8 Measuring Resonant Success

Eric Klopfer, Jason Haas, Scot Osterweil, Louisa Rosenheck

Published on: Feb 19, 2019

As both creators and researchers of educational games, we are often asked, "Do games work?" This is, of course, an unanswerable question. Similarly, we can't answer (and haven't tried to answer) the question, "Do books work?" It depends on the definition of "work." What are the desired outcomes and how are those being measured? And do those measurements matter? We must also consider which books, which students, which learning goals, which teachers, and which context we are asking the question about. There are too many variables to give any single answer to this question.

There is a rush to draw parallels between education research and medical research. The rationale is that medicine has defined a set of standards and protocols for determining whether medicines "work" or not. The randomized controlled trials of medicine are seen as a gold standard to many and can tell us whether something "works." There are many problems with this "health care envy" (Squire, 2011, p. 232). And those in the medical field know all too well the problems with this gold standard. It is hugely expensive and rife with contested, biased, and manipulated results (Ioannidis, 2005), as well as many cases in which the results are highly dependent on the context of the study (Open Science Collaboration, 2015). This isn't to say that randomized controlled trials are always a waste of resources, but that this methodology should be looked on with great scrutiny.

Instead we should embrace a diverse set of research methods, data sources, and analytical techniques that match the intervention, context, and outcomes. Our process involves asking what would make this particular game *resonate*? As we describe in the previous chapters, each of the games we have profiled embodies different principles, was deployed in different contexts, and had different desired outcomes. So how would we know if that game is resonating with the audience in the chosen context?

The first step is to *align* the measurements and data collection with the game goals. This may seem obvious, but often data are used because they are convenient rather than well aligned. School exams, surveys, and in-game data have their place, and are often convenient, but they don't always provide the best alignment with game goals. For example, *Vanished*was designed to help increase student interest in science and proficiency with using scientific data in argumentation. Even if we could have used school exams, they wouldn't have been aligned with these goals. We needed to use more appropriate data to measure these outcomes.

Games scholar James Paul Gee has often said, "Nobody gives a Halo player a test afterwards to find out what they learned. The evidence of their learning comes from their success in the game itself." In other words, being great at *Halo* is evidence of *Halo* performance. But many times, we as teachers and researchers are interested not only in how the students did on a particular task, but also in how well they transfer some of that skill and understanding to another context. We might want to know how well they transfer their in-game genetics skills to solving genetics problems on paper, or how well they transfer their skills in forming hypotheses to laboratory projects.

Such transfer is in many ways the holy grail of teaching and learning, as well as a hugely challenging problem (Bransford & Schwartz, 1999). It turns out in general, people are not good at applying something that they learn in one context (e.g., a mathematical concept like ratios and proportions set in a context of pizzas) and applying that in an entirely different context (e.g., the same mathematical concept of ratios and proportions applied in a context of weights of objects). We need to be taught how to bring the ideas from that one context into the new one.

Gee (2003) has also argued that games are really good at promoting such transfer. They do this by constantly asking players to transfer their learning from one context within the game to another context within the game. For example, players might need to stack boxes at one point in the game to climb out a window. Later in the game, they may need to stack barrels to reach a seemingly unreachable object in that room, or to build a bridge of boxes to cross a river. The nature of getting better at the game is learning these game-specific skills and applying them again and again, often in novel ways. Thus, it may seem that doing well in *Halo* is actually doing well in a series of increasingly complex and diverse versions of a task, setting up players to at least transfer those skills to another, similar kind of game. Given how bad people seem to be at transfer, this kind of scaffolding and learning to transfer is a significant accomplishment for the medium.

The question is what do we want our players to transfer their understanding to? What could we measure and use as evidence that someone has learned something that may have meaning beyond the game itself?

It is most useful to think of these measures and outcomes in parallel with the design of the game itself. While some measures can be put into place post hoc, there are two big reasons to think about these measurements earlier in the design process. First, the game activity might change to help collect the necessary data. For example, in *Radix*, since we wanted data used as evidence, we needed to make sure there were opportunities within the game to use data as evidence. Second, the data that the game collects might be changed. So perhaps the desired activity (e.g., using data as evidence) is already a part of the game, but those data need to be kept in a useful and easy-to-parse way to be analyzed later. This is one of the big challenges of learning analytics—collecting the useful data and just the useful data.

In general we define five principles (the five A's) for data collection and measurement to provide evidence that a particular game is achieving the stated goals (figure 8.1). These principles are artifacts, affect, audience, assessments, and analytics.

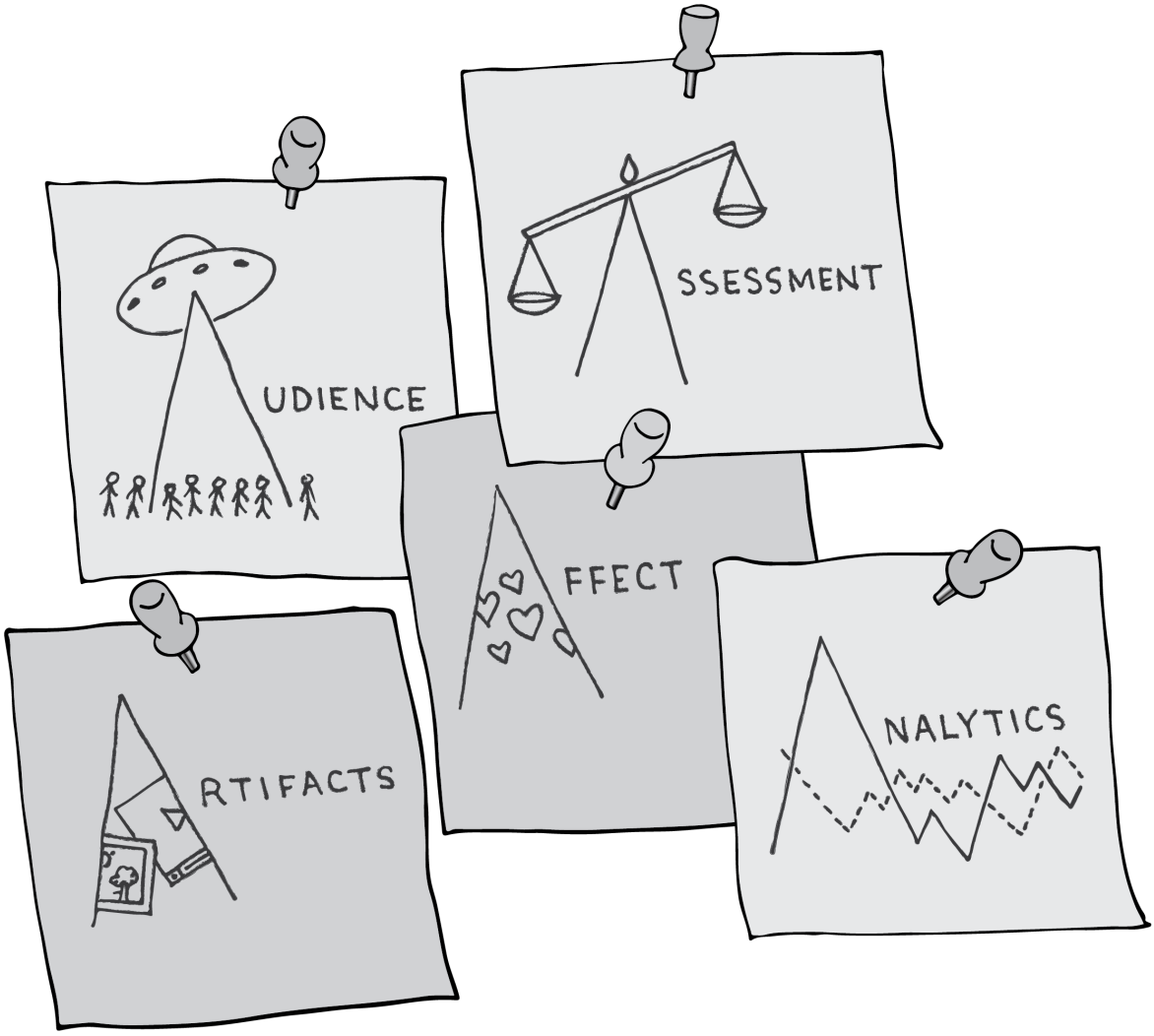

Figure 8.1 Five principles for data collection and measurement.

- **Artifacts**—The products that players produce within a game or a game context that can be captured. These are often complex products, like written dialogue, or multimedia (photos, video).
- **Affect**—The feelings players and their teachers/mentors have about the game and the game experience. Do the players like the game and come back to it?
- **Audience**—Who the game is reaching. Does the game reach its intended audience in terms of age, diversity, skill, and numbers?
- **Assessments**—Instruments provided to players (often outside the game) that measure player knowledge and understanding. These can range from multiple-choice tests to concept maps or choice-based tasks, often provided in a pre- and post-style (before and after the game).
- **Analytics**—In-game data that can be collected about player performance and usage. These data can be mined and analyzed to get insights into player patterns, learning, and engagement.

All five of these principles then need to be considered in *context*. Who is using the game? How are they using the game? Where are they using the game? Each of these variables is important for understanding the data provided by the five A's.

Let's take a deeper look at each of these, where they have been employed in our games, and where they find the most utility.

Artifacts

Artifacts are products that players produce in conjunction with game play. These can sometimes be produced in the game itself. In other cases, they are produced in activity around the game. Artifacts help to make thinking visible—to enable researchers and teachers to see what and how students are thinking about their activities.

Vanished involved a series of complex tasks performed by the players. These tasks often involved the actions of dozens, hundreds, or thousands of other players. Creating evidence from these complex tasks similarly required complex multidimensional data. Much of the action of the game took place in the forums. This forum activity provides ideal data about student activity.

Early in *Vanished*, the players were given a task to collect temperatures across the United States. This task led to a conversation about why they needed to collect the temperatures. Many of the students simply posted the temperature at their location on that day, sometimes without even noting anything about the location. Eventually the discussion (figure 8.2) turned to why they were collecting the temperatures. Most of the students seemed to respond with various forms of "because we have been asked to." One student (150103) provided some explanation of why they might be asked to do this particular task. "A lot can be determined from weather patterns. Did you know the fall of Rome occurred in a time of drought? We know this from the rings of trees. They are thicker when there is a lot of rain, and thinner when there is not much rain." The conversation after that, however, fell back to something more along the lines of "because they asked us."

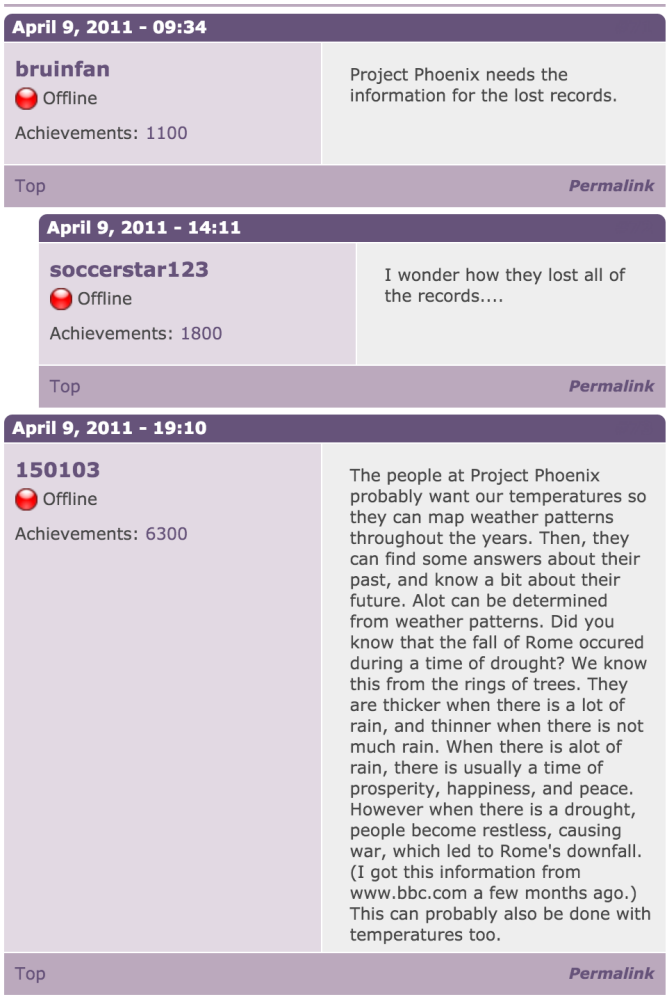

Figure 8.2 Discussion of temperature patterns in the Vanished forums.

Later in the *Vanished* adventure, the students discussed the role of a potential reversal in the magnetic fields of the Earth. This conversation showed a much greater depth of evidence citation (figure 8.3).

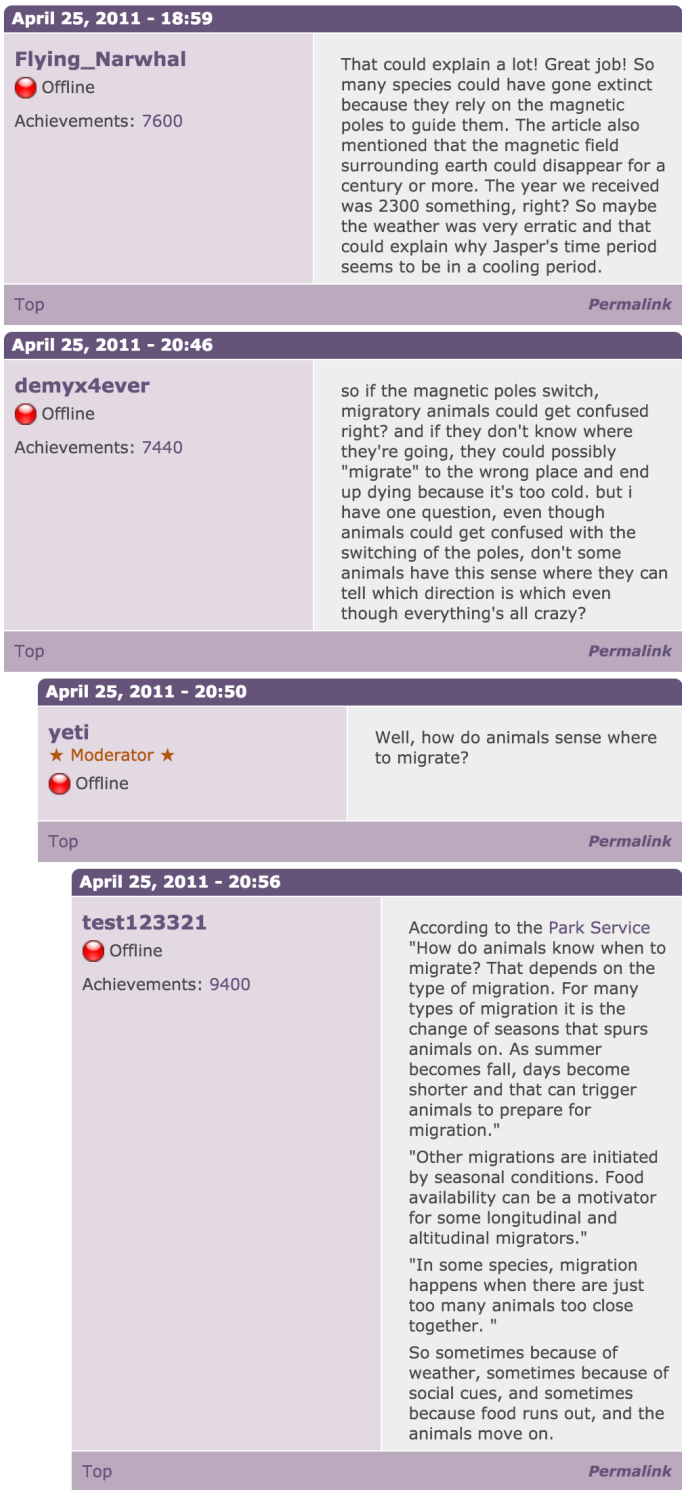

Figure 8.3 Discussion of the possible change in the magnetic poles in Vanished, using citations and evidence, both important in the discussion.

Another task involved collecting pictures and Latin names of local flora and fauna (figure 8.4). This task could not be assessed individually. It could only be looked at across the entire group. Was there

representative data from across the country? Did posting of some species spark the posting of other species? How did the discussion in the forums support the progression of species posting?

Figure 8.4 Some of the information about fauna collected by the players of Vanished.

Analyzing this kind of conversation is challenging, but clearly in this case, only these kinds of complex data could give us insights into the nature of argumentation from the students. These in-game

folk like blacksmiths, barrel makers, and slaves. Their task was either to provoke the revolution or stop it from happening. The goal of the game was to promote students' thinking from the perspective of the average people at the time, who were such an important part of the revolution.

As part of the research, players in many classes were asked to make screencasts (video screen captures from the game that they narrated) to document their experiences. This provided not only documentation of what they had done in the game, but rationales explaining why they had done these things. It was ultimately a way to capture a lot of data about player activity and thinking. It would have been extremely challenging to collect these kinds of data solely from in-game data (even if we had had the analytics that we have now). These video diaries captured a much more multidimensional set of data. They had the additional advantage of providing utility to teachers. Educators often struggle with how to "grade" the game play, but it is a more familiar experience to grade student projects, such as this. Most importantly, it also provided a richer experience for the student players themselves. They developed deeper insights into their characters and connected with the narrative in ways that they had not as they simply played the game. This represents the ideal for data collection—the collection itself not only doesn't detract from the experience, but actually enhances it.

Affect

Many conversations about the nature of games turn to the concept of "fun," though most scholars don't include fun in their definitions of games. Instead they focus on the structures of games, such as rules, systems, and outcomes (e.g., Salen and Zimmerman, 2004; Juul, 2003). But the result of these structures—the rules, systems, and outcomes—is intended to engage the player. For many game designers, including ourselves, "fun" may be the wrong word because of its associations. "Fun" conjures up notions of wild enthusiasm and continuous heightened engagement. While some games may be designed to elicit that response, many others are designed to elicit other kinds of engagement. The structures of the games challenge players so that they are deeply engaged.

We often recall Seymour Papert's (1998) previously noted concept of "hard fun." Papert was working with kids using his Logo programming language. He found that the students he was working with often wrestled with challenging problems. In the moment, the students would likely describe that feeling not as being "fun" but rather as being "challenged" or even "frustrated." Eventually they would overcome that challenge, either through some insight, help from a friend or teacher, or even sometimes trial and error. That feeling of solving the challenging problem was quite satisfying. While Papert called this "hard fun," a good argument could be made for thinking of this simply as "fun."

A similar feeling is obtained when playing many games. Games that are too easy are not engaging. Players who pick up games that provide no challenge often put those games down very quickly.

Instead, games that continuously provide players with challenges that are just a little too difficult elicit the most engagement. These games may be thought of as keeping players in their zone of proximal development (Vygotsky, 1978), the space between the things that they can't do at all, and the things that they can do easily by themselves. This is the space where learners can do things with help, and it is the space where the most effective learning happens. Good games keep players in that space, where they are engaged and learning.

The concept of "flow" (Csikszentmihalyi, 1975, p. 36), "the holistic sensation that people feel when they act with total involvement," is also associated with some types of games, particularly long-form games. It is a state reached when the player's skill and the game difficulty are well matched. If you've played a good long-form game, you certainly know the state. You are immersed in a task within the game, trying to solve a problem, and before you know it, hours have passed. This is related to and sometimes confused with a notion of immersion, though immersion (Jennett et al., 2008) may be induced by many design elements of a game beyond challenge, including narrative and aesthetics. Regardless, there seems to be a connection between these emotional states and learning that is useful to capture (see Hamari et al., 2016, for a discussion of this research).

While "hard fun" and "flow" can be distinguished from the naive notion of "fun" in a frivolous sense, they are emotional states that have the potential to be productive and keep a player engaged in the game. These are the states that we try to create in *resonant games*. We could even say that one way the resonance of games is apparent is in how it puts players in this state of being continually challenged, learning and overcoming those challenges. So one way we want to know if the game is succeeding is through evidence of this kind of engagement.

There are many new ways of trying to measure engagement, including galvanic skin response (sensors that are attached to the skin that measure changes in electrical conductivity that can be correlated with changes in mental state), eye trackers/sensors (cameras that can measure where someone is looking or the dilation of their pupils, or other kinds of cameras and sensors that correlate expressions with emotions). While these may have a place, we don't use any of them, for several reasons, including that they aren't particularly reliable at this time (at least for measuring something like flow, though some researchers have made progress on this, e.g., Nacke & Lindley, 2010). They are also expensive and challenging to scale. Perhaps most importantly, they primarily measure proximate measurements of engagement. What we really want to know is what the player thinks and feels.

The most common way of measuring engagement is through surveys. These are highly scalable but often not particularly reliable either. You ask players a series of questions about what they thought about the game, how they felt, whether they'd be likely to play again, and so on. These are often measured with a Likert scale (e.g., a scale of 1 to 5, with 1 being "would be highly unlikely to play the game again" to 5 being "highly likely to play the game again"). We often use these as one measure

because they are easy to obtain and they can be helpful, particularly when comparing games or parts of games that we have created and used with a particular audience. The problem is that the responses are usually between 3 and 4. While we might see statistically significant differences between groups, games, and contexts, these differences are often small in magnitude, and one has to question whether a difference of 0.1 on this scale is practically significant. Still these kinds of data can be a useful first pass on engagement when comparing similar products. For example, in UbiqBio, where we deployed a series of games in a similar format to the same group of students, we might be able to learn something about student preferences comparing their ratings across the games.

But there are other ways of measuring affect. Many researchers are now looking for direct biometric measurements of affect. As mentioned above, for some time galvanic skin response has been one such measure. As someone becomes excited, their conductivity changes, and sensors on the skin can measure such response. More recently researchers are using pupil dilation as detected by webcams to measure similar excitement. We have not employed any of these methods for the reasons we give above as well as another important reason: they are all intrusive. Such intrusion is not only off-putting for many players, but it may actually change their perception of the experience itself, which defeats the purpose of these measures in the first place.

In many ways, players' immediate response to the game is less important than their memory of the game anyway. We care about what they think they felt about the game—what they might tell their friends, or what might incite them to go back to the game again. As such, we look at measures like how much time they played the game (either measured through logs that the players keep or more recently through analytics), or how much they played the game beyond what they were required to do. We can look at how long they played each session for, or whether they did more tasks in the game than required. All of these give us insights into what matters, whether the game is appealing enough to play on its own. We can also look at the ultimate learning outcomes that we hope emotional engagement will bring

Direct observations are another method that we can employ, although this is limited for games that are played primarily outside school. For those played in school, we can directly observe the student experience. We can record conversations between students as they talk about the game, or note where conversation seems to stray far from the task at hand. We can watch for students who play until minutes after the bell has rung, or those who turn off the computer in advance and anticipate the bell. These provide powerful indicators of whether students do indeed like the game.

Interviews can also work well. We can interview students to find out not only whether they enjoyed the game, but what in particular they enjoyed or felt like could use improvement. This not only helps us determine whether the game is good or bad, but more importantly helps us iterate on our designs and form design principles that we can carry forward to future games. These are some of the most

important insights that we can receive. This feedback that allows us to change designs or synthesize principles across projects are the ones that can truly bring about a difference. In addition to interviewing students about their opinions of the game, we also interview them about the content in the game—what they are doing, what they think it's about, why they are choosing to pay attention to or ignore particular elements. Through this we can learn a lot about how they are problem solving in the game, whether they're enthusiastic about it or see it as just another assignment, and what elements they associate with those opinions.

The parallels in the commercial gaming world suggest that our methods are on the right track. What players look at in determining whether they will like a game are the reviews—both professional and from their peers. They look at the number of stars or ratings on a scale of one to ten or to one hundred that the game has received from sites they trust. They read the reviews that describe what players like or dislike. This combination of simple Likert scale results and qualitative data best describe the likability of a game.

Audience

The most obvious aspect of audience that we might be concerned about is its total size. And in many cases, this does matter. It matters that we can say that 33 million pets have been saved by players in *Lure of the Labyrinth*. Or that we have had nearly twenty-five thousand players in *Radix*. They show that we have been able to get to scale. But focusing too much on scale can be detrimental. It may be more important to reach a particular audience than a broad one.

Scale doesn't always matter. Sometimes dozens or hundreds of participants is entirely adequate to get the intended research results. This was the case with UbiqBio. We weren't doing research that required thousands of users, and the games themselves were made primarily for research (though we do try to maintain them for teachers). If the intended outcome is research, not product, then those numbers will suffice. What matters in those cases is which particular students were involved. If the claim that we want to make is that our game works well for low-achieving students, or students who haven't already identified as being interested in a topic, then we need to make sure that the audience that we have selected actually reflects those aspects. So we always need to get data about who the audience is, so that generalizations about results can be made in that context.

Often we are interested in differences between particular audiences. Is a game more appealing to males than females? Are teachers in low-resource schools as likely to pick up the game as those in high-resource schools? Are people who identify themselves as gamers as likely to stick with the game as those who do not identify themselves in that way? Collecting and comparing this information can be critical. We are always interested in these differences. Sometimes we may be targeting a particular

audience and we want to make sure that we got it. *Bite Club* and *Farm Blitz* targeted young adults. If the game wound up attracting a substantially older audience (even if it was large in scale), it might be considered unsuccessful. Other games might target particular kinds of learners, those with a particular science misconception, or perhaps students with a learning challenge. *Vanished* was targeting middle school students outside school. If the audience primarily came from in school, then we would not only interpret results differently, but interpret the success differently as well given that it was designed to be a model that could attract students to science outside the classroom.

But scale can matter for several reasons. Sometimes scale matters because the funder wants to make sure it is funding not a niche research project, but something that ultimately reaches a mass audience. A large scale means that their investment is affecting more people, which makes a lot of sense. Other times scale matters because the kinds of research questions that we are asking demand a large audience. Some of the analytics research that we discuss below, while not always strictly "big data," is often big enough where the patterns can only be detected with enough data. Finally, some projects have a combined goal of research and product development. In the case of *Radix*, for example, one goal was to create a sustainable product, one that we could keep alive beyond the end of research funding. While we are still working on that goal, showing an audience that is large enough to potentially sustain a product is a critical first step. In many of these cases, audience alone is a necessary but not sufficient metric. We would also need to view it in the context of how those players were engaging with the game.

Scale isn't a simple function of the quality of the game. It is also a function of the nature of the game. (Is it about a narrow academic topic, or something broadly applicable? Think about the UbiqBio games that target primarily academic topics versus *Farm Blitz* and *Bite Club*, which are about financial literacy). Marketing, however, may be the largest contributor to scale. Unfortunately, marketing is something that isn't budgeted into most research projects. Many funders feel that marketing is not what research projects should be doing. But as more projects seek to create products or work in the realm of big data, we have a serious need to make sure marketing plays a role in the project.

Assessments

Assessments, whether they be customized or standardized, often play a significant role in determining the success of any of our academically oriented games. Usually people are concerned with what students can do beyond the game, not just what they can do inside the game. We (and many others) have argued that the game itself may be a better assessment of what students can do than a traditional written assessment. After all, the game can dynamically adapt to student performance, present them with multistep complex tasks, and set tasks in a more authentic context. Games that are designed using a form of evidence-centered design (Mislevy et al., 2003) try to do just that. In this approach, the designers need to start with a model of what they want the student to know (student model), work

toward what evidence they could actually see showing that a student knows those things (evidence model), design a task that could elicit that evidence (task model), and put together a system (a three-part process) that would enable the designers to implement these models. Balanced design (Groff et al., 2015) is a simplification of this idea that was created to make the ideas even more accessible to game developers. This is simplified to a content model, which includes what designers want students to know, the evidence model of how students would demonstrate knowledge of that content, and the task model of what the students would engage in. Evidence-centered design is for the design of all assessments, not just games, and illustrates good practices for assessment in general. As more game designers follow these processes, we may come to the day when well-designed games serve as the assessment themselves, but this is not the case today.

The story of *Radix* demonstrates this tension that still exists between data from the game and traditional assessment. When *Radix* was proposed, it did not include external assessments. We wanted to diagnose student understanding through their ability to complete the in-game tasks. Later, after the project began, we were asked to provide validation via external assessments. But what did this mean? Their performance on the state exam? On a test we made up?

The challenge with standardized assessments in cases like this is that they are very coarse-grained measures. They test a wide range of topics, most of them fairly superficially through short-answer and multiple-choice questions. In a domain like science, they also still primarily test factual knowledge, not more complex conceptual understanding, which is what we target in our games. So even if we do a lot better at teaching the things that we want students to learn, we won't see significant changes on the current battery of standardized tests (although there is hope that they are getting better, as tests designed around conceptual knowledge start to emerge).

The alternative is a custom test. While we could certainly write a test with relevant questions, it may not be particularly meaningful. Designing tests that are valid—where items are shown to be related to constructs, student understanding by item is correlated to other measures of understanding, the test is unbiased, and so forth—is difficult and expensive. A good assessment of the concepts we were targeting with *Radix* could have cost nearly the same as the game itself.

Instead we went with a hybrid approach (an approach we've used in other projects, including UbiqBio). We compiled relevant items from already validated standardized tests, but selected items that were only on the topics that we were targeting in the games. This provided enough specificity to allow us to try to measure the constructs of interests, without incurring the costs and complexity of a completely new assessment. It also provided an important validation for the teachers who were using the game. As it turned out, in most of the topic areas, students did improve by playing the game as measured by these assessments. But this is not true in all the topics, which once again demonstrates that it isn't about whether a game works or not but where, when, and why it works that matters.

Ultimately the assessments that we use must connect back to the policies and practices of the context in which they are being used. Right now in the United States, assessments that are economically efficient and reliable are still the norm. The resulting instruments are fairly static and knowledge oriented. Shifting to more dynamic and process-oriented assessments, including game-based assessments, requires an investment in research and infrastructure. There have been problems of approaching this issue at scale, because of the fragmented nature of state-based high-stakes assessments. The demand for assessments that align with new standards, frameworks, and most importantly desired twenty-first-century industry competencies, does apply some pressure that could result in real change. Game-based assessments stand to both benefit from and contribute to such a change.

Analytics

Big data is revolutionizing many fields, from medicine to social science to computation. In a recent report on big data and learning Dede (2015, p. 2) defines big data through the four Vs:

> The four "Vs" often used to describe what makes data big are (1) the size of data (volume); (2) the rate at which data is produced and analyzed (velocity); (3) its range of sources, formats, and representations (variety); and (4) widely differing qualities of data sources, with significant differences in the coverage, accuracy, and timeliness of data (veracity).

As more learning activities take place online, the significance of big data, and learning analytics (the analysis of big data as applied to learning) is rapidly increasing.

Games have long been looked to as a place for interesting big data. Game companies themselves have used this kind of information to improve game play. With reference to Gee's earlier comment about "no one testing Halo," in fact, Microsoft has done just that. As they launched *Halo 3* nearly a decade ago, they used a data-intensive approach to improving the game (Thompson, 2007). Through thousands of hours of tests, they were able to find where players struggled, where they were confused, where players had an unfair advantage, or where they simply got lost. They did this through observing players, but also recording their game play and analyzing visual and data summaries of where players were going and what they were doing.

In one example, they were able to look at a heat map (a map showing hotter colors where players spent more time and cooler colors where they spent less time) of game play to analyze where players were getting stuck. They were then able to change the terrain and the characters to alleviate this problem.

Now, analytics are an even more critical part of the game industry. In the age of "free to play" games (games that are free but you pay for upgrades, new levels, time savings, abilities, etc.), game companies want to know what tweaks to their games they can make to get more players to play, stick around, and spend money on those upgrades. So they look at volumes of data on what players are doing and what causes them to stick around and spend money. Sometimes this is done by looking at natural variation in patterns, but that approach is often combined with A/B testing, in which players may be randomly assigned two different conditions, and researchers look at the responses from the two conditions. For example, some players get a randomly assigned one-day wait for an upgrade to their virtual world, while other players get a two-day wait. Researchers can then examine whether the one-day or two-day wait led to more players paying for an upgrade.

With learning games, we are most interested in which patterns are associated with improved learning. Val Shute (Shute & Ventura, 2013) has done extensive work on a version of learning analytics in games she calls "Stealth Assessment." In one case she used a game called *Physics Playground* (originally *Newton's Playground*) to present learners with mechanics problems that they needed to solve by drawing various objects on the screen. For example, they might need to get a ball to a higher plane by drawing an inclined plane as well as a larger ball that might be used to knock it up the plane. All the students' actions in the game were logged, so when they drew a tool, used the tool, erased the tool, and so on, those actions were recorded. Other aspects such as time spent on task, time between moves, and so forth were also logged. Shute was able to show that the data taken from these logs were as predictive of players' physics knowledge as the formal tests that they took outside the game. This same technique has been applied by Shute and others in a variety of learning games, from 3D virtual worlds to role-playing games.

These analytics can be most useful when they are designed into the game from the start and focus on key insights that seem likely to be yielded from such data. As in many other fields, developers often decide to collect keystrokes or mouse clicks from their games rather than some higher-level action like completing the building of an item, or making a choice to go down a particular path. In these cases, you can still mine these data for interesting patterns through a process that our colleague Justin Reich (2013) calls "fishing in the exhaust." You can just look through and see if there are interesting patterns to be seen.

But fishing in the exhaust is unlikely to answer any interesting design questions. There are just too many different patterns to be seen, and when you find them, you need to determine which of those patterns are interesting, that is, meaningful and generalizable, not just spurious correlations (Vigen, 2015). If the meaning is built in from the start, through a process like evidence-centered design or balanced design, then these correlations are much more likely to be meaningful. On the other hand,

we should not entirely disregard the utility in understanding learning through such data streams. For certain kinds of outcomes, fitting data is available and can bring timely insights (Reich, 2014).

In our projects, we have used analytics in several ways. In the case of UbiqBio, we were able to look at play patterns and see how they correlated with success on external measures. Some of the very basics included how much time students played the various games.

The game requires a login and records how much time students spend actively involved in each game. Being "actively" involved is an important distinguishing factor, since this discounts time when the screen was just left on a page. This particular data set (figure 8.5) shows that the average (mean) amount of time spent playing each game varies quite a lot by game, with *Beetle Breeders* being played four times as often as *Chomp!* was played.

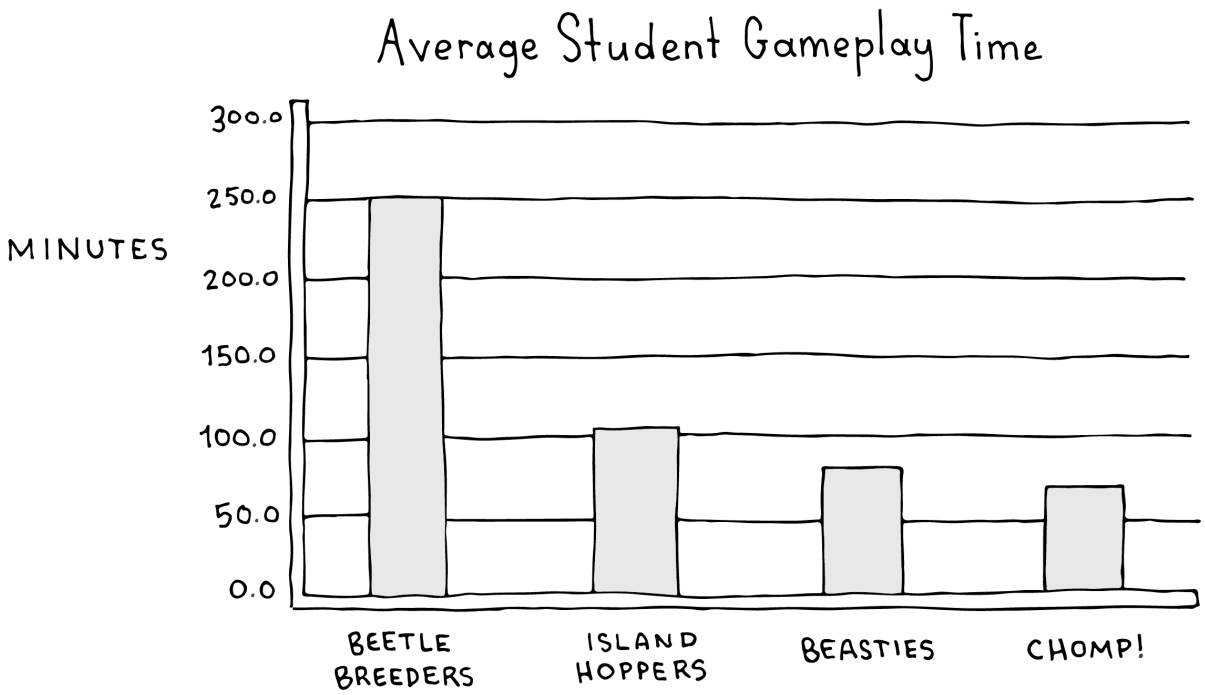

Figure 8.5 A graph showing the average student play time in some of the UbiqBio games (adapted from Perry & Klopfer, 2014).

Combining some of the analytics (data on both time spent and levels completed) with assessments (how well the players did on external assessments) provides interesting insights into the differences between the games and the relationship between design and outcomes (figure 8.6). In one game (*Beasties*), time spent playing was a predictor of performance, whereas in other games, time and level together (the first being positive and the second negative) were predictors of score (this essentially says students who spent less time completing the same number of levels did better). These kinds of

data are trivial to collect from the games and provide insight into both student performance and game design.

Coefficient Estimates for Play Time and Level Achieved on Test Score per Game

	Beetle Breeders	Beasties	Island Hoppers	Chomp!
Time (10K)	−.039**	.0647**	−.223**	x
Level	.245**	x	.0138**	x

Note: All models shown are statistically significant ($p < .05$). x = no statistically significant influence of that coefficient for that model. ** = significant difference at $p < .05$.

Figure 8.6 A table showing the relationship between performance gains, time played in the game, and the level achieved (adapted from Perry & Klopfer, 2014).

In *Radix*, analytics were built in as an essential part of the research and feedback to both teachers and students. We recorded data that included general things like which quests were completed and in what order, where players went, and which items they collected. But we also collected information pertinent to particular quests, like which tools were used doing those quests, what the students "turned in" for their quests, and specifics of how they used the tools. This is a fair amount of data, but it is designed to be recorded in useful chunks. These are the actions that we felt were meaningful within the game. Clicking on a flower may not be meaningful, but turning in a flower for a quest had some meaning intended by the student.

For example, in one of the genetics questlines, players were required to collect flowers and breed them to produce particular kinds of offspring. To complete the quest, players needed to turn in the

flowers that they had bred to exhibit the specified properties, as well as provide evidence of how they did this using a Punnett square. To understand not only whether they turned in the correct objects at the end, but also what process they used and how they revised their attempts, we can track these types of data:

- The genotype (genetic characteristics) and phenotype (visible traits) of the flowers they bred;
- Every attempt at creating the Punnett square and what they chose to put in it;
- Every use of tools to examine the genotype or phenotype of flowers and what they used it on;
- What they turned in when they tried to complete the quest, including the flowers and Punnett square.

In some of the questlines, specific detectors were built to identify patterns that we preidentified based on research. Often students have particular models in their heads of how systems work. By offering choices to express both the correct and the incorrect models of those systems, we can use analytics to identify the students who hold the incorrect models, and then offer those students alternate quests or relevant feedback based on their current state of understanding.

For example, in another questline around geometry concepts, players are required to draw scale maps. When players turn in their maps, we can determine specific mistakes like

- the correct scale but wrong map elements
- the wrong scale but correct map elements
- the correct scale but reversed in order

These could then help diagnose whether students were struggling with drawing geometric shapes (the first mistake), calculating scale (the second mistake), or converting units (the third mistake).

This combination of analytics makes it easy to show teachers where their students are succeeding and failing (figures 8.7 and 8.8).

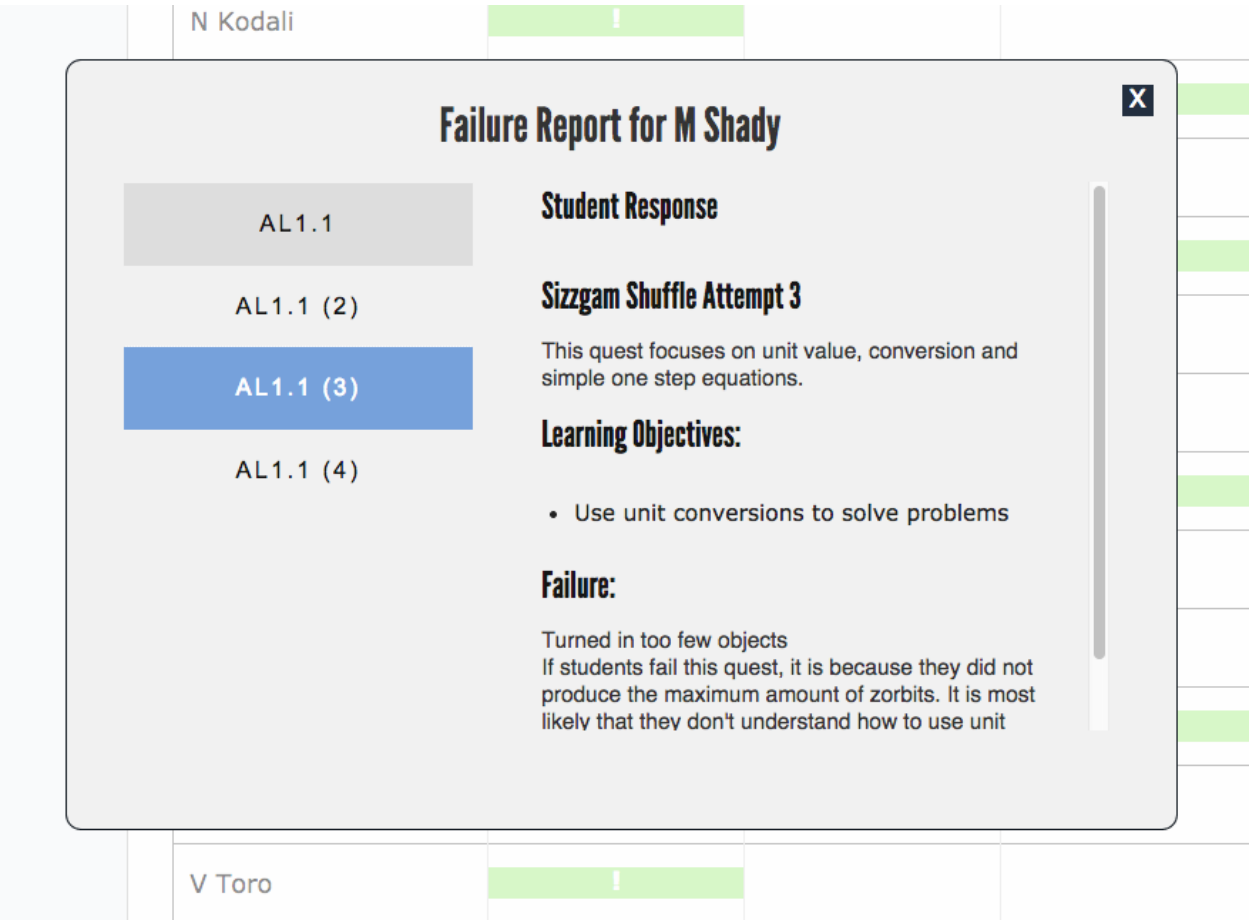

Figure 8.7 A failure report indicating what a student has done based on analytics.

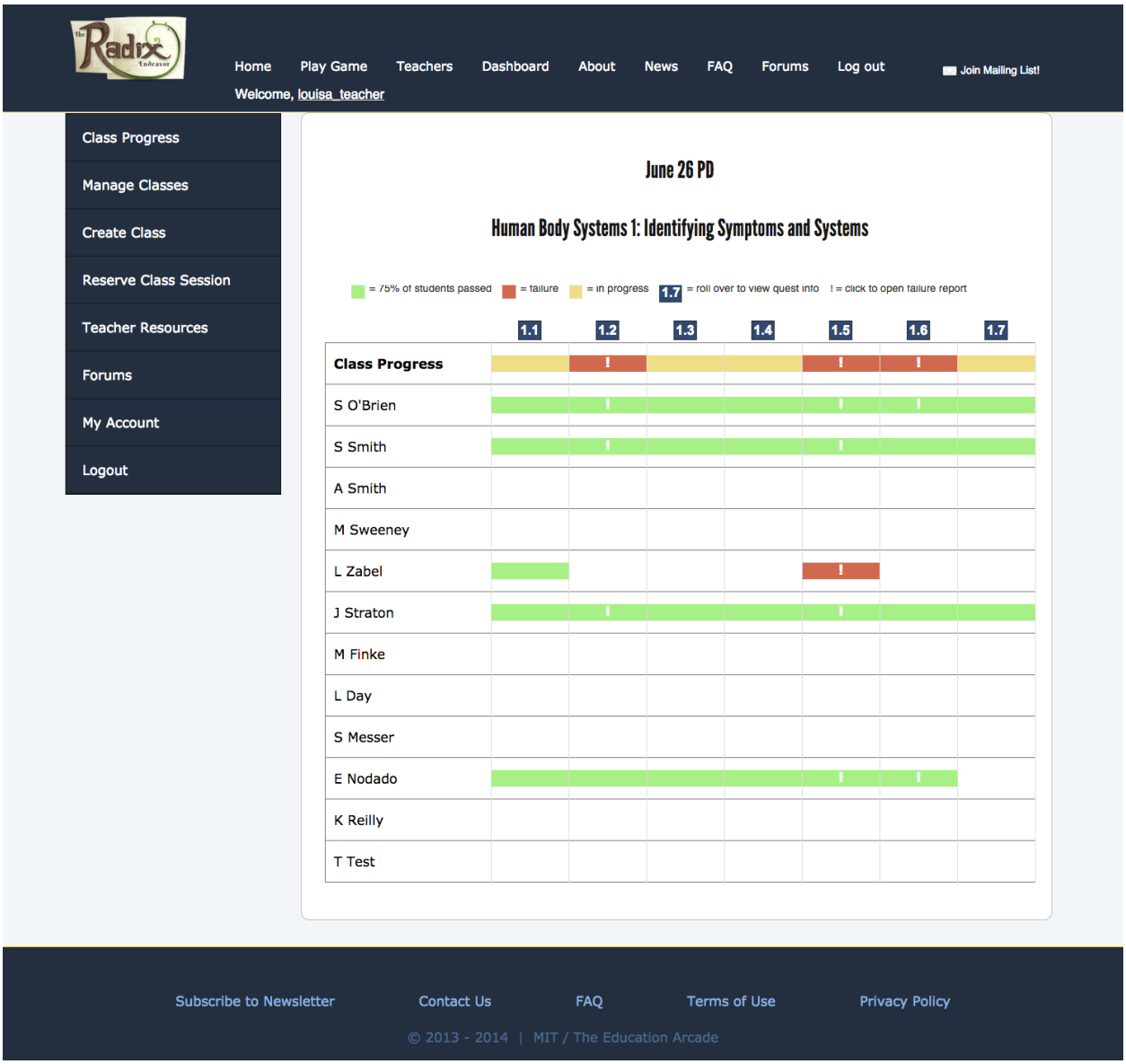

Figure 8.8 Dashboard showing student progress based on analytics.

But these analytics also provide the opportunity to gain additional insights into behaviors and performance. For example, we can identify patterns of behaviors within the game that are more likely to lead to success or failure of a quest. Or we might simply be able to see whether the game is successful at providing some of the choice that we would like to have in games. We don't want to create a game with one fixed pathway to a solution, but instead want one that has multiple pathways. Analysis by one of our colleagues, Montzy Cheng, did just that. She looked at the different tools that students chose for getting to a solution. Looking at one of the genetics quests and the tools that students used, she noted one highly successful pathway (figure 8.9) in which students used the examiner tool at least once and then bred the creatures. But there was another pathway in which the students didn't use the examiner, instead using the decoder and using it more times. This also led to

success. These kinds of analytics can help us determine whether our designs are achieving the intended goals.

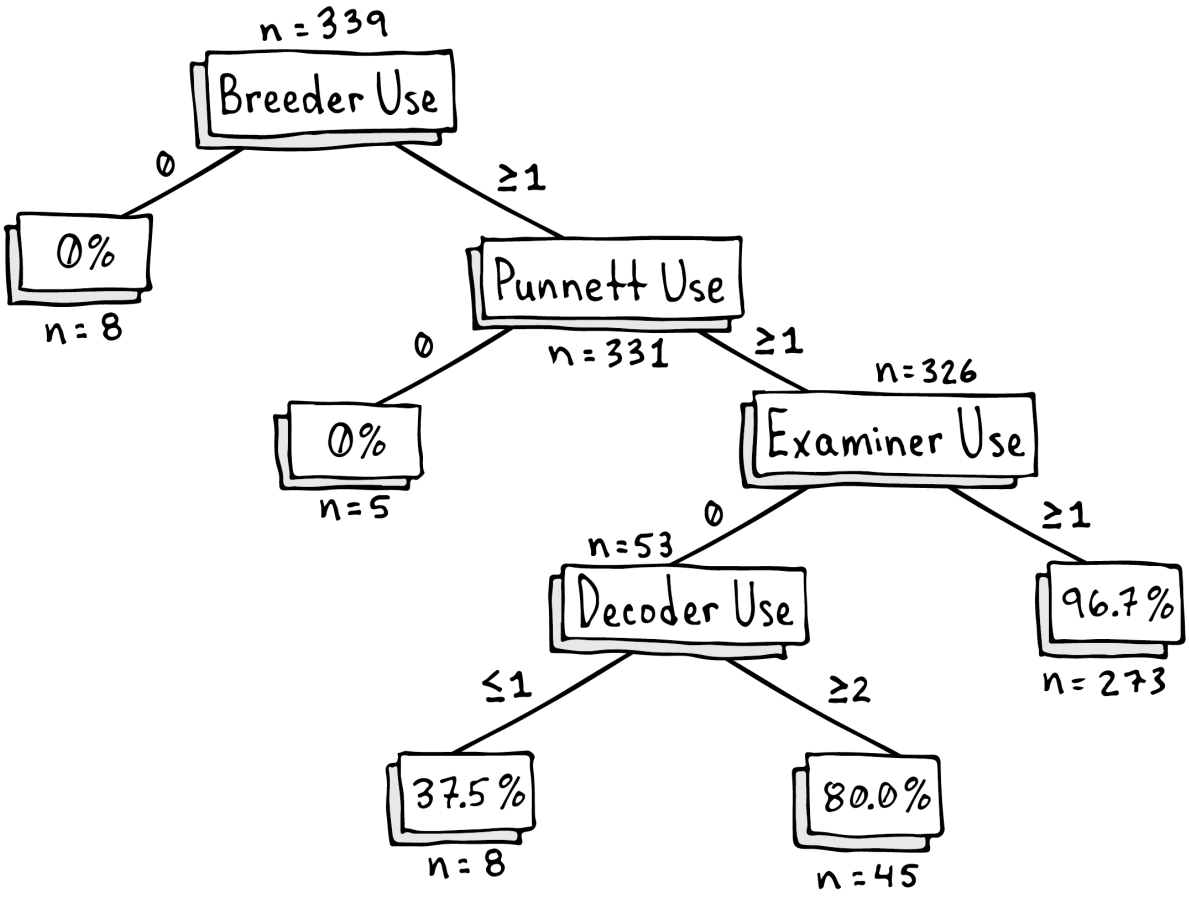

Figure 8.9 A tree showing different pathways of tools used to solve a quest (adapted from Cheng et al., 2017).

Context

All these metrics—artifacts, audience, affect, assessments, and analytics—are mediated by context. When was the game used? Where was it used? How was it used? We always seek to capture these data so that we can situate the findings within the right context. Often we think of our games as preparing students for future learning (Bransford & Schwartz, 1999), not necessarily teaching them the concept outright. For example, in one study (Arena & Schwartz, 2014) researchers designed a game called *Stats Invaders* to be used before students learned statistics. The idea was that students would develop some intuitions for statistics through playing the game that they could draw on when they learned the concepts formally. That study found that in fact the game did prepare students for this future learning.

Many of our games are built on this idea—that the game should come before the formal lesson, not as reinforcement. A great example of this was in *Celebrity Calamity*, in which our partners at Commonwealth ended up doing a small study. One group received a financial lecture first and then played the game, and another group played the game first and then had a lecture. The Commonwealth team members tested the knowledge of each player in the middle (after only one activity) and again at the end (after both). They found that only the students in the group that played the game first, and then saw the lecture, improved their understanding. And for that group, only after doing both activities did they see improvement.

At a broader scale, the partnerships surrounding the Commonwealth games provide interesting examples of different contexts that contribute to substantially different outcomes. Partnerships that put the games in the right place at the right time (as measured by people who were primed for the message and with the ability to act immediately) had a substantial impact. In other cases, even when the game was played, if the context wasn't suitable for action, then there was no change.

So, when we implement games in the classroom, we track not just whether but how the games are being used. Is it used as homework or classwork? Before the unit or after the unit? Through observation or self-reporting teachers, we can also get information on what teachers did to support concepts in the game. Did they make direct reference to the game in a class lecture? Or provide a follow-up activity that related to the concepts in the game? While some of these variables help us determine how well the game worked, they also provide us with insights into where students and teachers like to use the games and how we might be able to change them to suit the classroom environment better.

Application

Applying these principles and techniques can be challenging. Each research project calls for a unique suite and application of methods and metrics. This holds true whether the project is embarking on research for the purpose of publication or simply to learn more about what is working or not working in a particular implementation. We do research when we are developing products, but we also embark on a less formal research process when we implement any games in classes to learn about how they are affecting students. This process enables us to improve the game design, implementation approach, and even outreach strategies for a game. Regardless of the nature of the research, in considering the methods to choose and how to apply them, it is important to consider several steps.

Plan in advance—There is a great temptation to rush into development and implementation, with the thought that research can be considered later. This temptation is especially strong in the day of big data, when you can collect every keystroke and figure out what it all means later. In practice, this

purely post hoc research approach misses many of the most important insights. Those insights are revealed when you can identify meaningful actions based on theory, practice, or other empirical data.

Don't be seduced by numbers—Another mistake that is easy to make is to value quantitative data over qualitative data. After all, quantitative data can provide "answers." In fact, qualitative data may be more appropriate for some research scenarios, and, at the very least, qualitative and quantitative data can be complementary. Quantitative data can help answer the question of whether something happened or not (e.g., did students learn more using this game?), but they don't tell us why that result was obtained. Qualitative data (e.g., data from interviews, observations, or student responses) can help shed light on why that result was achieved. They can elucidate student thinking, motivation, and experience.

Include diversity—No game works everywhere for everyone every time. But you won't be able to tease apart details of who it works for, and the context in which it works, without variation. Part of that variation can be intentional, as part of designed research, but natural variation is also important. This includes students, teachers, schools, implementation scenarios, and many other implementation variables. While you may think it is a better experiment to control these (and it may be at a late stage), early in a project, this variation can help bring insights into the conditions for success and failure.

Diversify your methods—Diversity of methods also matters. No single method will answer your questions completely. You will need to employ suites of methodologies and consider how they complement each other. Consider the qualitative methods that can answer the whys to your quantitative measures' whethers. Collect analytics data that may indicate actions that produce student artifacts of importance. Together these data can help triangulate findings that no single measure can.

Connect research and design—Most importantly, deeply and iteratively connect research and design. Research is not the same thing as evaluation—to determine whether something worked. Research can and should be formative: it can help shape the trajectory of design and development to lead to a better product. Design-based research (Brown, 1992) should be the backbone of learning games development. This methodology from the learning sciences promotes an iterative approach to design, development, and research, leading to better products and findings that help communicate why elements of the product are or are not working. Connecting research, design, and development makes this methodology more effective.

Connect research and implementation—There are many reasons that a game might succeed or fail in a classroom, only some of which are related to the game itself. The details of implementation are also critical for the success of the game. So research should also investigate the details of implementation. This includes information on what professional development was provided, the

classroom materials that were used, and even the marketing messages to the intended audience. Documenting these factors so that they can be improved is also a key pathway to success.

Resonating Research

Research on learning games is as complex as the design and development of these games. Good research is also equally as important as good design and development for the success of learning games. Research helps inform teachers, administrators, policymakers, and, most critically, the designers and developers themselves. Engaging seriously in this process has allowed us to not only develop products that we are proud of as products, but understand what is unique and interesting about them. It helps us abstract the details of what we are developing so that they are not just instances of games, but instances of ideas. The products may come and go, but the ideas behind them are what we expect to endure. They are what we hope resonates.