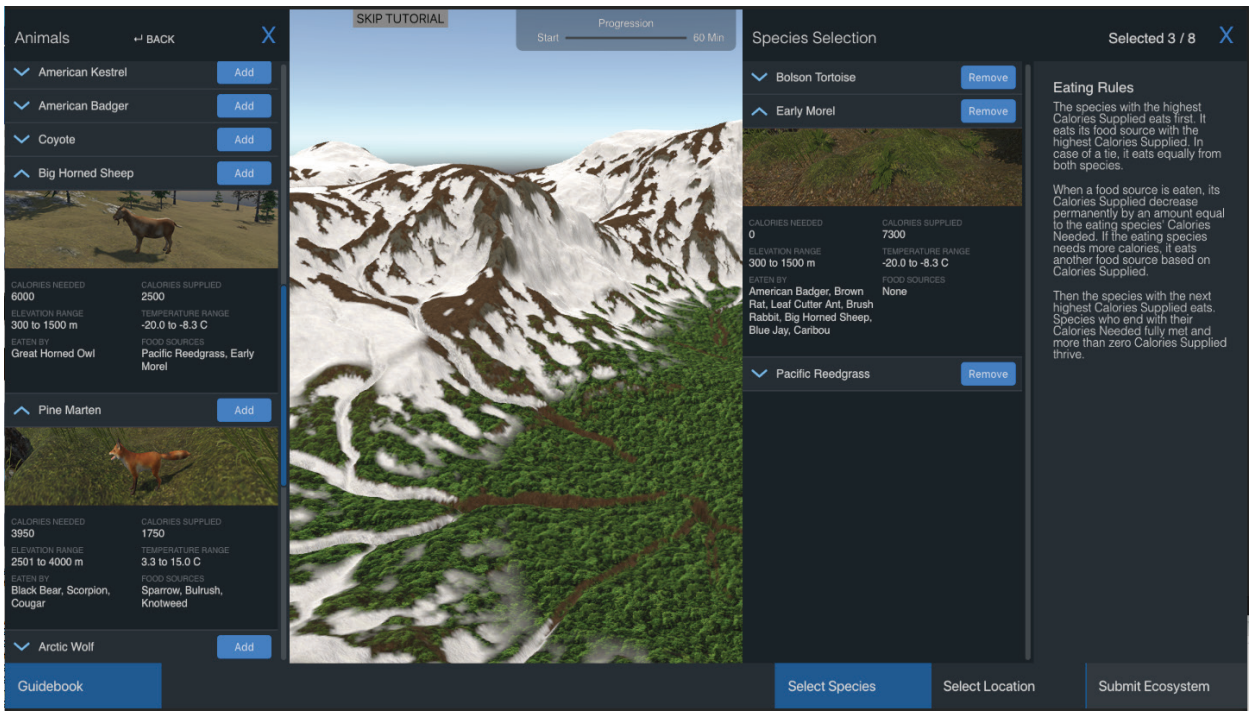
Beyond the important dimension of fairness, developers of GBA must build a broader validity argument supporting particular uses of their assessments in the classroom or workplace. As in the case of traditional assessment, this argument must cover the breadth of validity research, including but not limited to face validity, content and construct validity, concurrent and predictive validity, and consequential validity (Ferrara et al., 2016). Since GBA may be novel to both test-takers and classroom or workplace decision makers using the results, some types of validity may be challenging but important to demonstrate. We highlight some specifics in the examples below.

Examples of GBA: Imbellus

Before coming to Roblox, our team worked at a small GBA startup, Imbellus. Using the processes and techniques outlined above, we developed a hiring assessment to select new business analysts for the global consultancy, McKinsey and Company. For this assessment we had two primary tasks that were operational and part of the selection process: Ecosystem Placement (EP) and Pathogen Spread (PS). Both tasks were designed to measure cognitive skills that had been shown to be important for success at McKinsey.

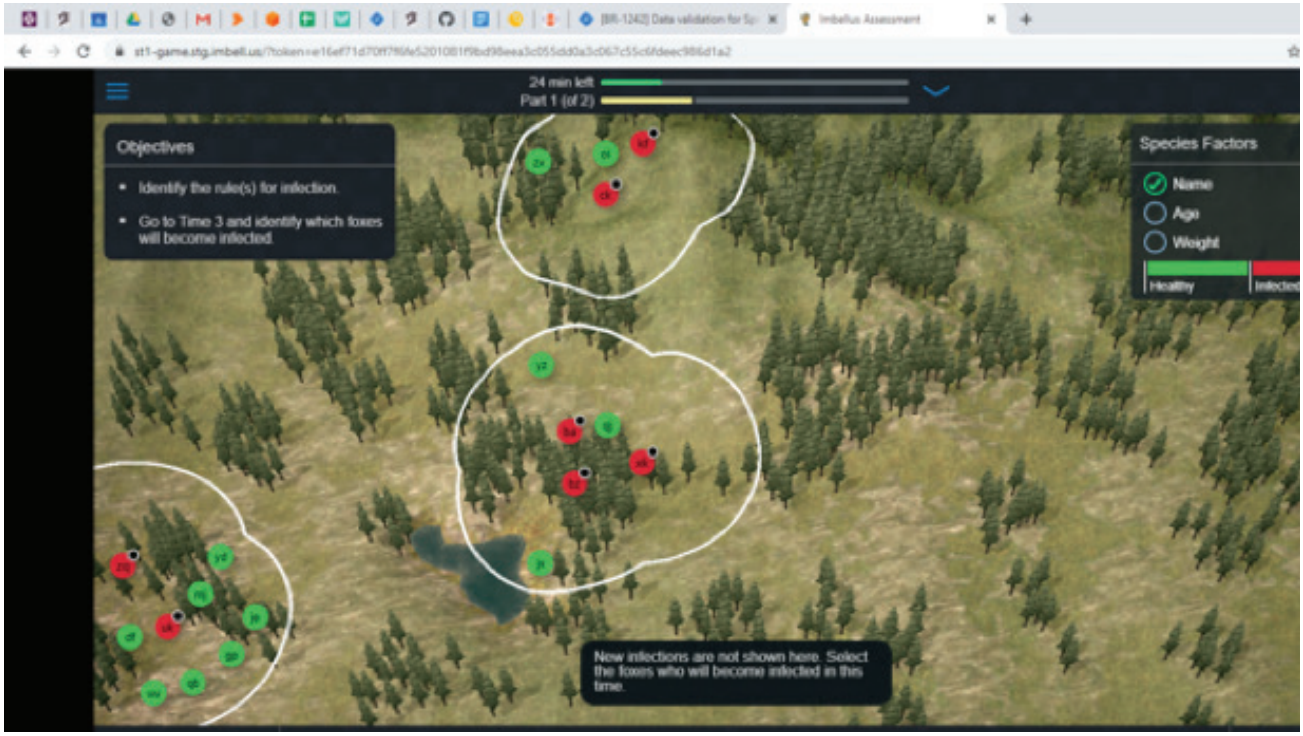
The Ecosystem Placement task measures test-takers systems thinking and situational awareness. In this task, test-takers are presented with a 3D landscape and given the goal to create a sustainable ecosystem within that environment. Test-takers are given a list of possible species that they can use to build out their ecosystem. Each species has caloric needs, environmental requirements, and predator-prey relationships that they must consider as they engage in the task.

Figure 1. *A screenshot of the Imbellus Ecosystem Placement Task.*



The Pathogen Spread task measures test-takers' situational awareness and reasoning ability. In this task, test-takers are presented with a scenario where a pathogen is spreading through an animal population. Test-takers are given the goal to predict the pattern of the pathogen based on evidence given to them within the scenario such as animals' infection statuses across time, space, and other variables such as age, weight, and temperature.

Figure 2. *A screenshot of the Imbellus Pathogen Spread Task.*



The Validity Argument for Ecosystem Placement and Pathogen Spread

While specific details of the validity research supporting the use of these tasks in hiring at McKinsey must remain confidential, we can provide an overview of the framework of the overall validity argument. Briefly stated, the argument demonstrates that:

- The assessment content is based on skills required by the job;
- The GBA tasks demand that players demonstrate these skills;
- This use of skills is observable and scored appropriately;
- The assessment structure reflects target content coverage;
- Recruiters are able to interpret and use assessment scores to make appropriate decisions;
- Applicants perceive the tasks as measuring relevant skills at the appropriate level of difficulty, and
- Scores on the assessment are associated with concurrent and predictive measures of candidate quality.

On the last point, during development, pilot/field testing, and operations, we were able to demonstrate concurrent validity through domain expert/novice contrast, correlations with existing instruments measuring at least part of the same domain (systems thinking, situational awareness, deductive reasoning), and predictive validity through comparing GBA performance to hiring outcomes and early job performance.

Adapting for Education

In 2019, we began to expand into the educational space by developing an adaptive, game-based assessment focused on life science content and science standards. PEEP—Project Education Ecosystem Placement was a staged adaptive GBA task aimed at measuring and providing feedback on problem solving processes for K–12 learners. Within PEEP, test-takers were asked to construct sustainable ecosystems based on the constraints of the game-based environment. PEEP was funded by the Walton Family Foundation, and was adapted from the original ecosystem placement test developed for McKinsey.

Unlike the industry version, PEEP was adapted to be more aligned and reflective of accurate life sciences content taught in schools, particularly a subsection of the Next Generation Science Standards (NGSS Lead States 2013). It was also designed to be developmentally appropriate for secondary school-aged children and also integrated elements of accessibility that would be necessary for it to be used in a school setting. PEEP was initially designed to be used as a high-stakes, summative assessment that adapted to the student's skills as they engaged with the task. PEEP was modular, where students would be asked to build out multiple ecosystems across varying environments. Each module would vary in its levels of difficulty and complexity. Complexity and difficulty would be scaffolded based on the students' performance in the previous module.

Piloting PEEP

We piloted the PEEP task in late 2019 with students from 8th to 10th grade at various school districts across the United States. Two studies were conducted to better understand students' and teachers' perceptions of the task, underlining scoring distributions. Information gathered from these studies was used to iterate and further improve the PEEP assessment task.

Over 80% of students who engaged in the task expressed positive sentiment towards it and felt it was relevant to their school work.

First, we conducted think-aloud studies where students would play through the task and, as they did, they would be prompted to describe what they were doing, why they were doing it, and their experiences with the game interface. Results revealed that students found the task enjoyable, engaging, and relevant to what they were learning in school. Interestingly, the younger students expressed more interest and engagement in the task compared to the older students, however both groups had overall positive sentiment. Teachers found the task engaging and a fun supplement to add to their curriculum. However, teachers did express concerns about the GBA's alignment to Next Generation Science Standards. They also had reservations about the scoring, interpretations, and reporting functions of the task

After think-aloud testing, we also conducted a small scale pilot where students went through the PEEP task at their own pace. This was done in classrooms and without researchers or teachers asking the students to explain what they are doing or why. This simulates a test taking environment for the student to give us more accurate data. Similar to the think aloud findings, results from this study revealed that over 80% of students who engaged in the task expressed positive sentiment towards it and felt it was relevant to their school work. Initial results showed that the underlying scoring for PEEP was working and showing variance in score distribution across students.

While these results were promising, PEEP was never implemented in schools beyond this initial work. In 2020, Imbellus was acquired by Roblox and the team transitioned toward working on the Roblox platform to develop hiring assessments as well as contribute to the educational community that is growing at Roblox.

Examples of GBA: Roblox

The "Roblox Problem-solving Assessment" (PSA) is a GBA designed to evaluate the problem-solving competencies of applicants for a variety of technical positions at Roblox, a US-based digital gaming platform technology company where the authors work. Our hiring assessments are developed and tested specifically for Roblox and the needs of our workforce. The assessment development and testing process are guided by rigorous scientific frameworks and best practices from the fields of Learning Science, Psychometrics, and Data Science. The use of an automated, standardized assessment provides an equitable opportunity for all candidates, regardless of background, to demonstrate job-relevant skills.

Roblox chose to develop GBA for hiring selection for both of the reasons cited above: measuring different constructs and measuring familiar constructs differently. First, the Roblox PSA is designed to ensure that each candidate is given the opportunity to demonstrate critical skills and abilities that are important to their prospective role at Roblox. These include hard-to-measure competencies like systems thinking and creative problem-solving, which are amenable to GBA. Second, even for some target constructs that have non-GBA, off-the-shelf assessments available (e.g., aspects of personality and computer coding ability), Roblox wanted an engaging assessment that showcases its own technology as part of the hiring process—hence GBA.

Construct Identification

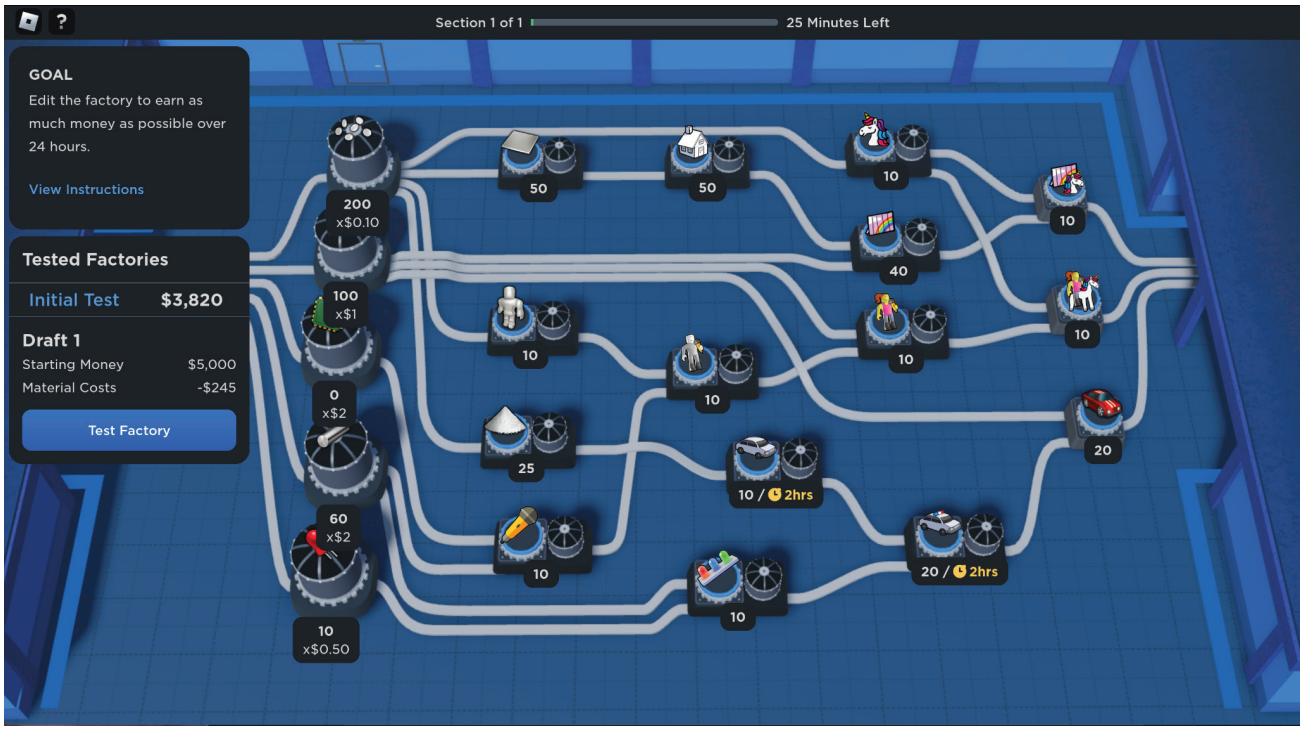
The first step in developing the Roblox PSA occurred in 2021, when we identified the constructs necessary for success in the roles of interest. To accomplish this our psychologists conducted a broadly-scoped job analysis, including over 100 interviews with Engineer and Product hiring managers and leaders and collected data and artifacts on their job duties. During these interviews, respondents identified KSAs that are targeted during the selection process, important for success at Roblox, and that distinguish experts from novices across various roles. The major themes across the interview responses were summarized for both junior and senior roles across the Engineering and Product functions.

Complex skills such as creative problem solving and systems thinking are necessary for success in the target roles, and high levels of ability in these areas indicates potential to make a long-term positive impact at the company.

The identified KSAs were then ranked as most viable for a game-based medium using a literature review and whether or not the KSA is already being measured as a part of the hiring process. There were four categories of KSAs or competencies identified: cognitive, intrapersonal, interpersonal, and practical. Based on a literature review, market research, and the signals already being collected during the interview process, we decided to develop two game-based tasks that focus on key cognitive skills and abilities of applicants, which we built using the Roblox game engine and platform over two years using the ECD approach described above.

When evaluating candidates for roles at Roblox, we are interested not only in strong technical ability, but also in the application of those skills and abilities during complex cognitive processes. Our job analysis demonstrated that complex skills such as creative problem solving and systems thinking are necessary for success in the target roles, and high levels of ability in these areas indicates potential to make a long-term positive impact at the company. There are currently two tasks that are in-use operationally: "Robots" and "Factories." The Robots task is designed to measure creative problem solving, specifically ideation and divergent thinking. The Factories task is designed to measure systems thinking skills. Both creative problem solving and systems thinking were identified as critical skills for success based on an extensive job analysis done between 2020–2022.

Figure 3. *Factories Task within Roblox.*



The Roblox GBA uses the same development and scoring techniques mentioned above. Within the Roblox GBAs performance is measured based on patterns of behaviors that applicants exhibit within the task while they engage in the problem solving process. Generally, there are no "right" or "wrong" answers like one would see on a traditional test. Instead, we look to quantify how they get to a solution and the steps they took to get there. These item scores are not based on machine learning or black box techniques, using ECD (Mislevy et al., 2013), we outline the items during task development so we know what actions a player can take and how we will develop a scoring code around various patterns.

Validating the Roblox Problem-Solving Assessment

Similar to our work with McKinsey, we have continued to develop a program of research leading to a multi-faceted validity argument supporting the use of our GBA tasks for hiring at Roblox. The framework of this research has been largely the same, ranging from measuring the face validity of the tasks through applicant survey to comparing scores concurrently with external non-GBA measures of the same or similar constructs (creative problem solving, systems thinking), to prediction of candidate quality and performance (correlation with expert-scored resumes, prediction of performance at later stages of the hiring process, prediction of performance on-the-job).

Reducing Anxiety: Roblox Practice Test

There is a large literature in assessment extolling the virtues of practice tests as a key part of assessment (e.g., Adesope et al., 2017 for a review in the education context). Allowing test-takers to engage with test content and format can reduce anxiety and improve measurement validity. At Roblox, a key component of the use of GBA has been to also provide an opportunity for applicants to familiarize themselves with the Roblox PSA environment, especially the UI/UX aspects of a GBA, which might be unfamiliar to some candidates.

In 2023, we launched "Kaiju Cats," our practice GBA task that encourages candidates to familiarize themselves with game-play elements used in the hiring assessments in a pressure-free environment. The goal of this tool was to provide candidates with an easy (and stress-free) way to get familiar with the test format and reduce test anxiety for those who may not feel comfortable with game-based elements. Initial pilot results revealed that Kaiju Cats lowered test anxiety among applicants (through pre/post measurements), particularly for those applicants who did not have prior Roblox experience. The practice test is live on the Roblox platform and open to the public and we advertise it heavily in recruiting events as well as all applicant communications. As of late 2024, over 300,000 users have engaged with the task on the Roblox Platform.[1]

Figure 4. *Screenshot of Kaiju Cats available publicly on Roblox.*

Roblox Community Fund - Education

In 2021, Roblox created a Community Fund to provide grants to pairs of developers and educational organizations to develop new, educationally focused experiences on the Roblox platform. Many of the grant recipients were educational partners who already work with thousands of educators and millions of students across formal and informal educational settings.

Our team at Roblox supported this work by developing artifacts, tools, and acting as consultants on many of these projects. Many of the developers working in this space have limited experience with building educational games and simulations and even less experience with GBA. Our team was asked to step in and help fill the gap by building out ECD documents, leading workshops, and meeting on a regular basis to talk through measurement strategies and data collection techniques. At the end of 2024, we have contributed to 5 separate experiences that are currently live on Roblox that are accessible by students, parents, and teachers.

One of these experiences is Mission: Mars, a free educational experience available on Roblox and developed in collaboration with the Boston Museum of Science and Filament Games.[2] In Mission: Mars, students are astronauts on Mars and have to engage in a variety of problem solving tasks while they explore the planet. Our role in supporting this work included meeting with members of both the design team and Museum of Science content team to talk about stealth assessments, proactive evidence design, and potential scoring strategies within the task.

1 Anyone with a free Roblox account can try Kaiju Cats at https://www.roblox.com/games/13977123257/Kaiju-Cats.
2 Similarly, anyone can try Mission: Mars at https://www.roblox.com/games/10840095864/Mission-Mars.

Beyond Selection: Workforce Learning & Development at Roblox

Recently, we have begun to develop an in-house game-based conversational simulation tool as a general engine for workplace learning and development (L&D), again using the Roblox platform as the foundation. Our first use of this L&D simulation is as a way to provide our managers with training and practice delivering feedback to employees as part of a simulated employee performance review conversation—a key area of improvement identified by our internal employee listening program. The new tool provides an interactive environment to help managers practice this skill (particularly giving difficult feedback) and transfer what they learn into their actual performance conversations with employees. Similar to the assessment development and testing process, the L&D development has been guided by rigorous scientific frameworks and best practices from the fields of Learning Science, Psychometrics, and Data Science.

This game-based L&D tool specifically focuses on four areas of development for Roblox managers and leaders: how to structure a performance conversation; how to build conversations around feedback that is the most specific and relevant to the current "situation, behavior, and impact" (Bommelje, 2012); how to work with their employees to construct goals, and how to maintain supportiveness and openness throughout even difficult conversations. The primary mechanism is a series of simulated conversations with both immediate feedback to the learner (typically a Roblox people leader) after dialogue choices and end-of-conversation feedback telling the learner what they are doing right and how they could improve.

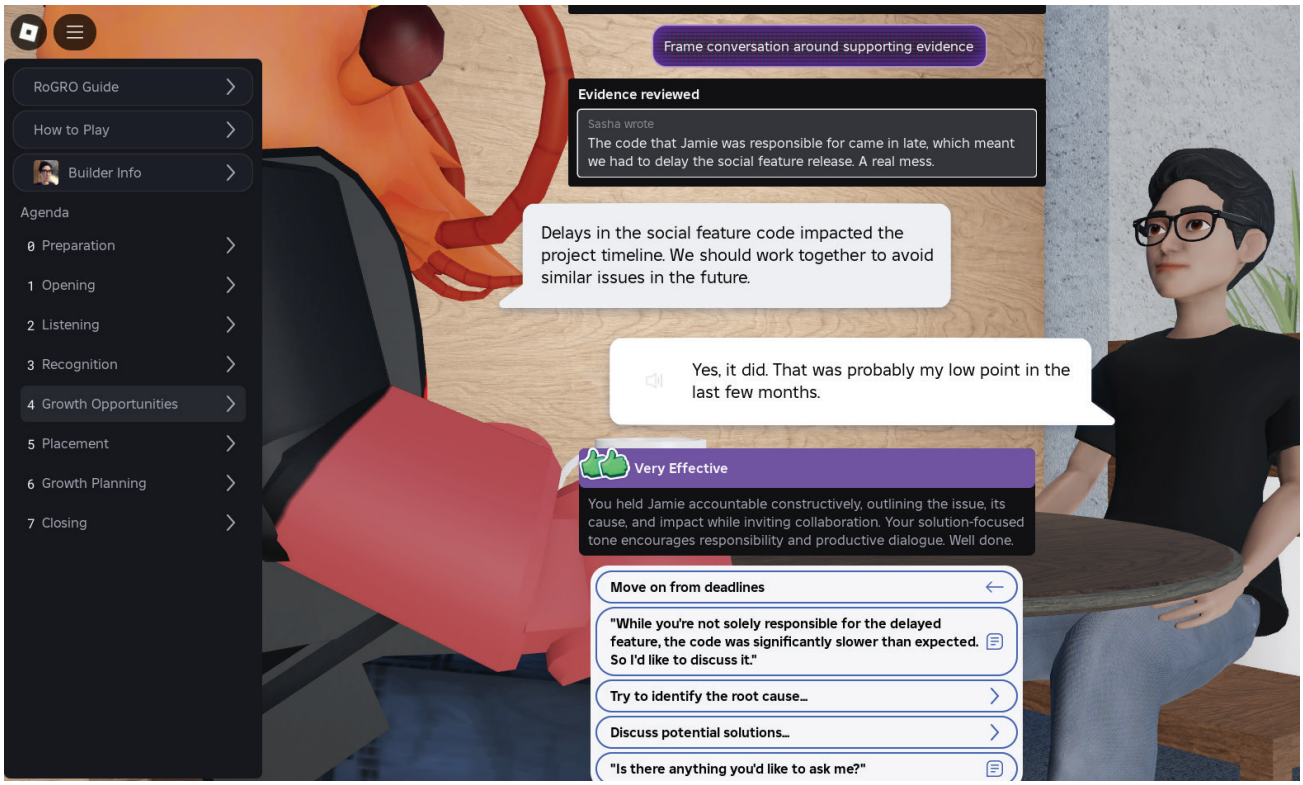
The tool is built on the Roblox platform and is designed to be easily accessible to current employees. Upon entering the task, the employee is presented with a conversational scenario, usually around giving feedback to their direct report. Employees are walked through a tutorial which outlines how to interact with the various UI elements they see during the task. The employee must complete the tutorial and then begin to prepare for the conversation with their colleague or direct report. They will use examples, peer feedback, and other evidence to support the conversation.

Once the employee enters the tool, they see their Roblox Avatar seated at a table across from a simulated direct report. Employees then begin a conversation with their direct report by selecting prompted dialogue options. Each option elicits a response from their direct report as well as real-time feedback from the tool on the effectiveness of their choice. Feedback provides the employee with areas to improve on as well as reinforces the positive behaviors they demonstrated.

The new tool provides an interactive environment to help managers practice this skill and transfer what they learn into their actual performance conversations with employees.

As the employee progresses through the conversation, they are reminded to use evidence to provide performance feedback to their direct reports on their accomplishments and growth areas. The choices that the employee makes while engaged in the tool are recorded and scored based on their alignment to specific learning goals. After the employee exits the tool they are provided a summative report of their time in the experience and specific areas of improvements they can focus on that are tied to their performance on the learning goals.

Figure 5. *Screenshot of Roblox L&D Game.*



We are currently integrating this game-based L&D simulation into our existing manager development courses, which cover a range of topics including providing manager feedback. The L&D simulation is designed to be used alongside the manager feedback course to reinforce what is being taught in the lectures and workshops through hands-on game-based practice. Managers will attend this course and then have access to the simulation where they can work through various scenarios and try out the techniques they had just learned. Practice-based learning has been shown to increase the probability of mastery compared to workshops alone (Ogrinc, 2003). This provides our managers with real-time classroom support as well as a way to practice at their own pace with the ability to return as often as they desire.

Our new prototype went live in early 2025 and we are currently conducting a series of validity studies to make sure that the game-based simulation is engaging, relevant for our managers, and ultimately improves the quality and frequency of the feedback employees receive. A full study design will be implemented in 2025 that will collect employee and manager perceptions of the tool and of the quality and frequency of the conversations they are having, performance data within the tool, and overall product usage (i.e., how do managers use the tool).

If this work shows promise we will be expanding this simulation beyond feedback conversations into other areas such as structuring effective 1:1 conversations, preparing and delivering presentations, and "managing upwards." We are also exploring the use of an integrated LLM to allow for more fluidity in the conversation as well as adaptations over time.

Concluding Thoughts on the Future of GBA

Almost a decade of designing and developing GBA in both the education and workforce environments has taught us that the approach can be very challenging, but also very rewarding. Looking ahead, we believe that the use of GBA will continue to expand and become a familiar component of many testing programs as long as the field can continue to drive development costs down and improve the underlying technology.

Controlling Costs

Compared to more traditional assessment, GBA is still very expensive on a cost-per-item (or unit of information) basis. There are several reasons for this cost differential. First, as we outline above, developing GBA requires an interdisciplinary team with a broad range of skills (game design, software engineering, cognitive scientists, assessment experts, psychometricians, etc.). Some of these disciplines, like engineering, are highly in-demand in the labor market. Second, testing and development cycles are long and early stages require frequent iteration (and, often, expensive pilot and field test data collection). Third, fully-immersive game experiences are difficult to make accessible for test-takers with disabilities, requiring either new technology or the development of equivalent means of assessment. Finally, there may be hardware and bandwidth requirements for some GBA that require investment in infrastructure.

The use of game-based assessments will continue to expand and become a familiar component of many testing programs as long as the field can continue to drive development costs down and improve the underlying technology.

The good news is that there are a variety of innovations and strategies that can be combined to reduce GBA development costs and ensure the method is more feasible for broad adoption. First, the explosion of generative artificial intelligence in recent years, while not useful for everything its proponents claim, does appear to be very useful at producing medium-quality code, reducing engineering costs and accelerating development. As this capability continues to improve, the costs of game engineering and UI component development will continue to decrease.

Without question, the explosion of generative AI promises to increase the efficiency of GBA development and may fundamentally change the work of many of the disciplines required. However, we have yet to see the ability of current-generation tools to completely eliminate any job function entirely. One very interesting area to watch is the application of generative AI to 3D and "4D" (animated) art, an essential part of game-based assessment development. Roblox recently introduced an open-source foundational generative model, "Cube 3D," which generates 3D models and environments directly from text and, in the future, image inputs. The generated objects are fully compatible with game engines today and can be extended to make objects functional for use in GBA.[3]

Beyond generative AI, there are several additional ways to lower GBA costs even further. First, developers can build extensions to existing game development platforms to support educational and workforce assessment and release them free to the broader assessment community. Our experience harnessing the Roblox platform for GBA is a proof-of-concept experiment that demonstrates the feasibility of adapting existing technology for assessment. Second, as those existing platforms and game engines improve their ability to run on low-end hardware and slow networks (something on all technology companies' roadmaps given the need to expand their customer base globally), the cost to implement GBA in educational settings will decrease. Finally, we believe that GBA developers working on the frontier of this area can help others by sharing or licensing relevant artificial intelligence and machine learning methods and novel psychometrics methods and code libraries.

3 Code is available at https://github.com/Roblox/cube and you can try an interactive demo at https://huggingface.co/spaces/Roblox/cube3d-interactive.

Increased Formalization

We believe there is enormous potential for the development of a more rigorous science of game-based assessment, building on the century-plus of academic and industry work that has created the foundation of modern psychometrics and measurement. Particularly promising is the emerging "General Game Playing" subfield of AI research that has led to development of multiple Game Description Languages including: S-GDL (Genesereth et al., 2005), RBG (Kowalski et al., 2017), and Ludii (Soemers et al., 2022), among others.

This is analogous to the idea of *design patterns* in architecture (Alexander 1966) or software development (Beck & Cunningham 1987), with similar potential for improving the efficiency of GBA development. This improved mathematical formalization of game elements ("ludemes") could improve scoring design and cut development and testing time. For example, equating "forms" of GBA tasks is currently complicated and data-intensive; improved formalization might get us closer to equating with little or no data (Mislevy et al., 1993). Further work in this area may also make it possible to generate games rapidly for prototyping and assessment use simply by describing a limited set of variables.

Concluding Thoughts

Almost a decade of designing and developing GBA in both the education and workforce environments has taught us that the approach can be very challenging, but also very rewarding. The combination of increased interest in measuring cross-cutting or complex cognitive constructs and durable skills in education and the workforce, coupled with the desire to make assessment more engaging, suggest a growing demand for game-based assessment, despite the relatively high start-up costs and need for an interdisciplinary development team. Looking ahead, we believe that the use of GBA will continue to expand and become a familiar component of many testing programs as long as the field can continue to drive development costs down and improve the underlying technology.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701.

Alexander, C. (1966). The pattern of streets. *Journal of the American Institute of Planners, 32*(5), 273-278.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Baker, E. L., Everson, H. T., Tucker, E. M., Gordon, E. W. (2025). Principles for Assessment Design and Use in the Service of Learning. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *Handbook for Assessment in the Service of Learning, Volume I: Foundations for Assessment in the Service of Learning.* University of Massachusetts Amherst.

Beck, K., & Cunningham, W. (1987). Using pattern languages for object-oriented programs. Paper presented at OOPSLA 1987 Conference. No. CR-87-43.

Bommelje, R. (2012). The listening circle: Using the SBI model to enhance peer feedback. *International Journal of Listening*, *26*(2), 67-70.

Deterding, S., D. Dixon, R. Khaled & 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.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled approaches to assessment design, development, and implementation: Cognition in score interpretation and use. In *A. A. Rupp & J. P. Leighton (Eds.), The Handbook on cognition and assessment: Frameworks, methodologies, & applications* (pp. 41-74). Malden, MA: Wiley.

Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2010). Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology*, *19*, 105–114.

Genesereth, M., Love, N., & Pell, B. (2005). General Game Playing: Overview of the AAAI Competition. *AI Magazine. 26,* 62–72. https://doi.org/10.1609/aimag.v26i2.1813

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, 54, 170–179.

Hellerstedt, A., & Mozelius, P. (2019). Game-based learning—a long history.Proceedings of the Irish Conference on Game-based learning, Cork, Ireland. https://www.researchgate.net/profile/Peter-Mozelius-2/publication/336460471_Game-based_learning_-_a_long_history/links/5da1c24745851553ff8ad248/Game-based-learning-a-long-history.pdf

Kowalski, J., Sutowicz, J., & Szykuła, M. (2017). Regular boardgames. arXiv. https://arxiv.org/abs/1706.02462

Lindley, C. A. (2002). The gameplay gestalt, narrative, and interactive storytelling. *Proceedings of Computer Games and Digital Cultures Conference,* June 6–8, Tampere, Finland.

Liu, M., & Haertel, G. (2011). Design patterns: A tool to support assessment task authoring. *Large-Scale Assessment Technical Report*, *11*.

Sweet, S. J., & Rupp, A. A. (2012). Using the ECD framework to support evidentiary reasoning in the context of a simulation study for detecting learner differences in epistemic games. *Journal of Educational Data Mining*, *4*(1), 183–223.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, *2003*(1), i-29.

Mislevy, R. J., Sheehan, K. M. & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30,* 55–78. https://doi.org/10.1111/j.1745-3984.1993.tb00422.x

Narayanasamy, V., Wong, K., Fung, C., & Rai, S. (2006). Distinguishing games and simulation games from simulators. *Computers in Entertainment (CIE), 4*(1), 9. https://doi.org/10.1145/1129006.1129021

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States.* The National Academies Press. https://doi.org/10.17226/18290

References

Oranje, A., Mislevy, B., Bauer, M., & Jackson, G. T. (2019). Summative Game-based Assessment. In D. Ifenthaler & Y. Kim (Eds.), *Game-based assessment revisited.* Springer.

Ogrinc, G., Headrick, L. A., Mutha, S., Coleman, M. T., O'Donnell, J., & Miles, P. V. (2003). A framework for teaching medical students and residents about practice-based learning and improvement, synthesized from a literature review. *Academic Medicine*, *78*(7), 748–756.

Pellegrino, J. W., Baxter, G., & Glaser, R. (1999). Addressing the two disciplines problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (pp. 309–355). AERA

Plass, J. L., B. D. Homer, & C. K. Kinzer. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283.

Porter, A. (2007). Rethinking the achievement gap. *@PennGSE: A Review of Research*. https://www.gse.upenn.edu/system/files/u10/Fall_2007.pdf.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals.* Cambridge, MA: MIT Press.

Seelow, D. (2019). The art of assessment: Using game-based assessments to disrupt, innovate, reform, and transform testing. *Journal of Applied Testing Technology, 20*(S1), 1–16.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, *55*(2), 503–524.

Simons, A., Wohlgenannt, I., Weinmann, M., & Fleischer, S. (2021). Good gamers, good managers? A proof-of-concept study with *Sid Meier's Civilization. Review of Managerial Science, 15,* 957–990. https://doi.org/10.1007/s11846-020-00389-8

Smith, P. S., & C. L. Plumley, with L. Craven, L. Harper, & L. Sachs. (2022). *K–12 Science Education in the United States: A Landscape Study for Improving the Field.* Written by P. Sean Smith and Courtney L. Plumley. Carnegie Corporation of New York.

Soemers, D., Piette, É., Stephenson, M., & Browne, C. (2022). *The Ludii Game Description Language is Universal.* https://doi.org/10.48550/arXiv.2205.00451

Stecher, B. M., & Hamilton, L. S. (2014). *Measuring hard-to-measure student competencies: A research and development plan.* RAND Corporation. https://www.rand.org/pubs/research_reports/RR863.html.

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

Wing, R. L. (1967). *The production and evaluation of three computer-based economics games for the sixth grade: Final report* (Report No. ED014227). Westchester County Board of Cooperative Educational Services. https://eric.ed.gov/?id=ED014227

About the authors

Dr. Sean P. "Jack" Buckley is Vice President of People at Roblox, where he oversees several teams including People (HR) and People Science and Analytics. He was previously President and Chief Scientist at Imbellus, Senior Vice President at the American Institutes for Research (AIR), and Senior Vice President of Research at The College Board. He also served as Commissioner of the U.S. Department of Education's National Center for Education Statistics (NCES) and as an Associate Professor at New York University, and an Assistant Professor at Boston College. He began his career as a surface warfare officer and nuclear reactor engineer in the U.S. Navy and has also worked in intelligence analysis. He holds an M.A. and Ph.D. in Political Science from Stony Brook University and an A.B. in Government from Harvard University.

Dr. Erica Snow is the Senior Director of People Science and Analytics and Early Career Recruiting at Roblox. Previously, she was Director of Learning and Data Science at Imbellus, a game-based assessment startup acquired by Roblox. She also worked at SRI international as the Lead Learning Analytics Scientist before joining Imbellus. Dr. Snow has over a decade of experience evaluating the implementation and impact of a variety of educational technologies(i.e., ITSs, MOOCs, LMS, and blended learning courses) within K–12, postsecondary education, and workforce training. Her work has been presented both domestically and internationally to both scientific and non-scientific colleagues and has been published in over 70 peer-reviewed publications. She holds a Ph.D. and MA in Cognitive Science from Arizona State University and a BA in Psychology from Ball State University.

About the Study Group

The Study Group exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy; uncover future design needs and opportunities for educational systems; and generate recommendations to better meet the needs of students, families, and educators.

Date of Publication
March 2026

Citation
Buckley, J., & Snow, E. (2025). Game-based assessment: Practical lessons from the field. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning*. University of Massachusetts Amherst Libraries.