Game-Based Assessment:
An Integrated Model for Capturing Evidence of Learning in Play

V. Elizabeth Owen, Rich Halverson, & Nate Wills, University of Wisconsin-Madison
Email: vowen@wisc.edu, halverson@education.wisc.edu, natewills@gmail.com

Abstract: This paper presents a Game-Based Assessment model (GBA) designed to capture relevant information on play and testing 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 game-play itself can provide a powerful new form of assessment. The GBA model has two layers: 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 between 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 game-play 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 for evidence generation to document learning.

This paper presents a Game-Based Assessment model (GBA) designed to capture data on player learning in the midst of game-play. The GBA model 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.

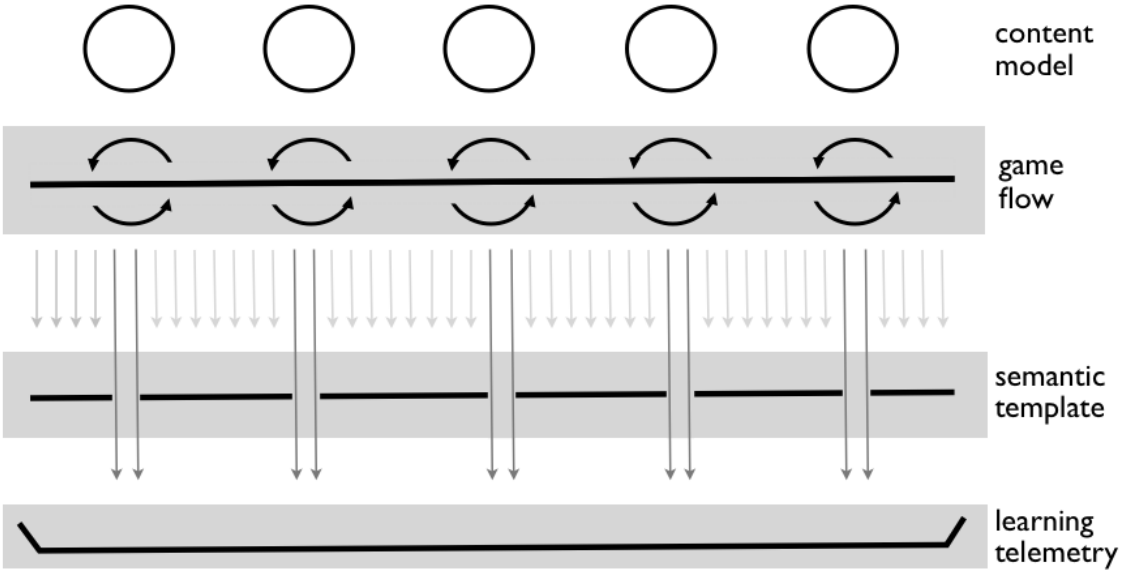


***Figure 1*: Game-Based Assessment Model**

The GBA model is designed to draw significant game-play moves from the game-context. The model has is integrated into an overall 4-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 game-play; 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.

Progenitor X invites players to dissect, collect, cultivate, differentiate and treat diseased tissues via adult epithelial stem cells. 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. Based on the content model above, *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 2: 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.

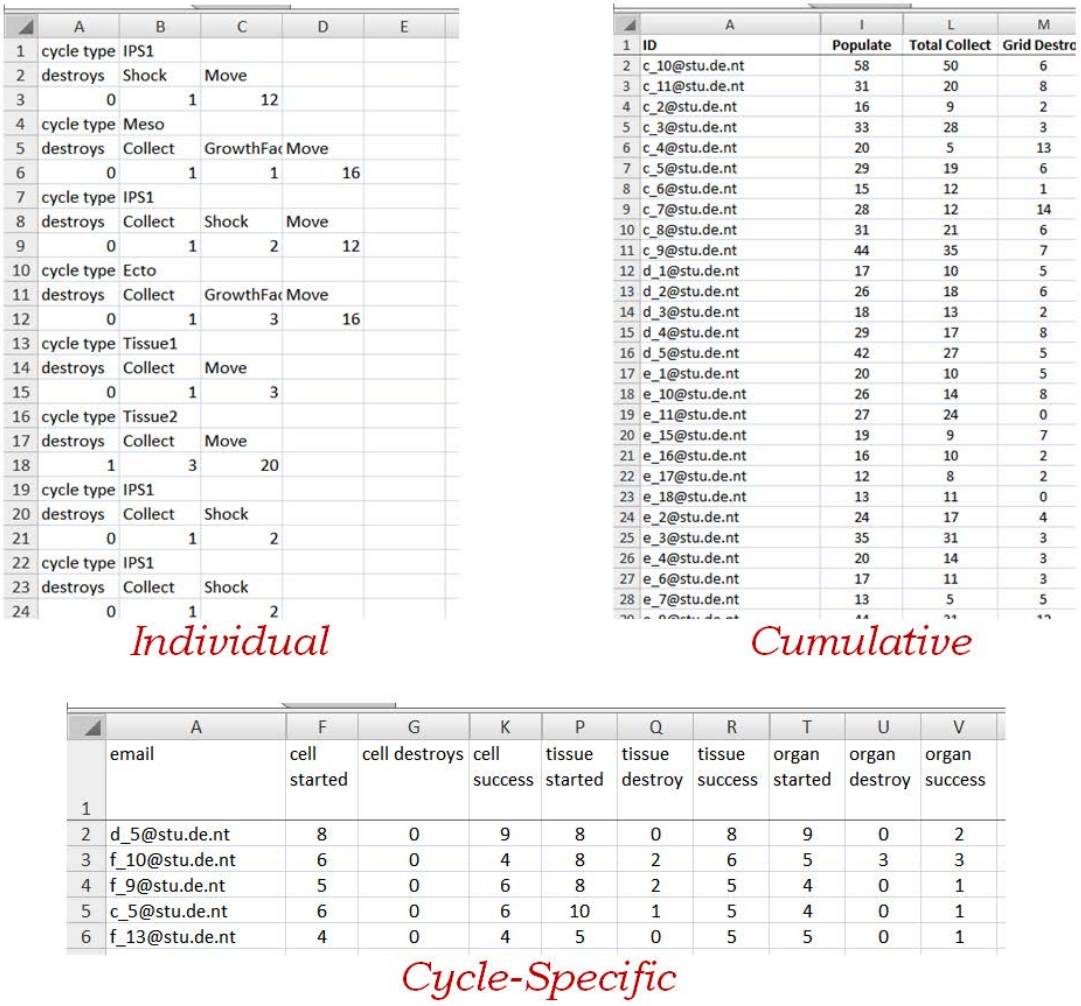
	A	B	C	D	F	G
1	**In-Game Sequence**	**Mission**	**Cycle****	**Cell population type (Stage 1)**	**Treatment tool type (Stage 2)**	**Collection cell type (Stage 3)**
2	1	1	A. ips	fibroblasts	electroporate	ips
3	2	1	B. meso i	ips	growth factor	meso
4	3	1	A. ips ii / C. ecto i	fibroblasts	electroporate	ips
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
16	15	3	E. tissue v	meso	N/A	meso / tissue
17			**(A = ips, B-D = diff cells, E = tissue, F = organs)			

Figure 3:* Detailed Semantic Template 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.



	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			
17	destroys	Collect	Move		
18	1	3	20		
19	cycle type	IPS1			
20	destroys	Collect	Shock		
21	0	1	2		
22	cycle type	IPS1			
23	destroys	Collect	Shock		
24	0	1	2		

Individual

	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
18	e_10@stu.de.nt	26	14	8
19	e_11@stu.de.nt	27	24	0
20	e_15@stu.de.nt	19	9	7
21	e_16@stu.de.nt	16	10	2
22	e_17@stu.de.nt	12	8	2
23	e_18@stu.de.nt	13	11	0
24	e_2@stu.de.nt	24	17	4
25	e_3@stu.de.nt	35	31	3
26	e_4@stu.de.nt	20	14	3
27	e_6@stu.de.nt	17	11	3
28	e_7@stu.de.nt	13	5	5

Cumulative

	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

Cycle-Specific

Figure 4: **Learning Telemetry - Processed Telemetric 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. For example, if a player successfully collected the required number of cells in a cycle 2 times, and tried to complete the cycle 5 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 stem cell 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
Aggregate results revealed intriguing reasons to look further into the "black box" of the game. First, with an 11% average increase in pre-post content scores, the game seemed to be a noteworthy learning vehicle. Interestingly, the aggregate efficiency ratios told us little about learning outcomes as measured by the post-test. Only in the last organ cycle of the game (the "boss level") was the efficiency ratio correlated with pre-post gains (r = .3219). Thus, by the end mission of the game, being good at the mechanics was associated with learning the content model. However, we were unable to identify overall game mastery (as measured by player efficiency ratio) with content learning (measured by pre-post tests) elsewhere in game-play data. This led us to investigate what was going on with players within the specific game cycles.

	Pre-Post Gains
Total Gameplay	11% average increase (t-test sig = .0098)
Efficiency Ratio	no significant correlation
Boss-level Efficiency Ratio	significant positive correlation (r = .3219)

***Table 1*: Aggregate *Progenitor X* Data Summary** n=39, α = .10

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, 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 (see red X in *figure 5*). 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.

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 5*). We grouped 3 possible player outcomes: correct collection (successful); correct set-up but health runs out (near failure); incorrect setup and/or incorrect collection (far failure).

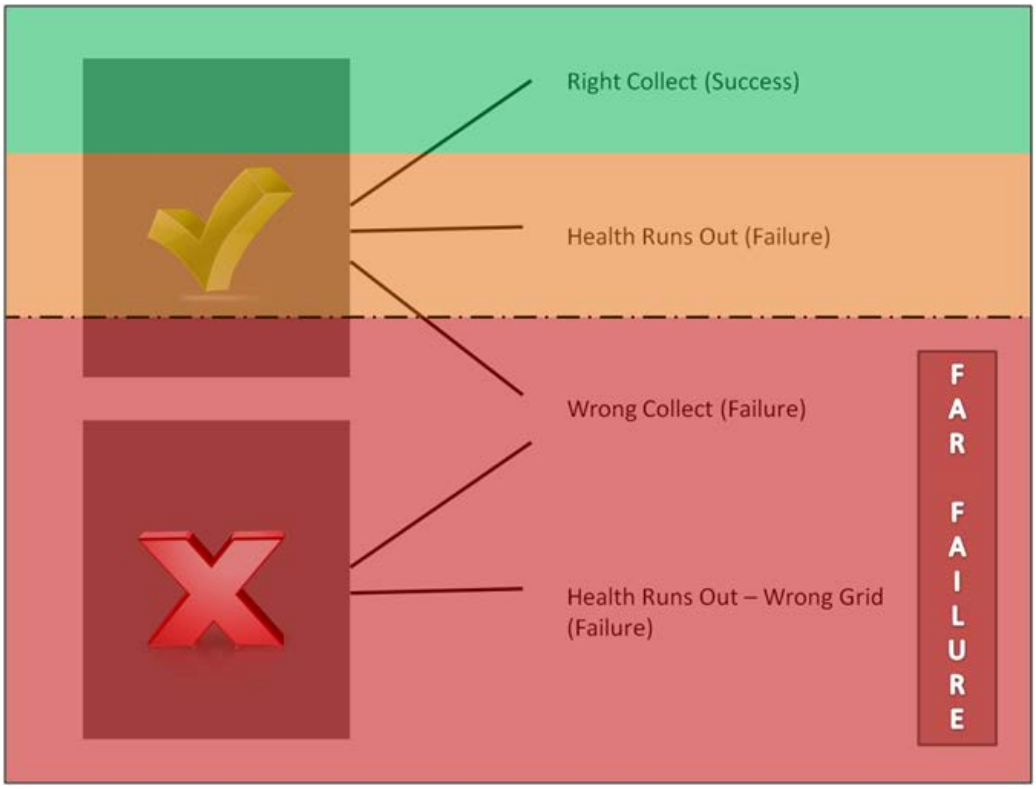


Figure 5: **"Far Failure" In *Progenitor* Gamespace**

The analysis of far failure gave considerable insight into the player data. We found that the total number of "far failures" for players *across all cycles* was significantly negatively correlated with learning as measured by the pre-post tests (r = -.2788). Other indicators of play, including the number of cycles started, number of successful collects, and total cycles completed had no correlation with pre-post gains.

To deepen our understanding of far failure, we divided students into quartiles according to pre-post change; the upper quartile had the largest gains in content question scores, while the lower quartile had the smallest. On average, lower quartile students had 7 cycles of "far failure," while upper quartile students only had 2.3 (difference significant at α = .1). Concerning the number of "far failures" and pre-post change, the top students' were positively correlated, while the lower quartile had strong negative correlation. Since both groups had comparable total NUMBERS of failures, the lower quartile had a greater proportion of "far failures;" thus, the latter's losses in learning the content may be linked to their lack of responsiveness to the game cues. The pre-post correlation with this number suggests that certain types of failure, not failure itself, inform learning.

	Upper Quartile	Lower Quartile
# of average "far" failure cycles (p=.0768)	2.3	7
	Sig. + correlation (r = .5796) with pre-post	Sig. - correlation (r = -.9408) with pre-post
Total failures	No significant difference.	No significant difference.

Table 2: **Summary of Quartile Findings** * n=20, α = .10

Conclusion and Significance

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 game-play; 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 failure in *Progenitor X* game play. Games allow players to experiment with failure without real-world consequences. However, the kinds of failures players experience matter. Productive failure (Kapur, 2008, 2012) 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 happens when players bridge game-mechanic knowledge to content-model knowledge through game-play; non-productive failure happens when players ignore the content model and treat *Progenitor X* solely as a colorful puzzle game with zombies. 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*
Educational 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

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, 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 ERIA Interactive team, including Kurt Squire, Ben Shapiro, Mike Beall, Ted Lauterbach, Meagan Rothschild, and Shannon Harris.