Gameplay as Assessment: Analyzing Event-Stream Player Data and Learning Using GBA (a Game-Based Assessment Model)

V. Elizabeth Owen, University of Wisconsin-Madison (*GLS Center*), vowen@wisc.edu
R. Benjamin Shapiro, Tufts University, ben@cs.tufts.edu
Richard Halverson, University of Wisconsin-Madison (*GLS Center*), halverson@education.wisc.edu

Abstract: This extended study (building on pilot research) presents a Game-Based Assessment model (GBA) designed to capture relevant information on play and test whether it can constitute reliable evidence of learning. A central challenge for videogames research in education is to demonstrate evidence of player learning. Assessment designers need to attend to the ways in which gameplay itself can provide a powerful new form of assessment. The GBA model has two key layers which build on content-based educational game design: a semantic template that determines which click-stream data events could be indicators of learning; and learning telemetry that captures data for analysis. This study highlights how the GBA was implemented in a stem-cell science learning game, and shows how the GBA demonstrates a relationship among success, kinds of failure, and learning in the game.

Objectives and Theoretical Framework

A central challenge for videogames in education is to demonstrate evidence of player learning. A typical approach to assess learning in games is to measure the quality of player learning in terms of independent, pre-post instruments. This process can compare game-based learning against other kinds of interventions, but, in treating the game itself as a black box, we lose the unique characteristics of the games as a learning tool. James Gee has suggested that games themselves provide excellent models for designing the next generation of learning assessments. Well-designed games reward players for mastering content and strategies, scaffold player activities toward greater complexity, engage players in social interaction toward shared goals, and provide feedback (through gameplay) that allows players to monitor their own progress (Gee, 2005). Rather than ignore the motivating and information-rich features of games in capturing learning, assessment designers need to attend to the ways in which gameplay itself can provide a powerful new form of assessment. This requires learning researchers to think of games as both intervention *and* assessment; and to develop methods for using the internal structures of games as paths to generate evidence of learning.

This study's framework is the Game-Based Assessment model (GBA), designed to capture data on player learning in the midst of gameplay. It's an extension of GBA pilot research, which introduced the model and preliminary findings around gameplay patterns and learning (Owen et al., 2012). The GBA framework has been developed by the Games, Learning and Society (GLS) research group as a process for capturing relevant information on play and testing whether it can constitute reliable evidence of learning. The GBA model draws on concepts and tools from evidence-centered design (e.g., Mislevy & Haertel, 2006), stealth assessment (Shute, 2011) and educational data mining (e.g. Baker & Yacef, 2009) to describe a strategy for building assessment tools into game design from the ground up in order to use game play itself as the barometer of player learning.

GBA Model and Methods

The Game-Based Assessment model is grounded in the content model and game-flow design of the game development process, and emphasizes two key layers: the *semantic template* and *learning telemetry*. Below, we describe each feature of the model in context of *Progenitor X*, a GLS game about regenerative medicine.

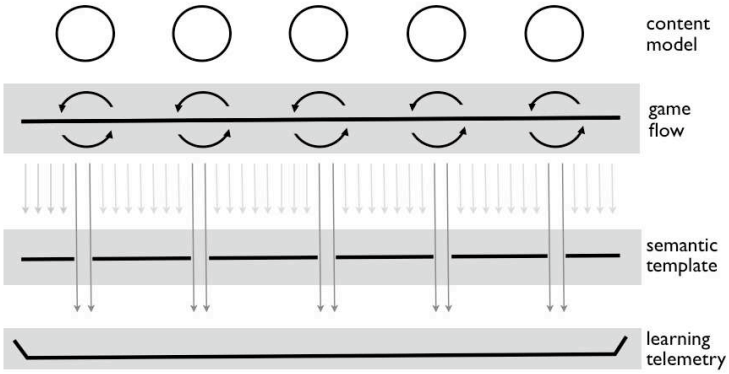


Figure 1: Game-Based Assessment model

The GBA model is designed to draw significant gameplay moves from the game-context. The model is integrated into an overall four-layer GLS game design strategy: the *content-model*; the *game flow* design; the *semantic template*; and the *learning telemetry* (Figure 1). The first two layers, the content model and the game flow design, constitute the game design process. The content model outlines the learning goals for the game. The game flow design builds player interaction opportunities around these learning goals to create a gaming experience. The final two layers, the semantic template and the learning telemetry, form the assessment process. The semantic template selects relevant data from the click-stream generated by gameplay; the learning telemetry layer collects and organizes the resulting data-record into player-profiles. Here we provide a brief overview of how these layers, using the game *Progenitor X* as an example, comprise a generic blueprint for our approach to assessment-driven game design.

Content Model

The content model for a GLS game consists of several content chunks that string together a series of core concepts along a process that represents current thinking in a domain. Because the resulting medium for interaction is a game, rather than a simulation, the design team is concerned with creating motivating conditions of play as well as the representational accuracy of the content model. *Progenitor X* provides an example.

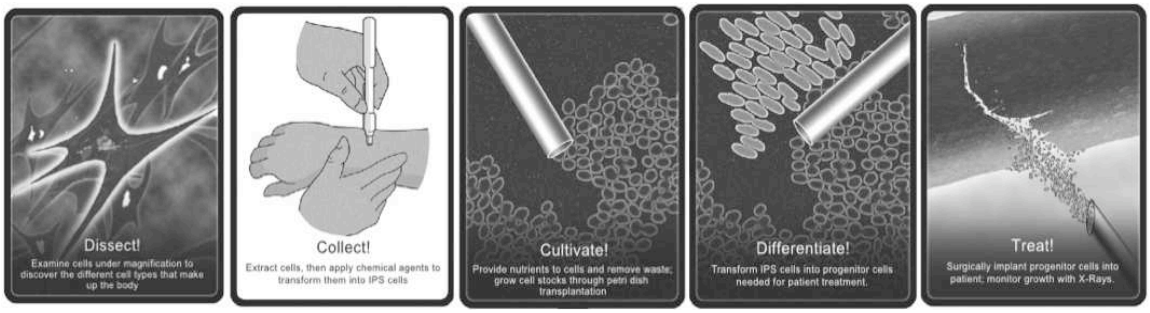


Figure 2: *Progenitor X* content model

Progenitor X invites players to dissect, collect, cultivate, differentiate and treat diseased tissues via adult epithelial stem cells (Figure 2). Each verb in the content model provides an occasion for interaction. A process derived from professional practice provides a simplistic but coherent account of real scientific procedures, designed for accessibility to the study demographic of secondary school students.

Game-flow Design

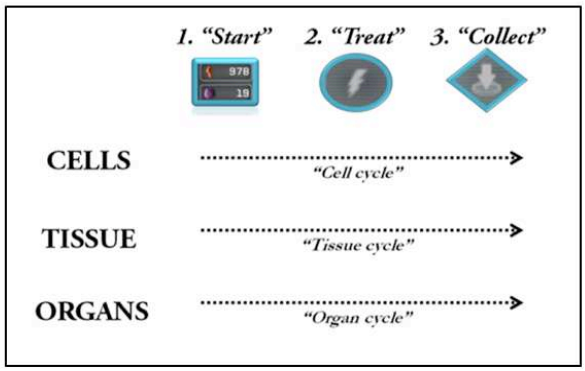
The game is designed to motivate player interaction and learning. Through the iterative design process, the content model is embedded in a world that allows players to interact with the core ideas. The verbs of the content model are translated into key moments in interactive gameplay. *Progenitor X* embodies this process, taking the verbs of the content model and creating a turn-based puzzle game in which players assume the role of a regenerative biologist to prevent a zombie apocalypse. Given a series of content-based objectives, *Progenitor* players perform three main actions in game-flow: cultivate (or ***start*** a cycle of) cells, ***treat*** them, and then ***collect*** the resulting target material.

Figure 3: *Progenitor X* game-flow design

Semantic Template

The semantic template defines conceptual windows of interest in the game that represent key moments of learning. It is designed around the intersection of the content model with the game-flow design. The key question for semantic template design is: of all the clicks that players make in the game, which ones indicate learning? The semantic template represents a hypothesis about which in-game actions can generate interesting evidence of learning.

In *Progenitor X*, the semantic template revolves around the *start*, *treat*, and *collect* verbs of the content model. The first sequence of player action is the **cell cycle**, in which players *start*, *treat*, and *collect* a group of vital cells. These new cells are used to create tissue in the next cycle (i.e. **tissue cycle**), where players use the same action sequence. Then comes an **organ cycle**, where the player uses the newly collected tissue to *start, treat, and collect* their way to a whole, healthy organ. Two views of the semantic template in *Progenitor* are shown below.



	A	B	C	D	F	G
1	In-Game Sequence	Mission	Cycle**	Cell population type (Stage 1)	Treatment tool type (Stage 2)	Collection cell type (Stage 3)
2	1	1	A. ips	fibroblasts	electroporate	ips
3	2	1	B. meso i	ips	growth factor	meso
4	3	1	A. ips ii / C. ecto i	fibroblasts	electroporate	ips
5	4	1	C. ecto ii	ips	growth factor	ecto
6	5	2	E. tissue i	meso	N/A	meso / tissue
7	6	2	E. tissue ii	meso	N/A	meso / tissue
8	7	2	A. ips iii / D. endo i	fibroblasts	electroporate	ips
9	8	2	D. endo ii	ips	growth factor	endo
10	9	2	E. tissue iii	endo	N/A	endo / tissue
11	10	3	F. organ i	N/A	scan	necrotic tissue
12	11	3	E. tissue iv	meso	N/A	meso / tissue
13	12	3	F. organ iii	N/A	scan	necrotic tissue
14	13	3	A. ips iv / B. meso ii	fibroblasts	electroporate	ips
15	14	3	B. meso iii	ips	growth factor	meso

Figure 4: A) Basic and B) Detailed semantic templates of *Progenitor X*

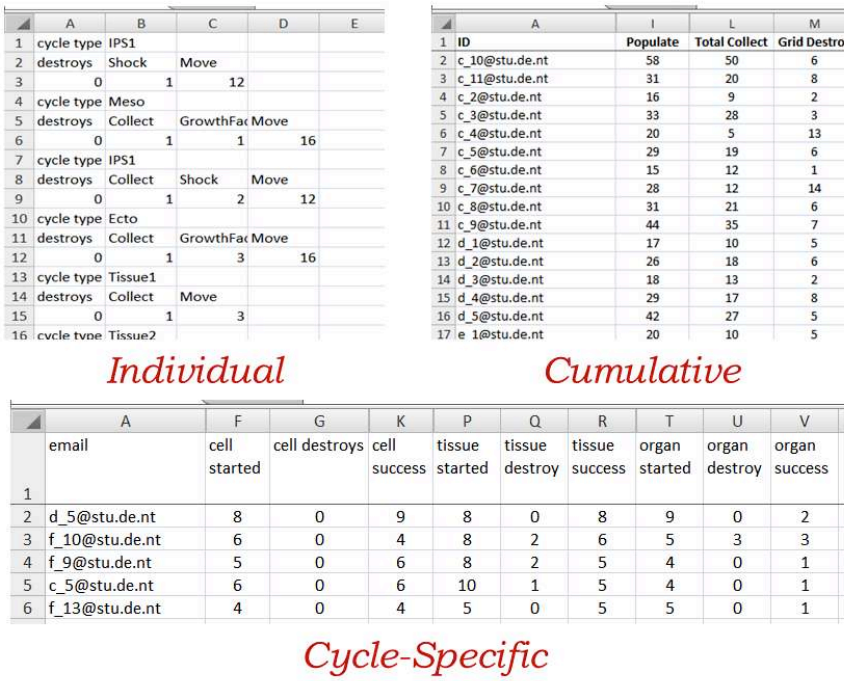
Learning Telemetry

The learning telemetry layer collects the data specified by the semantic template and organizes it for analysis. It is a mechanism of the game environment that coordinates the components of the game world into a sequential data-stream that enables analysts to track player paths across the game-world.

In *Progenitor*, capturing telemetry started with identifying gameplay moments within the semantic template on an event-stream level. Significant click-stream events (over 400) around the action sequence (*start*, *treat*, and *collect*) were documented and flagged for recording. Then, search parameters were constructed, allowing reconstruction of interface cues as context for player actions. Lastly, a query schema was developed to pull the specified event-stream data from the massive database. Ultimately, through synchronizing GBA's semantic template and learning telemetry, we were able to identify and collect three kinds of telemetric action-sequence data: cycle-specific, cumulative, and individual.

▼ _id	4fbd2a48cf86304117000004	ObjectId
_id	4fbd2a48cf86304117000004	ObjectId
created_at	2012-05-23 13:19:52 -0500	Date
gameName	ProgenitorX	String
key	ToolSelectedData	String
schema	4-18-2012	String
session_token	2012-05-23_115607	String
timestamp	854	Int
toolName	Move	String
updated_at	2012-05-23 13:19:52 -0500	Date
user_id	235	Int
wasFlashing	YES	Bool
▶ _id	4fbd2a48cf86304117000005	ObjectId

Figure 5: Learning telemetry - raw click-stream data output *(information from one click)*



	A	B	C	D	E
1	cycle type	IPS1			
2	destroys	Shock	Move		
3	0	1	12		
4	cycle type	Meso			
5	destroys	Collect	GrowthFac	Move	
6	0	1	1	16	
7	cycle type	IPS1			
8	destroys	Collect	Shock	Move	
9	0	1	2	12	
10	cycle type	Ecto			
11	destroys	Collect	GrowthFac	Move	
12	0	1	3	16	
13	cycle type	Tissue1			
14	destroys	Collect	Move		
15	0	1	3		
16	cycle type	Tissue2			

	A	I	L	M
1	ID	Populate	Total Collect	Grid Destro
2	c_10@stu.de.nt	58	50	6
3	c_11@stu.de.nt	31	20	8
4	c_2@stu.de.nt	16	9	2
5	c_3@stu.de.nt	33	28	3
6	c_4@stu.de.nt	20	5	13
7	c_5@stu.de.nt	29	19	6
8	c_6@stu.de.nt	15	12	1
9	c_7@stu.de.nt	28	12	14
10	c_8@stu.de.nt	31	21	6
11	c_9@stu.de.nt	44	35	7
12	d_1@stu.de.nt	17	10	5
13	d_2@stu.de.nt	26	18	6
14	d_3@stu.de.nt	18	13	2
15	d_4@stu.de.nt	29	17	8
16	d_5@stu.de.nt	42	27	5
17	e_1@stu.de.nt	20	10	5

	A	F	G	K	P	Q	R	T	U	V
1	email	cell started	cell destroys	cell success	tissue started	tissue destroy	tissue success	organ started	organ destroy	organ success
2	d_5@stu.de.nt	8	0	9	8	0	8	9	0	2
3	f_10@stu.de.nt	6	0	4	8	2	6	5	3	3
4	f_9@stu.de.nt	5	0	6	8	2	5	4	0	1
5	c_5@stu.de.nt	6	0	6	10	1	5	4	0	1
6	f_13@stu.de.nt	4	0	4	5	0	5	5	0	1

Figure 6: Learning telemetry - processed click-steam data output

Data Sources and Evidence

Data analysis required synthesizing learning telemetry data output with additional assessment. Specifically, we added two additional data sources to the core telemetric corpus: an adapted measure of success in gameplay, and data from an isomorphic pre- and post-test.

In order to sort the player data into meaningful patterns, we developed an *efficiency ratio* that measured the number of successful cycle completions by a player over the number of times the cycle was tried. (A "success" means the player has collected the right material at the end of the cycle.) For example, if a player successfully collected the required number of cells in a cycle 2 times, and tried to complete the cycle five times, the player's efficiency ratio would be 40%. (The higher the percentage, the more efficient the play.)

Efficiency Ratio = # of successes / # of tries

We also aggregated results from the pre- and post- content assessment, which included a series of questions about the stem-cell content model based on consultation with regenerative biologists Dr. James Thomson, Dr. Rupa Shevde, and Dr. Gary Lyons. Here, we specifically looked at change in player performance on content questions as measured before and after gameplay.

Results

Along with the telemetry data, players' efficiency ratio and the change in performance on the pre-post content questions became key data features. In this study, these features were analyzed within the aggregate data set of n=110 with nonparametric statistical methods, given the non-normal distribution of pre-post percentage change

and event-stream variables. The tests were directional (one-tailed) and conducted at a baseline alpha of .05, guided by our main hypothesis that on-task gameplay would result in increased learning outcomes (as measured by pre-post performance). Specifically, we used a paired-sample Wilcoxon rank test (Table 1) and Spearman's correlation test throughout the analysis. Multiple pairwise contrasts were conducted using the Holm procedure for assigned alpha, centering around constructs of play efficiency (see "Efficiency Ratio" above) and temporal game progression. The data was collected in the summer of 2012 from 110 randomly-selected middle-school students who played *Progenitor X* as part of a summer school curriculum *(*either in their Dane County school classroom, or on-site at the Wisconsin Institute for Discovery).

Aggregate results revealed intriguing reasons to look further into the "black box" of the game. First, with a 19.5% average increase in pre-post content scores, the game seemed to be a noteworthy learning vehicle. A first look at player progress through the game revealed a significant positive relationship between successive completion of game objectives and learning (r = +.272, p = .002). Success as well as game progression mattered; the number of successful cycles in gameplay was positively correlated with learning outcomes (r = +.216, p = .012). Thus, it seemed that player performance during specific points throughout the game held a key to deeper understanding of the relationship between gameplay and learning outcomes. This led us to investigate what was going on with players on a micro level within each given cycle.

Table 1: Aggregate results summary

	Pre-Post Gains
Total Gameplay	19.5% average increase
Objectives Added	Significant positive correlation (r = +.279, p = .002)
Objectives Completed	Significant positive correlation (r = +.272, p = .002)
# of Successful Cycles	Significant positive correlation (r = +.216, p = .012)

In order to examine player interaction, we mapped all possible cycle outcomes. Within a cycle, players populate (*start*) an initial grid with the right kinds of cells (green check, Figure 7), and then transform those cells (*treat*) into a target cell/tissue to *collect*. After initial population with the right cell, the cycle can end in three ways: collecting the right cell (success), collecting the wrong cell (failure), or over-manipulating/treating the cells so that the Ph becomes toxic (failure).

Additionally, a player could have also initially populated the grid with the wrong cell (red X, Figure 7). In this case, there are two options for ending the cycle: collecting the wrong cell, or over-manipulating the cells until the Ph levels (health) becomes toxic.

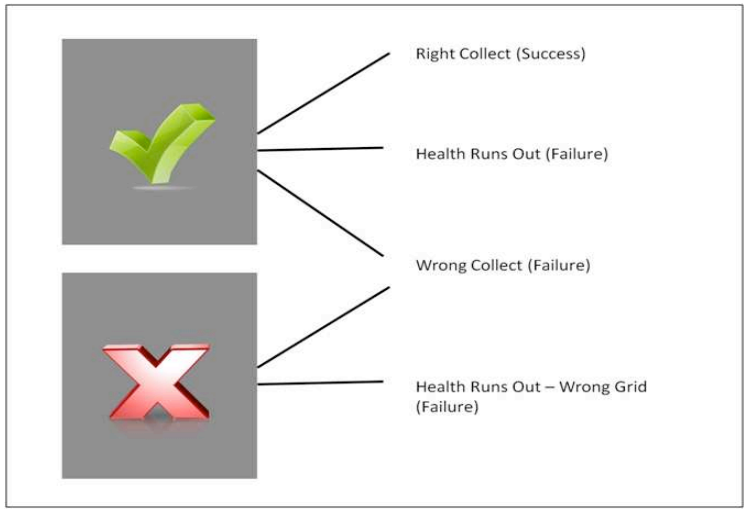


Figure 7: *Progenitor* gamespace - incorrect AND correct initial grid population

The possible outcomes imply varying degrees of player compliance with multiple in-game cues (e.g. flashing buttons & in-game narration). To explore this idea, we clustered the types of failures into "near" and "far failure" (Figure 8). We grouped three possible player outcomes: correct collection (successful); correct set-up but health runs out ("near failure"); incorrect setup and/or incorrect collection ("far failure").

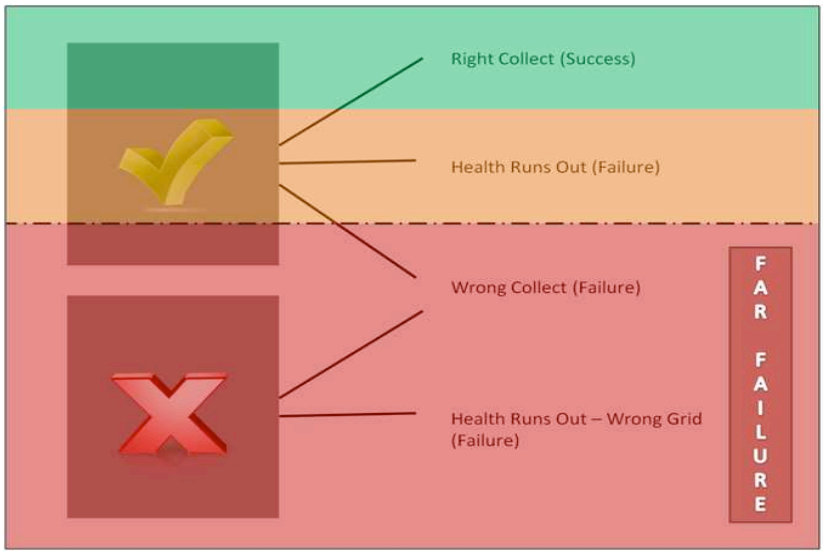


Figure 8: "Far failure" in *Progenitor* gamespace

The analysis of far failure gave considerable insight, especially after parsing game data by level sequence (from the beginning Objective 1 to the final Objective 8). Far Failure and success in the first and last objectives in the game proved vital, with implications for early scaffolding and "boss-level" assessment. Objective 1, the game level immediately after the tutorial, seemed to be a critical filter point for those trying to "game the system" by not attending to instructional cues (expressed in far failure numbers). Within Objective 1, far failure had a significant negative correlation with learning outcomes (Table 2). In fact, for players with extreme numbers of Objective 1 far failure (over two standard deviations from the mean), the average increase on the pre-post biology assessment was only 3.6% (as opposed to a 19.5% in the aggregate group). Conversely, success in Objective 8 (the "boss" level) had significant positive correlations with pre-post gains, both in terms of raw number of successes (r = +.272) and efficiency ratio (r = +.193). (The final stage, this Objective 8 "boss" level, required a cumulative performance of all lab skills learned in the game.) Thus, far failure and success in these bookend levels seemed to hold critical significance for learning outcomes.

Table 2: Objective-level aggregate results summary

	Pre-Post Gains
Objective 1 Far Failure	Significant negative correlation (r = -.167, p = .04)
Upper Extreme: Players w/ Obj 1 Far Failure	3.9% average increase *(15.6% lower than aggregate)*
Objective 8 Successes	Significant positive correlation (r = +.272, p = .002)
Objective 8 Efficiency Ratio	Significant positive correlation (r = +.193, p = .022)

To explore these phases of learning connection more deeply, we divided players into quartiles according to pre-post change. The upper quartile (33 students) had the largest gains in stem-cell content question scores, while the lower quartile (41 students) had the smallest. Interestingly, the patterns of play in each quartile supported the trends from the aggregate analysis: overall game progression and success was positively connected with pre-post gains, as was performance on the boss level, while far failure in critical early game cycles was negatively associated with learning outcomes.

Overall quartile trends revealed key differences in play progression, success, and far failure. During the same duration of gameplay, Objective progression was significantly different for the two groups (p = .019). The upper quartile, on average, got to Objective 7 (out of eight total), while the lower quartile made it to Objective 6 (Table 3). Like total playtime, the number of total successful cycles between these quartiles was not significantly different. However, upper quartile successes had positive correlation with learning, while the lower quartile's had none. Proportionally, the lower quartile also had twice as many off-task failures – per objective, low performers had two far failures, while the upper quartile averaged one. Within similar duration and success counts, the upper quartile got further in the game, had fewer far failures, and had contextual success that supported learning gains. Thus, each group seemed to be using their time very differently, prompting further level-specific investigation.

Table 3: Overall quartile results summary

	Upper Quartile	Lower Quartile
Timestamp	No significant difference.	
# of Objectives Added	7	6
# of Objectives Completed	6	5
# of Far Failures per Objective	1	2
Total # of Successful Cycles	No significant difference.	
Successful Cycles vs. Pre-Post Gains	Significant positive correlation (r = +.377, p =.015)	NO correlation

Objective-level play data exposed telling differences between the quartiles. The first trend was that early far failure was associated with learning losses. Compared to the upper quartile, the lower quartile had twice as many far failures (on average) in Objective 1 of the game. These Objective 1 far failures had a significant negative correlation with learning for the lower quartile (r = -.277). Since each quartile had similar numbers of total failures, these far failure proportions are a stark contrast. Essentially, the lower quartile had more frequent early far failures, which then associated with poor learning outcomes. Conversely, in the second trend, players with the greatest learning gains performed very well in the final objective. The upper quartile had significantly higher Objective 8 successes than the lower quartile (p = .023). Cell cycle success, a skill specifically taught in early levels, was also significantly higher in the final level for the upper quartile (twice as many as the lower quartile; p = .023). These endgame contrasts imply that the top learners' on-task performance in early levels provided the gameplay mastery necessary to excel at the boss level, and demonstrate knowledge of the baseline biology lab practices that underlie core game mechanics.

Table 4: Objective-level quartile results summary

	Upper Quartile	Lower Quartile
# of Obj 1 Far Failures	2	4
Obj 1 Far Failure vs. Pre-Post Gains	NO correlation	Significant negative correlation (r = -.277, p =.040)
# of Total Failures	No significant difference.	
# of Obj 8 Successful Cycles	3	2
# of Obj 8 Successful Cell Cycles	2	1

Overall, four major trends emerged in the data. In the aggregate set, general gameplay progression as well as total gameplay success had a positive relationship to learning. Far failure in tutorial levels of the game was negatively correlated with learning outcomes, and boss-level performance was positively associated with pre-post gains. Quartile analysis reflected these trends, reinforcing that far failure and success in the bookends of the game were crucial differentiation points for learning.

Important implications arise from these data. First, gameplay progression and success seem to effectively support content learning. Secondly, early-game correlations suggest that certain types of failure in context - not failure itself - inform learning. Far failure in the crucially-scaffolded Objective 1 could be a critical indicator of players losing learning due to "gaming the system" (signaled by lack of attention to instructional input). Thirdly, boss-level performance seems to be an effective gauge of overall content knowledge gains. These bear major impact for generalized game design on two counts. In early levels, far-failure-based adaptive feedback may be key in changing off-task players' learning trajectories. Finally, the boss level, designed as an effective indicator of learning content via game mastery, may render pre-post content tests needless – ultimately making *gameplay itself* the only necessary assessment.

Conclusion

The GBA model allowed us to move beyond a simple pre-post comparison of game play to learning outcomes by providing data on how players interacted with the game environment. The design of the semantic template allowed us to collect data at key moments in gameplay; the learning telemetry allowed us to tag and assemble these click-stream data points into play profiles we could use for analysis. The resulting data allowed insight into the role of success and failure in *Progenitor X* game play. As we have seen, games allow players to experiment with failure without real-world consequences. However, the kinds of failures players experience matter. Productive failure (Kapur, 2008) suggests that effective learning environments encourage students to activate prior knowledge as a condition for direct instruction. *Progenitor X* introduces players into an unfamiliar subject matter context (regenerative medicine), but in a familiar game-genre context (puzzle-based videogames). Familiarity with the game-conventions invites players to interact with a system in order to learn programmed

relationships between cells, tissues, tools and cultures. One interpretation of our analysis is that productive failure and success happen when players bridge game-mechanic knowledge to content-model knowledge through gameplay; non-productive failure happens when players ignore the content model and treat *Progenitor X* solely as a colorful puzzle game with zombies. Specific junctures of connection between content learning and in-game performance, shown in this study, have major implications for educational game design. In early gameplay, differentiating non-productive "far" failure in vital tutorial levels may be key in guiding off-task players towards better learning outcomes. In final stages, these data highlight the potential of a well-designed boss level to be a comprehensive, naturalistic, summative assessment of content knowledge. In future game design as well as research, the richness of the data generated by the GBA will allow us to further explore the relations between player interaction and learning.

References

Baker, R. & Yacef, K. (2009). The state of educational data mining: A review and future visions. *Journal of Education Data Mining,* 1(1), 3-17.

Gee, J.P. (2005). Learning by design: good video games as learning machines, *E-learning*, 2(1), pp. 5-16.

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, 26(3), 379-424.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence-Centered Design for Educational Testing. Principled Assessment Designs for Inquiry Technical Report 17. SRI International Center for Technology in Learning. Accessed May 31, 2012 at http://padi.sri.com/downloads/TR17_EMIP.pdf

Owen, V.E., Wills, N., & Halverson, R. (2012). *CyberSTEM: A Game-Based Evidence Model.* Presented at Games+Learning+Society, Madison, WI, June 13-15.

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. 503–523). Charlotte, NC: Information Age.

Acknowledgments

This work was made possible by a grant from the National Science Foundation (DRL-1119383), although the views expressed herein are those of the authors' and do not necessarily represent the funding agency. We would also like to thank the *GLS Center* team, including Kurt Squire, Mike Beall, Ted Lauterbach, Allison Salmon, Kevin Alford, Meagan Rothschild, Shannon Harris, Keari Bell-Gawne, Greg Vaughan, and Nate Wills.