CHAPTER 14

Assessment

Assessing Video Games for Learning

David Simkins, *Rochester Institute of Technology, Rochester, New York, U.S., www.davidsimkins.org, dwsigm@rit.edu*

Key Summary Points

1. As we assess games for learning, we should create an inclusive environment that allows for a wide variety of methods.
2. As we publish on assessment, we should be forthright about our foundational assumptions, particularly our epistemologies. We should judge the contribution of assessments taking the stated foundational assumptions as given, but with an eye for improving methods and practices of assessment given those assumptions.
3. Design and development of learning games needs to similarly own their epistemology, design learning goals in keeping with them, and create experiences that conform closely to those epistemologies.

Key Terms

Assessment
Games
Learning games
Serious games
Epistemology
Methods

Introduction

The increased interest in the use of games for instruction in formal and informal learning spaces has led to an increased interest in assessment of learning games (Annetta, 2010; Clark, Tanner-Smith, Killingsworth & Bellamy, 2013). These assessments originate from many different fields, many of which have differing beliefs about what constitutes knowledge. These differences are indications of significant

epistemological diversity within the field. That is, we do not all share the same premises about what it means to know, and therefore we do not all agree on what methods we should use to gain knowledge.

Epistemological diversity is not unique to the study of games and learning. It is a characteristic source of discussion and debate throughout education (Pallas, 2001). It is also present in the media studies and computer science divide within games studies. Having discussion around what it means to know is common in all academic fields, even in hard sciences such as physics where there are commonalities among most scientists about how one might consider something "known" within the field. Diversity of epistemologies is much greater within humanities, social science, and education. In education research, epistemology is a vexing issue because it not only relates to how we study learning, it relates to how we believe people learn in the first place. Epistemology is central to our inquiry on all levels, and differences among epistemological commitments lies at the center of what we believe we should do and expect from learning (Greeno, Collins, & Resnik, 1996). For reasons beyond the scope of this chapter, we cannot solve this problem by simply determining the correct epistemology. It is unlikely, perhaps impossible, which we will ever resolve all questions about what it means to know. If we accept that we will not simply resolve our differences, this chapter is an attempt to allow us to move forward together as a field. Given that we will not resolve epistemological differences, how should we go about sharing our research findings with each other and building useful understanding?

Three core questions stand out in assessment research, and regardless of the epistemological approach used, any study should first consider these three questions. The three include: 1) the game's learning goals, 2) the core mechanics/core gameplay, and 3) the out-of-game context of play. These are, in part, derived from central studies of games and leaning that form the foundation of the growing field (Clark, Tanner-Smith, Killingsworth & Bellamy, 2013; D'Angelo, Rutstein, Harris, Haertel, Bernard, & Borokhovski, 2013; Shute & Ventura, 2013) and from my own experience assessing video games for learning (Simkins & Steinkuehler, 2008; Simkins, Egert, & Decker, 2010; Steinkuehler et al., 2011). For many of us, the goal of sharing studies about assessment of games is to create better learning games. These three questions, collectively, address the core of that inquiry, and finding a way to work together toward finding helpful answers to these questions may help us to find a way to act as a field of study.

Question one: What are the learning goals?

Before making any other determinations about a study, it is helpful to identify the learning goals. For games developed for learning, these may be easy to determine. Hopefully the designers were prioritizing learning goals throughout their design and development process. Not all games, however, which show promise for learning are specifically designed for that purpose. Researchers must often examine the game's design and conduct prior or simultaneous research to identify the potential of a game as a tool for learning. For example, Squire's work on the *Civilization* series of games shows that some games created for entertainment may have excellent potential as learning games (Squire, 2011). In various studies with diverse populations, it is suggested as a useful tool for explaining alternate theories of historical process, explaining historical contingency, and offering a "modding" environment for creating scenarios for *Civilization* that highlight historical processes, environments, and facts.

Occasionally, even a game designed for learning may be effective in more areas than intended. What happens when a game about scientific hypothesis testing is also very good at developing scientific collaboration and communication skills among group members? Sometimes in the process of research we discover learning affordances we had not intended to test, opening opportunities for future inquiry. Researchers studying complex environments in the wild must often use a variety of research tools to identify potential learning affordances in an open-ended way. Ethnographic, grounded theory, and various qualitative and mixed methods inquiries that apply to research "in the wild" can be extremely helpful in identifying potential learning goals that could be used in conjunction with more targeted game assessments.

Question two:
What are the core game mechanics and other important aspects of play?

In any good learning game, the gameplay should be aligned closely with learning goals (South & Snow, 2012). If we are studying unintended learning goals, this may be a little more complex. Still, to assess a game, it helps to deeply understand the game. The amount of time we need to spend understanding a game may vary dramatically depending on the game itself. The simpler the game, the less time one needs to spend with it to understand what is really going on in play. Even simple games can have surprising affordances for learning, so it is helpful to have a deep understanding of the game's play.

SimCity provides one of the earliest examples of a game created for entertainment and studied for its potential learning content (Betz, 1995-6; Squire 2005). In the early versions of *SimCity*, the core mechanic was the placement of blocks representing areas zoned for a particular purpose. As a city planning game, the focus was on developing the infrastructure for a city, building it over time and balancing constraints and resources, such as money, pollution, and population growth. There was no single objective for *SimCity*; rather, it was a sandbox in which to explore the constraints and affordances of the tools provided. As a learning game, it was a good example of learning through exploration. To understand its play, however, we would need to have an understanding of how the constraints and affordances worked together to make a challenging environment for *SimCity* players, challenges that could be overcome, potentially providing a sense of accomplishment even when there were no explicit goals.

We would also benefit by understanding the cheating mechanisms built into the game—one could enter a code to give oneself more money, for example. I place cheat in quotes because though we may see this as an inappropriate way to play the game, the developers did not. The ability to "cheat" was provided to allow players to continue to have fun in the way they wish. Many early *SimCity* players believed cheating was cheating and would never use the ability to give themselves extra money. For others, it was a normal part of the practice, perhaps even a necessity for their play style. Because it is a sandbox game, it is up to the player to identify his or her own goals and create engaging play in the space provided. The ability to cheat or not cheat is a core aspect of its gameplay; it is not just coincidental to it.

In the *SimCity* example we can quickly identify the core mechanics of the game. Understanding the potential practices and cultures of play, however, may require us to play the game more extensively. Even after exploring the play in depth, we may also need to explore the culture around play, observing other's play and discussing play with others to get a wider sense of the varieties of play.

The goal of this second question is to understand the in-game context of the assessment. How we use this in-game context will depend on our methods of assessment. Researchers may also need to understand when and where to constrain the player's options to facilitate the study of learning, based on the target learning goals. For example, researchers may need to disable *SimCity*'s cheat commands if managing resources is central to the learning goals of a particular curriculum that uses the game as a tool for learning.

Question three: What is the out of game context?

The in-game context is not the only context relevant to learning. It is important to take account of the environment in which learning takes place. Often, the environment is a critical component for facilitating learning, though it can also have the effect of distracting players away from the learning goals. Some methodologies may require us to minimize this out of-game effect. Other methodologies focus on understanding the out-of-game effects, requiring us to see the gameplay within a more natural environment. The learning context is the out-of-game environment in which the game is played. In Gee's terminology (2012), everything in-game is the game, with a small "g." The context around play is part of the Game, with a big "G." That is, it is not only the physical environment around the game but also the virtual environment. Some companies that develop learning games are also developing social spaces around the games, sometimes called affinity spaces (Gee, 2012), which encourage the collaborative and social aspects of learning. A positive and constructive learning environment can be crucial to the achievement of learning objectives, and understanding or controlling context is a part of a complete assessment of a learning game.

When we talk about complex games, we really cannot understand the entire learning environment without being deeply steeped in the practices that surround the game. Such is the case with any large, multiplayer game environment. Looking at the learning that takes place in *Whyville* (Kafai & Fields, 2009) or *World of Warcraft* (Steinkuehler, 2007) both require intimate knowledge of the game's social structures and communities, not just what the players are doing on the screen. The games are just two examples of game contexts that are driven by community. They are influenced by designed structures that facilitate constructive community and where player self-organization creates opportunities for mentoring among players. This network of player interaction facilitates a player's access to information about expert play. While this development of networks to promote expert practice is player generated for many games, some games for learning are designed specifically to mirror established practices, such as Shaffer's epistemic games (2006), which leverage professional practices to create contextualized play.

We can see similar affinity spaces grow around games in ways that do not directly tie back to the game at all. In the work of Black (2008), Magnifico (2010), and Jenkins (2012) on fan fiction communities, we can see the development of literacy practices through participation in the writing, sharing, and commenting on fan fiction related to, but not specifically supported by the game. In situations where the game we are assessing contains "unsanctioned" or completely player-generated and operated content sites, it may be necessary to find expert informants to introduce the researchers to the player communities.

The entire context of a game can be described in terms of circles representing different spaces where players interact with games (see Figure 1). Researchers implementing assessment protocols need to be aware of each of them, though it may restrict assessment, or even player access, to a subset of the three. The innermost circle is the immediate physical context of the players playing the game while they are playing. The second is defined by the opportunities for interaction around the game built specifically by people with special authority over play, which could include researchers, instructors, game designers, or publishers. Whether moderated or unmoderated, these spaces are to some extent controlled by and the responsibility of non-players. The third circle is defined by the social interactions around the game by the community of game players—the students or players themselves. The power structure of these spaces is less formal, often with greatest influence by those players with the greatest social influence within the group.

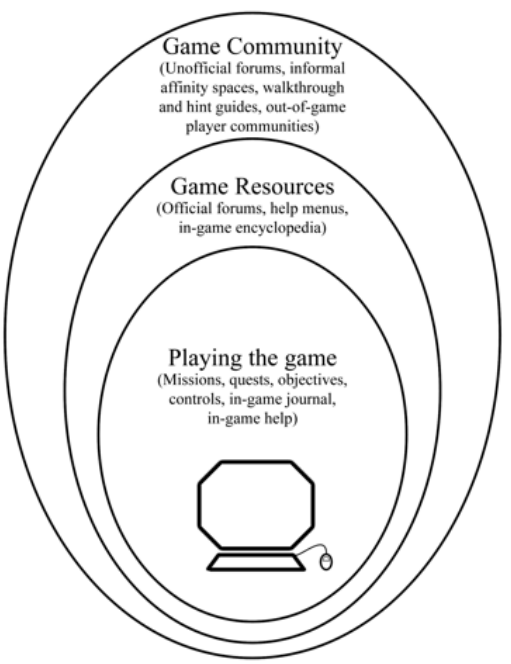

Figure 1: Game context

Identifying the circles is merely a heuristic, and the boundaries between them are not always clear-cut, but understanding that communities of practice around games are formed in several layers can help identify which environments the assessor needs to access to complete a full assessment of play. In identifying the circles, one can begin to determine where your players are playing and in what game-related activities they are participating. After which you can determine what access researchers, instructors, and designers can or should have to complete an assessment.

Case Study One: Dark Gold: Analyzing Big Data Through Quantitative Stealth Assessment

As an outside researcher, it is not always possible to examine a massive data set. Fortunately, Sony Online allowed a select group of researchers to access much of the in-game transaction data for *Everquest 2* (Keegan, Ahmed, Williams, Srivastava & Contractor, 2011). The resulting data set allowed for in-depth queries about actual gameplay for hundreds of thousands of players carried out with minimal, if any, direct effect on the players' behavior. This level of authenticity is in line with the ideals of stealth assessment (Shute & Ventura, 2013).

To understand the study, the researchers needed to know a great deal about the game. For example, researchers needed to know that in *Everquest 2*, it was against the rules of the game to trade in-game items for out-of-game currency or goods. They also needed to know that the rule was commonly broken, and not merely by individuals, but by organized groups of players who would collect in game gold and goods and sell them. The gold and goods were traded online, but no in game money was transferred. Instead, the transaction was completed in game when first the receiving player completed an out-of-game real money transaction.

To complete the study, the researchers used existing research on massively multiplayer online games to identify opportunities within the game that could be compared to out of game examples. In this case, the researchers hypothesized that in-game "criminal organizations" will respond to game policing in ways analogous to how criminal organizations in our society as a whole respond to policing. The in-game "criminals" are gold farmers, who are only criminals by analogy because they do not break laws but they do break the end-user license agreement of the game they are playing. These data were compared to data from criminology research about how criminal organizations respond to the threat of enforcement. These responses to law enforcement included a variety of ways in which criminals learned and adapted to policing so that the impact of the enforcement was minimized.

The study investigated specific in-game behaviors and compared them with the out-of-game behaviors of criminal organizations—methods that included creating decentralized authority, and interacting through multiple tiers of operatives. To compare to out-of-game criminal organizations, the researchers found data with a similar scope: a longitudinal analysis of criminals and co-offenders collected through Project Cavier, a Canadian law enforcement task force.

While we can only assume the criminal's learning goals to be the avoidance of prosecution, such an assumption seems justified. Similarly, in-game gold farmers were assumed to be actively seeking to find ways to continue their efforts despite enforcement of in-game rules against selling in game items or characters for out-of-game money, established by the game's producers.

Even with a tremendous amount of in-game data at their disposal, any action occurring outside of the game was unavailable and needed to be inferred. The in-game markers of gold farming and information

about the variety of interactions that occurred outside of game space were both influenced by qualitative research, such as that carried out by Dibbell (2006) on the Chinese gold farming industry.

The study purported to show that there was an analogy between the reaction of physical world criminals to increased enforcement and the reaction of virtual world rule breakers to enforcement of the rules by the game companies. If true, this opens the door to using in-game enforcement to learn about how to investigate and enforce out-of-game criminal syndicates. It could also, potentially, indicate that in game breach of rules may prepare one for out-of-game participation in criminal syndicates, much as some first person shooters have been allegedly used as training tools for terrorists (SAIC, 2007).

Key Frameworks

The focus of this chapter is on how to make sense of and use of assessments across the field, despite our variety of methods and underlying assumptions within the field. To bridge differences within the field, we need to communicate and be flexible in our understanding of what good studies will look like. A principled study will be conducted within a stated point of view, with clearly indicated assumptions and following through with data collection, analysis, and interpretation of results that are coherent with the assumptions. No study is perfect, and all data is at least somewhat limited or compromised. The process of critique helps our community develop better tools for analyzing and discussing games for learning, but such critique needs to be consistent with the stated assumptions of the study—internal to the epistemologies of the researchers, not external to them.

To understand better how to approach creating community across assumptions, the chapter has followed two divisions within the field. The first is of epistemology. That is, the way we believe people come to know and understand which, implicitly, also indicates what we believe to be the limits of human knowledge. The second is overall methodology. Both I break into heuristics that are a bit rough. Neither of them can completely encapsulate the differences or approaches within our community, but they should provide a general sense of the communities and breadth of the field.

The end of the chapter discusses uses three case studies to introduce different models for assessing learning within gameplay. Each is grounded in an epistemology and therefore uses methods that differ from the others.

Methods

The goal of this chapter is to help us learn from each other. Part of this process is learning how to better study games and learning. This does mean evaluating each other's methods; however, the focus of evaluation should not be to prove one method better than another, but to improve and refine each of the methodologies and approaches to which our community is committed.

Within the last few years there have seen several significant publications about the assessment of learning in video games (Annetta & Bronack, 2011; Ifenthaler, Eseryl, & Ge, 2012). Generally, the methods for assessing learning in games are the same as assessing learning anywhere, and any quality methods text should help establish good methodological practices. Mirroring educational assessment in general, assessment of learning games tends toward statistical analysis, conversation analysis, discourse analysis, pragmatist methodology, critical theoretical approaches, and ethnomethodology. Education research has a general preference toward case study, which lends to meta-analyses as well. There are also relevant ways to use ethnographic methodologies, particularly in work on large social games played outside of formal learning environments. In a dense, but extremely helpful discussion of the field of cognition and learning, Greeno, Collins, & Resnick (1996) break the field into three epistemological views: the behaviorist/empiricist view, the cognitivist/rationalist view, and the situative/pragmatist-sociohistoric view. While grouping approaches can raise as many questions as it answers, the categories they use succeed in showing the major differences within the larger field of learning.

Greeno et al. (1996) do a better job than we can here of fairly describing the field within each general approach. At the risk of oversimplification the three tend to fall into epistemological and ontological camps—the largely post-positivist behaviorist/empiricist approaches that focus on repeatable experiments in controlled environments. The largely post-structuralist cognitivist/rationalist view in which knowledge is understood only within an individual and often non-reproducible context. Thirdly, the language and practice theoretical situative/pragmatist-sociohistoric view, in knowledge is shared and constructed—understood collectively and not individually.

Whichever viewpoint one holds will be crucial in determining which methods are best pursued. Generally, behaviorist/empiricist epistemologies tend one toward quantitative measurement and statistical analysis. The differences within cognitivist/rationalist or situative/pragmatist-sociohistorical viewpoints depend, in part, on the locus of knowledge. When knowledge is maximally localized and sensitive to context, in depth qualitative analysis is likely required. When knowledge can be generalized to any significant degree, qualitative, quantitative, or mixed methods approaches may be useful.

Quantitative methods

Quantitative methodologies are those that focus on measuring. That is, collecting and analyzing data in such a way that it can be measured against other data. This almost always means collecting numerical data and analyzing it statistically.

The general goal of most quantitative assessments will be in isolating and reproducing verifiable results. To do this, it is important to identify learning goals that are themselves measurable. In other words, the researcher needs to be able to identify when the goal has been met. This is often discussed in terms of mastery, with identifiable conditions for mastery. Once mastery of the learning goal can be correctly identified, the researchers need to create a use of the game that can be studied for attainment of mastery. There is a growing interest in the use of games to show the attainment of mastery. In this chapter we are concerned with assessment of games, not games as an assessment tool, but these can

sometimes go hand-in-hand because measuring the success of the game usually requires measuring student success within the game, to show that the player is improving through the use of the game as a learning intervention. To do this, it is important for the researcher to account for the role of play in a learning game. By understanding play, one can identify which indicators or behaviors on the part of the learner evince mastery, which suggest a lack of understanding, and which are distractors that could be misinterpreted by assessing researchers.

Though the intent is to create reproducible, verifiable knowledge, even in quantitative methods the learning context is important. Context affects outcomes, just as interventions affect outcomes. It may be desirable to create as neutral a learning environment as possible, allowing the assessment to be judged for its learning affordances without reference to outside influences. In many cases, however, a sterile learning environment may be counter-productive or misleading.

Table 1. Strengths and challenges in sample quantitative methods

	Strengths	Challenges
In-game instrumentation	Assesses player activity in-game. Time stamped and accurate, if implemented well.	Can be difficult to ascertain correct data to collec. Requires access to game code. Limited explanation of intent.
Pre/post-test	Relatively easy to create and implement. Can be targeted specifically to learning goals.	If not completed immediately before and after can potentially conflate results from other learning opportunities. If completed immediately before and after can suffer from participant exhaustion. Does nothing to help answer why an implementation works or fails only whether it seems to be beneficial. Takes time away from gameplay.
Eye/head tracking	Assesses what the player is attending to in game, not just what they are doing in game.	Equipment can be expensive. Can result in confusing data that shows attention but does not suggest reason for it.

While there are a variety of quantitative methods used in games assessment (see Table 1), the most common assessment method, for almost all quantitative learning and most mixed methods research, is a pre-test and a post-test. That is, using an assessment tool to measure what the student knows about the learning goals before the learning intervention, before playing the game, and what the student knows following the intervention. The gold standard for this is to use a normalized and validated research instrument, which is targeted specifically at the competencies addressed in the assessment's research questions. That is, we are looking to use tests that have been studied for their statistical reliability. That is, the questions (or other analysis tools) of the test(s) used as pre- and post- tests should be constructed so that an arbitrary selection of members of the target population of the study have the same chance

of answering a question on the pre-test as on the correlate question on the post-test. While this gold standard is optimal, it can also be nearly impossible to achieve within a reasonably scoped study. This is primarily because each intervention we create has its own learning goals, often, but not always, designed to fill a niche in a particular region or state's curriculum. Each assessment needs to be keyed to the precise research questions of the assessment and the research questions are tied directly to the learning goals. It is therefore quite expensive to create normalized and validated instruments that address each research question for each study. Still, when a pre- and post-test is used, it is important to understand and account for the compromises in the testing tool and process. It is possible to use some testing tricks to approach normalization even when optimal normalization cannot be assured. If the number of participants is sufficiently high, the most effective way of handling a non-normalized and validated test can be to randomize questions between the pre- and post-test. If each question type is represented in each test, the questions can be distributed across the pre- and post-test within each population studied so that half of people within each group, randomly determined within the group, receive the first of two equivalent questions in the pre-test, and the other half receive the other question. In the post-test, these are switched. Unless the tests are handled by a relatively sophisticated online process, this too can be difficult to manage, and online tests that include logins are not entirely anonymous and can be therefore difficult to use without risking some threats to protecting human subjects. It is also important to watch human impact on the study, including such things as participant exhaustion. A participant is not as able to answer test questions before a long game and after a long game session. One can help with this problem by administering pre- and post-tests on different days than the intervention itself, but this also poses many challenges, some examples include the potential loss of participants because they cannot or choose not to participate in each day of the assessment, organizational difficulty orchestrating the assessment across multiple days, and the potential for confounding results due to other experiences the students have during the elapsed time. The level of normalization is a decision based on research scope and potential impact, and compromises should be minimized, but expected.

Within a lab, a researcher might also have access to other methods of taking quantitative data. Tracking eye or head movement during play can be easily interpreted quantitatively and the results are often reproducible (Gomes, Yassine, Worsley & Bilkstein, 2013). Other forms of identifying what is happening during observed data can also be used to identify the instances or duration of behaviors. This can include taking video, audio, or transcribed data and marking the data with identifying marks, or codes, which indicate data where something is happening that is interesting and potentially relevant to the researcher's research questions. In addition to coding data, other measures can be used to identify what is occurring. Time on task and recording how long a player is engaged in each activity within the game, are methods of directly and quantitatively collecting data related to gameplay. Time on task data collection is, essentially, recording the amount of time spent playing the game, often breaking play down into specific tasks and determining how much time is spent with each task. Methods such as this can provide a quantifiable measure of key behaviors that provide evidence of mastery (Bell, 2008).

The methods used are generally those that provide the greatest confidence of the learner's level of attainment of mastery. In explaining the limitations of the study, an account should be made of conditions that prevented true mirroring of interventions. These may include such factors as the use

of similar but not identical classroom populations, limits in creating uniform content instruction, or potentially inconsistent levels of assistance given to students in performing in-game tasks. Given the realities of learning assessment, researchers cannot always control all variables, but a complete account helps future researchers to be aware of the limitations of the study.

Case Study Two: *Assessing Martha:* A mixed methods approach

The author was involved in mixed methods research on *Martha Madison's Marvelous Machines,* a learning game developed by Second Avenue Learning through a Small Business Innovation Research (SBIR) grant (Simkins, Egert & Decker, 2012). The game is a physics game targeted to teach the properties and uses of simple machines to middle school children, particularly middle school girls. This mixed methods approach used a combination of pre- and post-tests to show the effect of the intervention.

The study included three parts: a technical aptitude test, a Science, Technology, Engineering, and Mathematics (STEM) affinity test, and a content assessment targeted to the game's learning outcomes. The standardized technical aptitude test measured familiarity with web and PC applications. This test was given only during the pre-test and was used to determine if the population had the necessary skills to fully participate in the intervention without unintended effects due to unfamiliarity with the machines and controls. The STEM affinity test was derived from existing tests that show whether to what extent the participant sees STEM studies and STEM vocations as something in which they are capable and competent to engage. The content assessment tested the student's knowledge of the subject matter covered in the game and was adapted from standardized state assessments.

The pre- and post-tests were analyzed with typical statistical measures for test analysis, in this case one-tailed Wilcoxin signed-rank tests were used, given the sample size, controlled population, and types of questions. ANOVAs and t-tests are standard, when they are applicable.

In addition to pre- and post-tests, players were recorded playing the game and their in-game play was recorded. The in-game recording included a movie of each student's upper body as they played. This was matched with the in-game recording of their play, as their controlled characters moved through the game. The two video streams—in and out of game—were synchronized using pre-established markers as a beginning point for each. We did this by having the in-game characters perform a specific action that we recorded with the out of game camera by turning the camera on the screen. After the two streams were synchronized, we used Adobe Premiere to align the two videos into a single stream, side-by-side, for the purposes of data analysis. The side-by-side combined video was then coded using a pre-existing coding scheme.

Once the video was synchronized, researchers segmented the combined video into ten second chunks and codes were applied to each chunk. Using our pre-defined code set derived from similar research, researchers all coded one arbitrarily determined ten-minute section of video, recording all the codes

that were applied to each ten-second segment. This created 60 coded segments (10 minutes = 600 seconds = 60 chunks). This short 10-minute subset was used to determine the inter-rater reliability. That is, to determine if there was sufficient uniformity among researchers to treat one researchers coding of a segment as equivalent to any other researchers. Once our inter-rater reliability target of 95% accuracy was achieved (difference among codes <= 0.05), researchers then coded the thirteen videos, again breaking the video into ten-second chunks for the purpose of coding. The videos had an average of 219 segments, which equates with 36.5 minutes (2190 seconds) of play. Each of the codes either applied (1) or did not apply (0) to the ten-second chunk, using criteria finalized during the inter-rater reliability process. The coding spreadsheet included each code as a column, and researchers added chunks as rows filling out which codes applied for each chunk (see Table 3). Codes were non-exclusive and independent. Each code could either apply or not apply to each chunk with no assumption of positive or negative causality between any two codes.

Table 3. Sample coding segment including 13 codes in ten-second increments.

	PRO	PEX	EXE	PAS	COO	STP	IMP	WGO	ALG	WEX	DSF	PUZ	REC
5:00	x		x								X		
5:10	x		x										
5:20			x										
5:30	x												
5:40			x								X		
5:50	x		x										
6:00	x												

The outcome of this coding method was a large bank of data that could be used to test statistical hypotheses, a form of mixed methods data collection. Since the video was maintained intact, areas of note could also be evaluated through traditional qualitative methods, such as discourse or conversation analysis.

These methods of coding produce anonymized collections of data that can be used to compare among participants playing the same game or between games, so long as any of the same codes are used. Since each code is independent from the others, the entire code set does not need to be identical, as each code stands alone as either applying or not applying to each ten seconds of video. It is not clear as yet, but it is likely important that the chunk size not vary between data sets, as the size of the chunk has an effect on the relative complexity of the data coded. Longer chunks are likely to have more codes relevant to each chunk.

Once the data set is established, it is possible to identify relative percentages of each code, showing trends within a given intervention. It is also possible to run statistical comparisons among the codes to determine if there are trends of codes over time, if some codes tend to correlate with others.

There are limits of what hypotheses can be tested with this data, depending on what exactly is coded. For example, since nothing coded particular speech patterns among players, such as turn taking, there would be no way to look for turn taking in the coded data. Many hypotheses are relevant to the data coded, however, and the coding process creates a data set that is largely agnostic to the kinds of analysis one might want to perform on the data.

Mixed methods

As one might expect, mixed methods approaches supplement quantitative data collection and analysis with qualitative data collection and analysis. Generally, the intent is to approach data collection from a few directions, using complementary methods to heighten the strengths and mitigate the weaknesses of each approach. This variety of approaches, called triangulation, is the most common approach for gaining confidence in mixed methods assessments. The intent of triangulation is to accept the inability to completely control or understand the environment and to try and overcome this limitation by showing the alignment or disjunction of results gleaned from multiple types of inquiry. On one level, triangulation, or the use of multiple methods in coordination, occurs in most data collection, even within purely quantitative or qualitative approaches. The key of mixed methods research is that quite different forms of collection are used, such as using ethnographic interview and qualitative discourse analysis alongside traditional statistical methods, or by interpreting qualitative data as quantitative data through a process of numerically evaluating qualitative data. The goal should be to increase one's understanding of the whole learning experience by combining several methods across the qualitative-quantitative divide, acknowledging and working to enhance benefits and mitigate limitations of each approach to more completely describe the learning taking place during and around gameplay.

Pre- and post-tests are staples of both quantitative and mixed methods assessments of learning. As part of a mixed methods assessment, the test can be combined easily with other methods to triangulate effect. Another common method is to create a close but more generalizable read of qualitative data by identifying specific activities during play, called coding, which can then be potentially understood through quantitative analysis. These codes may be based on "top-down," pre-determined rubrics, or they may be based on "bottom-up" codes developed by the researcher from ongoing research into games and learning. A top-down coding scheme is used when the researchers already know what information they are looking for in the study. Bottom-up coding schemes are more often used in exploratory studies, or in ongoing process of creating sets of codes that can be applied to data. These approaches are not necessarily mutually exclusive, and researchers may choose to make multiple "passes" through codes, using top-down codes to identify what they know they will be interested in, and using a bottom-up

coding process to identify those activities they did not expect. This is time consuming, of course, but possible. It is also possible, especially in a new area of study, to use a bottom-up coding process on a substantial portion of the data to identify a set of desired codes, and then to use those codes in a top-down way to code the data.

Bottom-up coding schemes are often related to grounded theory approaches (Strauss, 1987; Glasser, 1992). In addition to grounded theory coding, thematic coding and clinical or standardized coding is common. Thematic codes identify tendencies or themes that recur in the coded data. Thematic codes are not justified exactly, but are designed by experts of phenomena, or in conjunction with expert insiders, to identify the interesting activities within a community practice. Standardized codes, including clinical codes, are systems of codes that have been used in previous studies and which have been given specific, normalized and reproducible definitions. All coding methods are also used to describe phenomenon in learning environments. While these coding schemes are not necessarily quantitative, they are compelling in part because they can be so easily converted into quantitative data, though quantitative interpretation of codes are not equally meaningful. The meaningfulness of this quantified qualitative data depends on the way the codes are determined and the quality and reproducibility of the coding scheme's results. As a result, clinical or standardized coding schemes are often produce the most meaningful and substantive quantitative data. There are several methods of achieving confidence in a coding system as a representation of quantitative as well as qualitative data. Central to them is the process of inter-rater reliability, which should be involved in all substantive coding processes. Developing inter-rater reliability involves testing the use of codes by multiple researchers coding the same content. These codes are then compared to determine that the coders are coding the same phenomenon the same way. To achieve parity of coding, researchers will need to engage in a process of learning and negotiating a uniform understanding of the precise meaning of each code within the group of researchers (Johnson, Penny, & Gordon, 2008). The somewhat arbitrary standard for acceptable inter-rater reliability is greater than 90% agreement when using a pairwise comparison of coded data. When comparing data between two coders, greater than 95% agreement is considered acceptable. The 95% agreement (>= 5% variance) is preferable in almost all studies.

As triangulation is generally central to mixed methods approaches, the qualitative and quantitative methods that are chosen are coordinated to complement each other. The key is to provide a convincing collection of data that can identify the successes and limitations of the learning environment and intervention. While more data may always seem better, it is important not to take data based on different initial premises and epistemologies and then interpret them as if they were coherent with each other. While pragmatist epistemologies may be able to interpret almost all methods as useful to an increased understanding of phenomenon, and could find useful comparisons among almost any sets of data, positivist empirical epistemologies would have use for most qualitative data, and most post-modern epistemologies would have little use for data claiming to be universal.

Qualitative methods

Qualitative methods involve collecting data on what people are doing within their context. This involves a very close read of the actions, speech, practices, and behaviors—words that may or may not mean the same things, depending on one's qualitative tradition. It is also important to qualitative researchers to provide a close read of the environment, social, cultural, and physical, and to provide, in analysis, an account of how these phenomena effect, correlate with or interact with each other. I use the term tradition here because multiple traditions exist within the same methodology. For our purposes here, tradition refers to one's qualitative research style. Methodologies relate to one's ontological and epistemological beliefs, which is "what is" and "how we can come to understand it," respectively.

Regardless of tradition or methodology, qualitative methods are used to tell the story of the intervention and the students' passage through it. In some traditions this storytelling is a metaphor, and the story is a description of what occurred. In other traditions, the researcher's role is quite literally to depict a story of what occurred through, for example, writing, film, or theater. In either case, the goal is to produce a substantive and knowledge-producing account.

Qualitative inquiry can be broken into two loose categories—ethnography and case study. Ethnography is the study of culture. It is an in depth, all-inclusive form of inquiry involving involvement in the practices of a culture, recording of field notes, and reporting out in a way that produces deep understanding of the target population. One major strand of ethnography follows Geertz (1973) methods for producing what he calls "thick description" of culture. Thick description is produced through a multi-layer account of many events that bring into analysis multiple perspectives, which are sensitive to and include within the description the role of as many contextual influences as possible. While much of quantitative analysis seeks to reduce the effect of outside influence from the description of the event, ethnographic analysis seeks to incorporate a rich and complete description of contextual influences into the description of the event. Whereas most quantitative analysis finds greatest utility in that which can be abstracted, qualitative analysis finds greatest utility in that which can be understood wholly only within a complex context.

In contrast to ethnography, case study is narrower in focus. Rather than studying culture as a whole, case study takes a narrower view, perhaps focusing on a single event, person, or group. Due to its narrower focus, case study is more often utilized in games assessment. There are ethnographies that focus on games and learning (Steinkuehler, 2004), but the focus on ethnography as a study of culture often precludes it from looking at a particular game as effective for learning. Assessment of the game for learning may be a part of the whole, but it is only a part of the whole.

Still, while ethnography is a larger enterprise, many qualitative case studies that study the efficacy of games make use of ethnographic methods to gain a rich understanding of what is occurring in and around gameplay. Ethnographic observation and interviews are methods used within many case studies.

Whether case study or ethnography, qualitative methods require the same three stages as all research—data collection, data analysis, and reporting. Data collection is dominated by traditional ethnographic methods, but analysis is varied in both ethnography and case study. Contemporary qualitative research in games and learning utilizes a variety of methods of analysis, including ethnomethodology (Garfinkle, 1967), conversation analysis (Sacks, 1992), discourse analysis (Gee, 2005), expert-novice study (Chase & Simon, 1973), narrative analysis (Bruner, 1990), and practice theory (Bourdieu, 1977). The diversity of methods is, in part, due to the descriptive nature of qualitative inquiry, and there is significant overlap and often non-distinctive lines between different approaches. The best methods will be those that help to make a case for the affordances and limitations of learning that takes place in play and helps others to create effective learning environments or games (see Table 2). The rigor of the method is reflected in the degree to which the data is included and interpreted fairly and completely within the complexity of its context. Its usefulness will be in the researchers ability to synthesize a meaningful narrative from that complex data such that the reader comes away with a deeper understanding of the subject of research, in this case the learning the occurred during play.

Table 2. Strengths and challenges in sample qualitative methods

	Strengths	Challenges
Ethnomethodology/ practice theory	Deep focus on practice can reveal ways in which learning turns into legitimate participation.	Time consuming. Focus on process and practice may be too restrictive for most games and learning assessments.
Conversation analysis	Reveals language as facilitation and constraint of activity. Identified meaning, norms, and action in text.	Requires focus on a small data set. Limited use of context.
Critical discourse analysis	Contextualizing text can expand understanding of language to understanding of practice.	Time consuming. Requires extensive understanding and engagement with context.
Expert-novice study	Can show process of how learners develop into experts. Can identify patterns in error as one develops understanding.	Requires a prior understanding of expertise in a practice.
Narrative analysis	Able to identify the conveyed reasons behind practice— the "why" of practice. Identifies ways that knowledge is shared.	Limited coherence between "objective" learning goals and the participant-focused assessment of meaning may limit use when specific learning goals are being studied, rather than the process of learning.
Ethnography (cultural, cognitive, etc.)	Thick description can provides deep understanding of phenomena.	Time consuming. Data collection and sifting and winnowing process of analysis places little or no value on research efficiency. Can treat culture as static rather than dynamic.

As with quantitative assessment, the qualitative researcher will need to attend to the central learning objectives of the game, and the assessment will hold the intervention to its ability to achieve these objectives. More than with quantitative research, however, qualitative research can identify previously unforeseen learning occurring within the phenomenon. This is in large part because qualitative research is concerned not with describing conformity to what was expected to happen, but to accurately describe what did happen. To do this it is important for the researcher to be able to be surprised, without necessarily trying to be surprised, by what occurs during observations, which allows for previously unexpected observations.

The greatest strength of qualitative research is in its ability to incorporate the effects of context and to explain the significance of context within learning. It is able to mark the process of learning over time while incorporating the context and to identify trends and changes within a single research participant. The cost of this is high in terms of time required for data collection and analysis. The time and amount of access required by qualitative researchers generally means that they follow a very limited number of participants. The method of describing learning process in qualitative inquiry is not seeking wide-scale verifiability and it would be impossible to recreate a learning environment exactly as it occurred in the qualitative analysis.

Case Study Three: Cognitive Ethnography of *Lineage*

Ethnographies are well-discussed, particularly within the field of anthropology. A rigorous, long-term ethnographic inquiry (Geertz, 1973; Malaby, 2003; Chen, 2012; Simkins, in press) may be the best possible approach to qualitative research, when time and access allows. Within educational research, researchers often lack either time or access to complete ethnographic inquiry, which requires months of intense participation within the community, but those that are completed can provide much needed insight into learning.

Conducting an ethnography requires a holistic approach to the understanding of culture. The social, cultural, and physical environment is explored from a contemporary and historical perspective. Analysis of observation involves a move between emic points of view, those from within the studied culture, and etic points of view, those from outside the culture, usually primarily focused on the perspective of the researcher.

One educational games ethnography was conducted by Steinkuehler (2004) over the course of two years, through interacting with a particular community of gamers who started together on *Lineage*, a massively multiplayer online game released by Korean company NC Soft. The community was on English speaking servers, and included participants from across the world as they played *Lineage*, and later moved on to *Lineage 2* and *World of Warcraft*.

Over the course of two years, Steinkuehler had the opportunity to play with a large variety of game players, learn specialist language around the game, and explore the specific practices that signify expert game playing. The work was largely analyzed through discourse analysis, and included a number of analytic tools including various kinds of expert-novice studies, studies of ethical play, cheating, ways of playing that mark one as an insider or outside to the core community of the game, and how one transitions from peripheral to central participant. A lot of time and energy is spent on understanding what it means to be full member of the practice, what it means for one's sense of being and identity as a leader, follower, and member of community.

Through this inquiry, a core group of practitioners became her core participants, and playing involved building trust and care relationships with her participants. Eventually, Steinkuehler became a community leader in her own right, having established herself as a trustworthy and valuable member of the community. This was not intended, nor particularly desirable for Steinkuehler as a researcher, but it did open doors to understanding all sides of the complex negotiations that underlie forming and maintaining a group of players within each of these games—each of which have particular challenges and affordances when it comes to developing meaningful connections between community members.

While tools of analysis vary widely, data collection is more uniform, and both are in ample evidence in Steinkuehler's ethnographic data. The first is an extensive record of observation of activities in and around play. The second is evolving interviews with key informants who can describe and explain the practices of the community, and help the researcher to interpret meaning of behaviors and patterns evident in observations. These interviews may be formal interview interactions where the researcher and participant are self-consciously engaging in an interview. It can also be informal interview interactions that occur during normal interactions in and around the games. As with almost any other modern research, each of the participants was aware that they were engaged in research, and each was identified by a pseudonym that protected them from potential social ramifications for what they might have said. The combination of interviews and observations allowed Steinkuehler to gain an in-depth understanding of these communities of practice as they played the three MMOs.

Best Practices

1. Identify the game's learning goals, both explicitly and implicitly.
2. Create assessment methods that maximize access to and understanding of data relevant to the learning goals. Ensure that the methods are compatible with researcher's epistemological commitments.
3. Assess the game using rigorous standards.
4. Report on your assessment, clearly identifying your epistemology, methods used, and learning goals assessed. Include a good description of the gameplay that shows the relevance of learning goals to the gameplay.

Resources

Epistemology and Methods

Cresswell, J. W. (2008). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches.* LosAngeles, CA: Sage Publications.

Greeno, J., Collins, A. M., & Resnick, L. (1996). *Cognition and Learning.* (pp. 15–46) In D. Berliner and R. Calfee (Eds.), *Handbook of Educational Psychology.* New York: Macmillan.

Levison, S. C. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press.

Assessment

Annetta, L. & Bronack, S. (2011). *Serious Educational Games Assessment: Practical Methods and Models for Educational Games, Simulations and Virtual Worlds*. Rotterdam, The Netherlands: Sense Publishers.

Shute, V. J. (2011). Stealth Assessment in Computer-Based Games to Support Learning. In S. Tobias & J. D. Fletcher (Eds.), Computer games and instruction (pp. 503–524). Charlotte, NC: Information Age.

Learning Game Design

Bogost, I. (2007). *Persuasive Games: The Expressive Power of Videogames.* Cambridge, MA: MIT Press.

Gee, J. P. (2007). *What Videogames Have to Teach us about Learning and Literacy.*New York: Palgrave MacMillan.

Salen, K., & Zimmerman, E. (2005). Game Design and Meaningful Play. In J. Raessens & J. Goldstein (Eds.), *Handbook of Computer Game Studies* (pp. 59–80). Cambridge, MA: MIT Press.

Squire, K. (2011). *Videogames and Learning: Teaching and Participatory Culture in the Digital Age.* New York: Teacher's College Press.

References

Annetta, L. & Bronack, S. (Eds) (2010). *Serious educational game assessment: Practical methods and models for educational games, simulations, and virtual worlds.* Boston, MA: Sense Publishers.

Bell, Ann (2008). *Game rubric: Assessing student learning in virtual simulations and serious games.* Downloaded on July 17, 2013 from http://www2.uwstout.edu/content/profdev/rubrics/gamerubric.html .

Betz, J. (1995-6). Computer games: Increase learning in an interactive multidisciplinary environment. *Journal of Educational Technology Systems.* 24(2), pp 195-205.

Black, R.W. (2008). *Adolescents and online fan fiction.* New York: Peter Lang.

Bourdieu, P. (1977). *Outline of a theory of practice.* Cambridge, MA: Cambridge University Press.

Bruner, J. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Chase, W. G. & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology.* 4, 55-81.

Chen, M. (2012). *Leet noobs: The life and death of an expert player group in World of Warcraft.* New York: Peter Lang.

Clark, D. B., Tanner-Smith, E. E., Killingsworth, S., & Bellamy, S. (2013). *Digital games for learning: A systematic review and meta-analysis.* Menlo Park, CA: SRI International.

Cresswell, J. W. (2008). *Research design: Qualitative, quantitative, and mixed methods approaches.* LosAngeles, CA: Sage Publications.

D'Angelo, C., Rutstein, D., Harris, C., Haertel, G., Bernard, R., & Borokhovski, E. (2013). *Review of computer-based simulations for STEM learning in K-12 education.* Menlo Park, CA: SRI International.

Dibbell, J. (2006). *Play money: Or, how I quit my day job and made millions trading virtual loot.* New York: Basic Books.

Garfinkle, H. (1967).*Studies in ethnomethodology.* Englewood Cliffs, CA: Prentice Hall.

Gee, J. P. (2005). *An introduction to discourse analysis: Theory and method.* London: Routledge.

Gee, J. P. (2007). *What videogames have to teach us about learning and literacy.* New York: Palgrave MacMillan.

Gee, J. P. (2012). *Big "G" games.* Downloaded on July 17, 2013 from http://www.jamespaulgee.com/node/63 .

Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In *The Interpretation of Cultures: Selected Essays.* New York: Basic Books, 3-30.

Glaser, B. (1992). *Basics of grounded theory analysis.* Mill Valley, CA: Sociology Press.

Gomes, J. S., Yassine, M., Worsley, M. & Bilkstein, P. (2013). Eye tracking analysis of visual spatial engineering games.In the proceedings of *The Sixth Annual Conference on Educational Data Mining* (EDM 2013). Memphis, TN, July 6-9.

Greeno, J., Collins, A. M., & Resnick, L. (1996). Cognition and learning. (pp. 15–46) In D. Berliner and R. Calfee (Eds.), *Handbook of Educational Psychology.* New York: Macmillan.

Hammersley, M. & Atkinson, P. (1986). *Ethnography: Principles in practice (2nd ed.).* London: Routledge.

Ifenthaler, D., Eseryl, D., & Ge, X. (2012). *Assessment in Games based learning.* New York: Springer.

Jenkins, H. (2012). *Textual poachers: Television fans and participatory culture.* New York: Routledge.

Johnson, R. L., Penny, J. A., & Gordon, B. (2008). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: The Guilford Press.

Kafai, Y. & Fields, D. (2009). Cheating in virtual worlds: Transgressive designs for learning. *On the horizon 17*(1), 12-20.

Keegan, B., M. Ahmad, J. Srivastava, D. Williams, N. Contractor (2011). Dark gold: Statistical properties of clandestine networks in massively multiplayer online games. *International Journal of Social Computing and Cyber-Physical Systems.*

Latour, B. (2004). Why has critique run out of steam? From matters of fact to matters of concern. *Critical Inquiry 30*(2),25-248.

Levison, S. C. (1983). *Pragmatics.* Cambridge, UK: Cambridge University Press.

Magnifico, A. (2010). Writing for whom? Cognition, motivation, and a writer's audience. *Educational psychologist 45*(3), 167-184.

Malaby, T. (2003). *Gambling life: Dealing with contingency in a Greek city.* Urbana, IL: University of Illinois Press.

Menand, L. (2001). *The metaphysical club.* New York: Farrar, Strauss, and Giroux.

Pallas, A. M. (2001). Preparing education doctoral students for epistemological diversity. *Educational Researcher* 30(5), pp 6-11.

Sacks, H. (1992). *Lectures on conversation, volumes I and II.* Oxford, England: Blackwell.

Science Applications International Corporation (SAIC) (2007). Games: A look at emerging trends, uses, threats and opportunities in influence activities. Downloaded from *ProPublica* on 2/20/2014. http://www.propublica.org/documents/item/889134-games

Shaffer, D. W. (2006). *How computer games help children learn.* New York: Palgrave Macmillan.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503–524). Charlotte, NC: Information Age.

Simkins, D. (in press). *The arts of LARP: Design, literacy, learning and community in Live Action Role Playing.* Jefferson, NC: McFarland Press.

Simkins, D., Egert, C., and Decker, A. (2012). Evaluating *Martha Madison:* Developing analytical tools for gauging the breadth of learning facilitated by STEM Games. In *Proceedings from the 2012 IEEE International Games Innovation Conference,* pp. 137-140. Rochester, NY, September 7-9, 2012.

Simkins, D. & Steinkuehler, C. (2008). Critical ethical reasoning & role play. *Games & Culture,* 3, 333-355.

Squire, K. (2011). *Videogames and learning: Teaching and participatory culture in the digital age.* New York: Teacher's College Press.

South, J. & Snow, B. (2012). Immersive game design: Aligning game mechanics with learning goals to maximize engagement and mastery. 18th Annual SLOAN Consortium International Conference on Online Learning, October 11-12.

Steinkuehler, C. A. (2004). *Learning in Massively Multiplayer Online games.* Downloaded on 11/1/2013 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.105.626&rep=rep1&type=pdf

Steinkuehler, C. (2007). Massively multiplayer online gaming as a constellation of literacy practices. In B. E. Shelton & D. Wiley (Eds.), *The Design and Use of Simulation Computer Games in Education* (pp. 187-212). Rotterdam, The Netherlands: Sense Publishers.

Steinkuehler, C., King, E., Alagoz, E., Anton, G., Chu, S., Elmergreen, J., Fahser-Herro, D., Harris, S., Martin, C., Ochsner, A., Oh, Y., Owen, V. L., Simkins, D., Williams, C., & Zhang, B. (2011). Let me know when she stops talking: Using games for learning without colonizing play. In C. Steinkuehler, C. Martin, & A. Ochsner (Eds.), *Proceedings of the 7th Annual Games+Learning+ Society (GLS) Conference.* Pittsburgh PA: ETC Press.

Squire, K. (2005). Changing the game: What happens when videogames enter the classroom? *Innovate Journal of Online Education*. 1(6).

Squire, K. (2011). *Video games and learning: Teaching and participatory culture in the video age.* New York: Teacher's College Press.

Strauss, A. (1987). *Qualitative analysis for social scientists.* Cambridge, England: Cambridge University Press.