

**Capturing In-Game Learner Trajectories with ADAGE (Assessment Data
Aggregator for Game Environments): A Cross-Method Analysis**

By

V. Elizabeth Owen

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Curriculum & Instruction)

at the

UNIVERSITY OF WISCONSIN-MADISON

2014

Date of final oral examination: 5/28/14

This dissertation is approved by the following members of the Final Oral Committee:

Constance Ann Steinkuehler, Associate Professor, Digital Media

Matthew Berland, Assistant Professor, Digital Media

Kurt Squire, Professor, Digital Media

Leema Berland, Assistant Professor, Disciplinary Studies – Science Education

Ryan Baker, Associate Professor, Cognitive Studies (Teachers College at Columbia)

© Copyright by V. Elizabeth Owen 2014
All Rights Reserved

For my family, in its many forms.

Acknowledgements

During the past several years in Madison, GLS has become my home in more ways than one. I want to deeply, deeply thank my advisor Dr. Constance Steinkuehler for seeing potential in me years ago and taking me into the fold. She has been a truly inspiring figure for me personally and professionally – her intelligence, courage, passion and remarkable presence have encouraged me to hold my head up and keep moving through more than a few tough periods. Dr. Richard Halverson has also been a close mentor, and I am truly thankful for his unwavering guidance and support. He, along with Dr. Kurt Squire, invited me into the CyberSTEM project to begin with, where the idea for ADAGE came into being. From the beginning, all three of these GLS leaders have been incredibly generous and supportive, and I am truly thankful. Also within the GLS core, Jim Gee has been a mentor from inception, and I would not be here without his launching guidance and support.

Allison Salmon is the brilliant programmer who made ADAGE as an API possible at all, and I still want to grow up and be like her someday. Thank you, Allison, for bringing ADAGE into real API existence, and for building it to the amazing place it is now. Thank you also to Dr. Ben Shapiro, who was a passionate member of the first CyberSTEM team and really helped open discourse about click-stream assessment possibilities from a very early stage. (This CyberSTEM work of the last few years has been made possible by a grant from the National Science Foundation (DRL-1119383).) Dr. Matthew Berland has been an incredible gift in this whole process, coming into the picture just as things were starting to get very hairy with the question...”so what do we do with all this data?” He has been a vital resource for GLS in delving more deeply into learning analytics and data mining, and his deep intelligence and expertise is absolutely fundamental for the trajectory of this work. Thank you, Dr. Berland. Dr. Ryan Baker, similarly, has given his invaluable expertise and time to this ADAGE research, and I thank him from the bottom of my data nerd heart! I feel incredibly lucky to have had such a helpful mentor and accomplished academic as an outside committee member. Dr. Leema Berland has also very generously given her time and considerable wisdom to the work, for which I am very appreciative. The entire GLS community has been amazing – Dennis, who was there from inception, Amanda, Shannon, Ryan, Meagan, and the whole gaggle of people I have commiserated with since I was a wet-behind-the-ears first-year grad student. I’m thankful for every one of you.

Last but certainly not least, there are many people in my own family – very close to my heart – without whom I simply would not have made it. Most of all, thank you to my incredible partner Ryan, and to my unconditionally loving parents, and to my dear brother (I finally forgive you for ripping the head off my Barbie doll 30 years ago. After all, you did help me proofread this dissertation, which should be penance enough.) :D

Table of Contents

Abstract	iv
Chapter One: Introduction and Overview	1
Chapter Two: Theoretical Framework	17
Games: Rich Microworlds for Learning and Assessment.....	17
Evidence Centered Design	20
Educational Data Mining	33
Chapter Three: ADAGE and <i>Progenitor X</i>	46
Chapter Four: Success and Shades of Failure – Feature Engineering and Applied Statistics.....	56
Chapter Five: Markov Modeling of Learner Progression Through the <i>Progenitor</i> Gamespace .	80
Chapter Six: Experimentation and Learning – Predictive Modeling with Detectors.....	109
Chapter Seven: Conclusions and Future Work	147
References.....	163
Appendices.....	189

Abstract

In learning and making meaning in this digital age (Steinkuehler, Barab & Squire, 2012), computer-based microworlds can be immersive contexts for supporting self-regulated learning (Papert, 1980). Games represent an important subset of microworlds, able to enhance player agency and endogenous narrative in learner-adaptive play (Rieber, 1996). Games offer pleasantly frustrating, well-ordered problems with timely scaffolding – thus providing new opportunities for assessment of complex skills in an authentic context (Gee, 2012). However, game-based assessment can be a challenge; instead of a few isolated, independent assessment points, evidence of learning is often manifested in a rich, electronic data stream of continual player interactions (Shute, 2011).

Game-based assessment thus needs to isolate specific, game-situated task performance – yet account for masses of context-rich, event-stream interaction data central to play narrative. Uniting these paradigms in an integrated assessment framework, the Games+Learning+Society (GLS) group has created ADAGE (Assessment Data Aggregator for Game Environments), a framework designed to transform click-stream data into evidence of learning. ADAGE integrates core game structures into a click-stream data schema, which is seeded with context vital to informing learning analyses (Owen & Halverson, 2013). Overall, it provides a rich, method-agnostic data yield, with scalability and cross-genre flexibility. ADAGE development has been guided by recent learning assessment research in Evidence Centered Design (Mislevy, 2011) and Educational Data Mining (“EDM”; Baker & Yacef, 2009).

This dissertation establishes ADAGE as an assessment data framework for learning games; empirically, it then investigates ADAGE-generated performance data to assess learner trajectories in a biology videogame. The overarching research question asks: what kinds of

organic player interactions (including play progression, in-game success, shades of failure, and experimentation) characterize learning? Three analyses (using statistics, machine learning, and EDM) investigate relationships between learning and 1) in-game success/failure; 2) core play progression; and 3) player experimentation. Ultimately, findings differentiate types of failure, reveal experimentation patterns, and demonstrate the positive relationship between strategic failure and learning. These ADAGE-based organic play trajectories have powerful implications for defining alternate learner pathways in new assessment paradigms, reconsidering the role of failure in formal learning evaluation, and informing iterative game design for the optimization of learner-adaptive play.

Chapter One: Introduction and Overview

Seymour Papert maintained three decades ago that “computers can be carriers of powerful ideas...they can help people form new relationships with knowledge” through “exceptionally rich and sophisticated micro-world[s].” Specifically, one “design criterion for our microworlds is the possibility of...games...that make activity in the microworlds matter” (Papert, 1980, p.4, 12, 126). More than ever, in our current technological era, these computer-based “video games have the potential to lead to active and critical learning” (Gee, 2003, p.46). However, Squire cautions, “games aren’t just open environments; they are carefully crafted learning experiences” (Squire, 2011, p.13). In other words, design matters. Indeed, good games encompass pleasantly frustrating, well-ordered problems (Gee, 2005) which reward higher-order thinking skills (c.f. Steinkuehler & Duncan, 2008; Halverson et al., 2011) – and provide just-in-time information in formative feedback cycles (Gee, 2003; Shute, 2011). Ongoing assessment thus becomes a vital component of maintaining the agency and endogenous motivation (Costikyan, 2002) in the designed experience of good games (Squire, 2006). It is also vital in leveraging interaction-rich game data for understanding of learning in the process of play, rather than simply seeing the game as a black box between pre- and post- measurements. However, game-based assessment of any kind can be challenge, since instead of offering a few isolated, independent assessment points, evidence of learning is often manifested in a rich, electronic data stream of continual player interactions (Shute, 2011).

Thus, game-based assessment needs to isolate specific, game-situated task performance – yet simultaneously account for masses of context-rich, event-stream interaction data central to play narrative. Uniting these paradigms in an integrated game-based assessment framework, the Games+Learning+Society group has created ADAGE (Assessment Data Aggregator for Game

Environments). ADAGE (Owen & Halverson, 2013) is an assessment data framework designed to turn click-stream data into evidence for learning. It integrates core game design structures into a click-stream data (telemetry) schema, which is then seeded with context vital to informing learning analyses. Overall, ADAGE provides a standardized game telemetry framework with a rich, method-agnostic data yield, efficient enough to have scalability, and flexible enough to use across games. In current development, ADAGE is both a game-based assessment data framework and an API with a data output engine. The ADAGE design and development effort has been especially guided by recent prominent research in measuring learning in digital environments: Evidence Centered Design (e.g. Mislevy & Haertel, 2006) and Educational Data Mining (e.g. Romero & Ventura, 2010; Baker & Yacef, 2009).

An ADAGE-based empirical study, this dissertation endeavors to better understand learning in the midst of play – a natural conductor for interest-driven, self-regulated exploration of knowledge (Vygotsky, 1930-1934/1978; Rieber, 1996). Specifically, this research establishes ADAGE as an assessment data framework for learning games; empirically, it then investigates ADAGE-generated authentic performance data to assess organic learner trajectories in the GLS biology game *Progenitor X*. In this application of ADAGE, interaction data informs three interlinked, cross-method analyses exploring the relationship between in-game performance, experimentation, and learning. Each analysis examines interlocking lenses of learning games as designed experience, grounded in defining characteristics of game microworlds. The overarching research question asks: what kinds of naturalistic player interaction with the educational gamespace (including play progression, in-game success, shades of failure, and experimentation) characterize learning? This question is central to understanding play experience in relationship to

learning – and thus, to harness the power of play in optimizing core design and learner-adaptive mechanics in future designed experience of educational games (c.f. Squire, 2006).

The remainder of Chapter One discusses this empirical research arc in detail, beginning with theoretical foundations of games as learning systems, and distinct microworld lenses of designed experience. Corresponding to each lens, the three analyses of this study (using statistical, machine learning, and educational data mining methods) are then described closely. Respectively, their research questions ask: 1) What is the relationship between learning and in-game success/failure? 2) What play progression patterns characterize learning? 3) What is the impact of player experimentation on in-game performance and, ultimately, learning?

Learning Games as Interactive Microworlds

Kurt Squire asserts that videogames offer “designed experiences” in which participants learn through “being” and “acting” within the gameworld (2006, p.19, p.22), an *in-situ* learning context in which there is no separation between knowing and doing (Brown, Collins, & Duguid, 1989). These designed learning realms support learner activity with built-in principles like just-in-time information and cycles of expertise (Gee, 2005a), and thus carry the embedded scaffolding characteristic of microworlds (c.f. Papert, 1980; Rieber, 1992). Modeling a “system or domain for the user,” microworlds by definition support “self-regulated learning” (Rieber, 1996, pp. 46-47; Zimmerman, 1989) through creating intrinsic motivation for learning in a relevant context – and in game form, provide a system of well-ordered problems which leverage player agency (Gee, 2003; Squire, 2011). Inherently, microworlds as games require behavioral action to progress (Rieber, 1996), manifesting gamespace “cognition as interaction” (Steinkuehler, 2004, p. 522) and interaction as an “authentic performance” measure in a situated context (Derry & Steinkuehler, 2003, p. 802; Boaler & Greeno, 2000). Game microworlds – thus

characterized by designed systems which scaffold self-regulated, intrinsically motivated learning – are important examples of “interactive learning environments where structure and motivation are optimized without subverting personal discovery” (Rieber, 1996, p. 44).

Interaction as vehicle for agency and learning is therefore a central theme in the study of context-rich learning worlds (Rieber, 1992; Greeno, 2005), including videogames as designed experience (Squire, 2006; Steinkuehler et al., 2012). As assessment, interaction data is a vital component of evaluating authentic performance in context from a situative perspective (Derry & Steinkuehler, 2003), reflected in the action-performance emphasis in games as learning context (Gee, 2012; Shute, 2011). This study’s learning theory is thus rooted in situated cognition in the authentic learning context of game microworlds.

Game Microworlds as Designed Experience: Three Lenses

ADAGE, the assessment framework of this research, is designed to structure and capture this critical in-game player interaction. These captured interactions enable study of the three defining elements of learning games as microworlds: games as systems of interaction (a play-based medium), games as scaffolded instruction (for educational content delivery), and games as intrinsic motivators (an endogenously engaging, player-directed experience). Figure 1 illustrates these three overlapping components below.

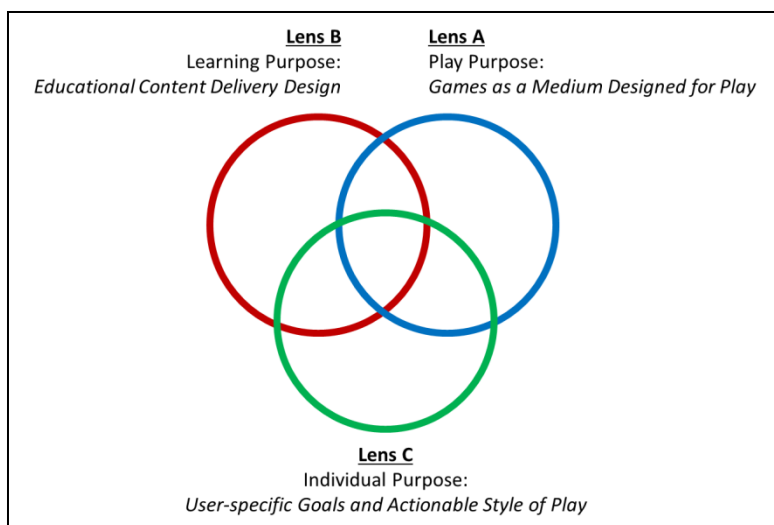


Figure 1. Three lenses of game microworlds as designed experience.

These three converging elements represent vital parts of the learning game experience. Figure 1 maps the intersection of each, visualizing three “lenses of designed experience” for educational games. Together, these lenses represent the tension between the game medium, educational content, and user-specific play goals that define a learning game experience – and which are so artfully balanced in good game design.

- *Play Purpose: Games as a Medium Designed for Play* is Lens “A”, representing the nature of games, designed (whether explicitly for learning or not) to offer a playful, fun experience. Games provide roles, goals, and agency, often engaging the players in a narrative and challenging them to discover an underlying rule system through play (Norton, 2006; Squire, 2011). In contrast to cognitive tutors, this results in a medium where one “right” answer is not always the goal, and discovery through play is an implied norm. Play purpose is originally manifested on the design side, by the development team building the foundations of a designed play experience.

- *Learning Purpose: Educational Content Delivery* as Lens “B” represents the learning game as a vehicle for delivering instruction. Content can range from domain-specific procedural and declarative knowledge (e.g. biology lab processes) to soft skills (e.g. empathetic social interactions). This lens considers content knowledge translation into specific verbs of play.
- *Individual Purpose: User-specific Goals* constitutes Lens “C”, representing gaming style and subjective play goals specific to each player. This lens focuses on player intent, manifested in the kinds of in-game behavior tendencies each user brings to the game. For example, one player may want to learn solely through experimentation, and engage with a game by immediately testing boundaries rather than strictly adhering to tutorial cues. Conversely, another player may wish to avoid failure and play “conservatively”. This might include following the game instructions to a tee, interacting exactly as the cues lead, and finishing the game with zero failure. (These patterns, of course, vary greatly by game and player, and can be characterized many different ways; play typologies for this particular study are detailed in Chapter Six).

Three Lenses, Three Analyses: An Empirical Arc

This research endeavors, through study of in-game interaction data, to capture learner trajectories of success, failure, and experimentation through each lens of designed experience (Figure 2). Each analysis, then, targets the investigation of one unique intersection of the three Lenses above. The first analysis, an evaluation of player performance on content-based verbs of play, focuses on the intersection between Lens B (content-centered) and Lens C (player-centered). The second analysis, tracking play progression and player attrition throughout the

course of the game, connects Lens A (focused on an trajectory of play) and Lens C. The last analysis is a synthesis of all three lenses. Through EDM methods, it examines player experimentation (Lens C) throughout game progression (Lens A) in relationship to shades of in-game failure and learning outcomes (Lens B). The analyses are described below, with the unifying Lens, research question, and corresponding method for each.

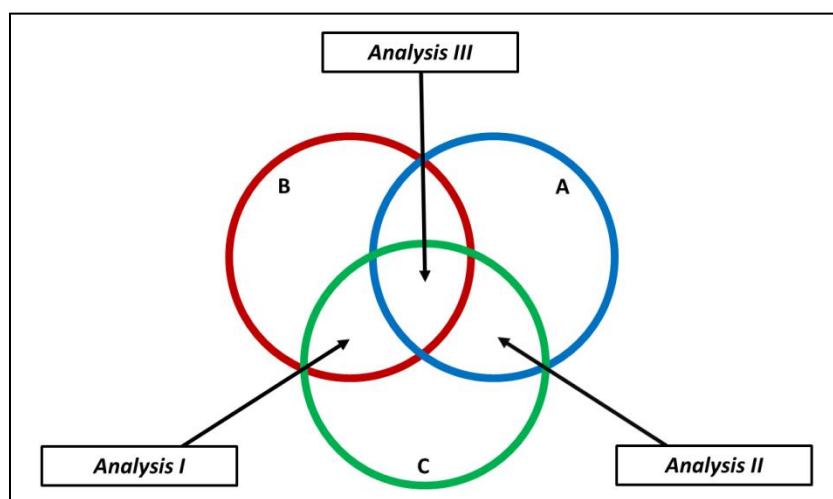


Figure 2. Analyses I, II, and III as situated in lenses of designed experience.

Analysis I: In-game Performance and Learning - Feature Engineering and Applied Statistics

The first analysis section uses feature engineering and applied statistics to connect context-specific performance trends and learning outcomes. This analysis rests at the intersection of Lens B and Lens C, investigating player choices in content-based performance (Figure 2). The goal is to get a non-reductive sense of nuanced success and failure, as marked by learning gains, across the full interactive landscape of the game. An extension of an earlier study showing statistical significance of failure and success constructs with learning outcomes (Owen et al., 2013), it uses informed, iterative feature engineering to more deeply examine these interactions.

Specifically, its research question asks how fine-grained, chronological performance data (including shades of failure and success) connect to learning outcomes.

Feature engineering, descriptive statistics, and nonparametric statistics are used to investigate patterns in this research vein. Feature engineering provides new telemetry indices with which to better understand the landscape of game performance, while statistics help connect the shape of those features to learning outcomes. First, feature engineering takes four base types of success and failure, and systematically applies six computational lenses to each to produce new fine-grained, objective-specific telemetry indices central to the research question. Then, these features are visualized through descriptive and nonparametric statistics, chosen because of the non-normal distribution of the data. Specifically, representation of these data in descriptive time series line graphs launch deeper investigation of sequential trends, and inform the use of bar graphs, scatterplots, and two-sample Wilcoxon tests to contrast learner groups. To corroborate these contrasts, Spearman's correlation is used on relevant features in relationship to learning outcomes. Figure 3 shows the flow of analysis for each new data feature.

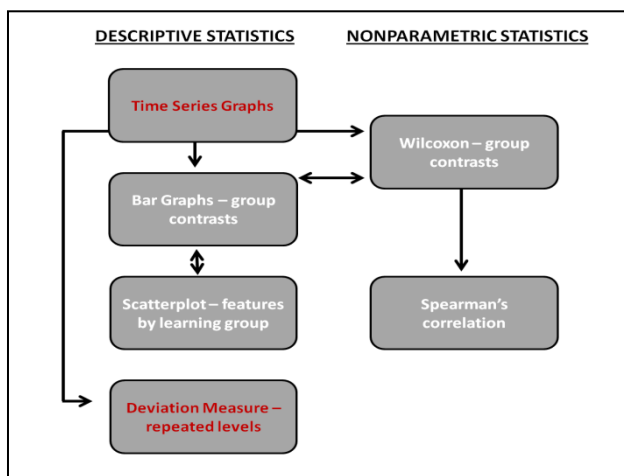


Figure 3. Analysis I flow of statistical visualization for feature-learning connection.

The visualized trends with the strongest relationship to learning, as corroborated by the nonparametric tests, shape final findings along three themes of performance and learning: frustrated failure, success in convergent-task game levels, and failure as learning strategy. Analysis I results provide a diverse range of features which feed analyses II and III, as well as highlight salient themes for further investigation in these subsequent analysis sections.

Analysis II: Understanding Play Trajectories: Markov Modeling of Learner Groups

Analysis two focuses on mapping learner navigation of the gamespace using machine learning methods. This analysis represents the overlay (Figure 2) between designed experience Lens A (game as a progressive play arc) and Lens C (player choice). Capturing basic player choices in the context of full game progression can give insight on interactions more characteristic of learning. To do so, this study traced play progression from level to level, visualizing whether players repeat a given cycle, move on to the next level, or quit. A sequential probability model was used, because it has the ability to illustrate the probability of players moving, in time order, from one level to another. Specifically, a first order Markov model was used as the probability model. Two Markov chain models were made, one for the upper quartile and one for the lower quartile of learners, as measured by learning gains on *Progenitor's* pre-post test of regenerative biology (see Chapter Three for pre-post detail). Contrasting the two models of play directly addressed the second research question: how does organic play progression differ between groups of learners?

Contrasting probabilities (of repeating, moving forward, or quitting) for each learner group gave insights into patterns of play most characteristic of learning gains. Each progression was illustrated in a Markov model, with quit states and each objective shown (Figure 4). A transition matrix, detailing the probabilities for each group in moving from one state to the other,

was then generated. Through comparing these probabilities between groups, play progression patterns characteristic of learning emerged along themes consistent with Analysis I: early game failure, mid-game scaffold-and-fade performance, and endgame strategic navigation.

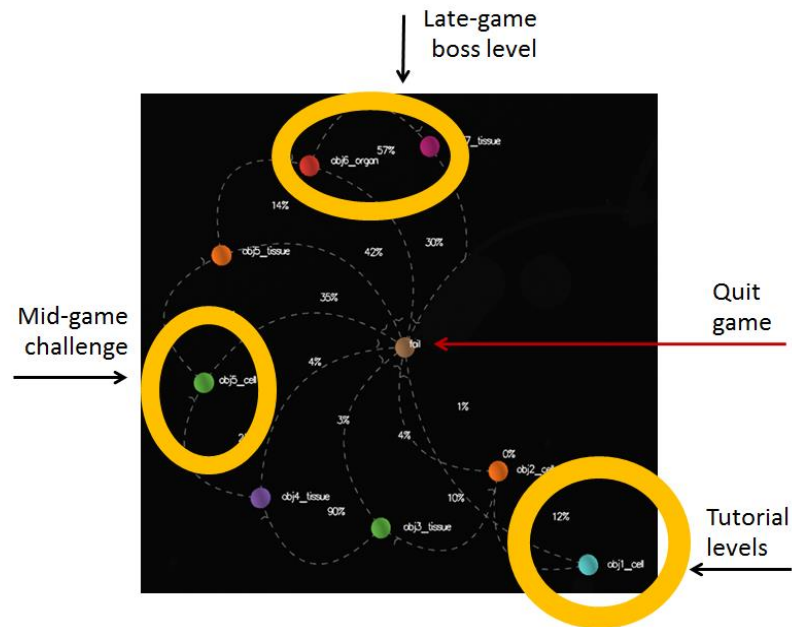


Figure 4. Example Markov model created in NetLogo (Wilensky, 1999) using the Narkov algorithm (Berland, 2012).

Analysis III: Experimentation and Learning: Predictive Modeling with Detectors

This data mining analysis builds on the previous studies, using performance and progression features to make inferences about learner behavior in the gamespace. This analysis represents a synthesis of all three designed experience lenses: Lens A (game as play-based medium), Lens B (game as content delivery system), and Lens C (learner-specific specific play goals and experimentation). The core research question for analysis three is: What play data features characterize experimentation in *Progenitor X*, and how does this behavior predict learning outcomes? Experimentation can be indicative of transgressive play, a natural element of

the game medium (Salen & Zimmerman, 2004) in which players may interact with the gameworld in ways unanticipated by the designers. To explore this construct, this study draws on educational data mining to build a detector of experimentation, and then connects the related player behavior with learning outcomes.

A detector is an automated model that can detect student behavior from log file data (e.g. Baker, Corbett & Koedinger, 2004). Here, it was built through holistic coding of “text replays” (Baker & de Carvalho, 2008), a series of player actions displayed in a snapshot, which was then evaluated for levels of player experimentation. Data features predictive of experimentation were then determined (using classifier algorithms J48, Naïve Bayes, and JRip) with RapidMiner¹ or WEKA data mining software (Hall et al., 2009). These detectors of experimentation in play were next investigated in relationship to learning, used as input variables in a predictive modeling of pre-post learning outcomes (using RepTree and M5’ algorithms).

Thus, detector building supports the leap from interaction data to behavioral inference about experimentation; next, the input of experimentation features into a model with predicted learning outcomes helps illuminate the relationship between play exploration and learning. These analyses yielded extremely interesting results along thematic findings of Analysis I and II, including the critical role of early failure in learning, shades of failure evolving in learning impact over time, and late-game strategic failure in relationship to learning gains (see Chapter Six for extensive results).

Study Assumptions and Limitations

In establishing ADAGE and using its data yield as game-based assessment, three main assumptions are made. Each are described as follows: 1) the learning by doing of games (Squire,

¹ <http://rapidminer.com/>

2006) can improve skills and knowledge; 2) different kinds of learning and learner characteristics can be measured during gameplay (Shute, 2011); and 3) ongoing player learning can be supported with formative feedback (Gee, 2003).

Because ADAGE is an API overlay of the core game programming, it is inherently limited to collecting in-game interaction data. However, its GLS creators consider it just one cornerstone of a much larger ecology of synchronous interaction data – e.g. sensor data, video logs, observation, and interviews (Owen et al., 2014; Halverson et al., in press). Triangulation of ADAGE log files with other forms of player interaction data can constitute powerful mixed-methods research, and is already being pursued at the center (e.g. Halterman et al., in submission; Beall et al., 2013). In a related constraint, because this particular study analyzes in-game data from the *Progenitor X* (a single-player game), it uses only player-game interaction data. Thus, this particular analysis does not analyze in-game social elements, because there are none built into *Progenitor X* core mechanics. In-game social interaction, however, is an area of great interest to GLS and is currently being built into ADAGE for the new multi-player game *Trails Forward*². In terms of studying out-of-game social discourse and artifacts, the center has plans to integrate ADAGE data and visualizations into a larger online interface which can serve as a community forum as well as a user portal for students, facilitators, and researchers. As these portals connected to ADAGE are developed, and multi-player telemetry capacity grows, it can better converge with study of the larger learning big “Game” context (Gee, 2012; Steinkuehler, 2004), and contribute to comprehensive assessment methods in situating big data.

For the example studies of ADAGE in *Progenitor X*, a main delimitation was locale-based data collection. Since our target audience for *Progenitor X* was 8th grade, IRB research guidelines necessitated parental/guardian consent forms signed by hand and presented in person.

² <http://gameslearningsociety.org/blog/?p=105>

This yielded a pool of subjects from the upper Midwest area (delimiting regional generalizability) and somewhat limited our sample size (n=110). However, the sheer volume of clickstream data offset this somewhat, yielding roughly 10,000 data points per player and over 1 million data points total. Ultimately, this research is largely concerned with putting forward methods for structuring and analyzing game-based assessment, and this dissertation's empirical study example represents just one framing of the enormous clickstream data yield possible with frameworks like ADAGE. However, in future studies, a much broader population sampling may be possible, since the GLS center just obtained IRB approval for remote data collection. This opens data collection for thousands of users – anyone who plays GLS games on the internet – across the globe, and could support very broad generalizability and large sample sizes of future analyses.

Implications of Adage and Game-Based Assessment

Maximizing learner engagement and support through good design is fundamental to fully leveraging games as a learning vehicle (Gee, 2003, Shute, 2011). One significant benefit of telemetry-based assessment is its ability to play a key role in optimizing learner-adaptive play experience in this iterative design process. Telemetry-based insights can support three development stages: core game creation, alpha user testing, and final-stage adaptive play design. During early design phases, building distinct mechanics of play which will further the narrative, teach the content, and provide moments of assessment is vital to designing an engaging, effective learning game (Asbell-Clarke, 2013; Plass et al., 2012). For example, mapping ADAGE structures of formative assessment (detailed in Chapter Three) to early core design efforts can inform the creation of play mechanics specifically designed to provide evidence of learning. In user-testing alpha and early beta phases, ADAGE-based visualizations and descriptive analytics

can be particularly helpful in refining UI design, as well as identifying bugs and player attrition points (e.g. Beall et al., 2013). After extensive post-beta playtesting, learning analytics can be used to predict in-game actions and performance most characteristic of learning. This knowledge of ideal player behavior can then inform the final design phase: user-adaptive, fully scaffolded play for optimized in-game learning (Owen, 2014). To this end, optimal player actions, sequential pathways, and assessment growth trajectories can each be defined through learning analytics (including visualization, prediction, and pattern mining methods). A layer of game mechanics – either enhanced in-game visual cues, for example, or an agent-based hint system – can then be built to help guide players toward pathways optimized for learning and engagement (Owen & Ramirez, in submission). In this dissertation, for example, insights into failure and transgressive play can inform the in-game highlight of unanticipated learning pathways while supporting experimental play. Overall, in-game interaction data and assessment structures can enable analytic insight vital to an optimized learning game experience.

Event-stream assessment frameworks like ADAGE can have impact both inside and *outside* the game experience. Game-based embedded assessment is a powerful tool able to capture authentic performance *not* decontextualized from an engaging learning environment (c.f. (Derry & Steinkuehler, 2003; Shank, 2011). Capturing in-game interaction can be a “quiet, yet powerful process by which learning performance data are gathered during the course of playing” (Shute, 2011, p. 505). Standardized, game-tailored assessment data frameworks like ADAGE – supporting cross-genre game application and multiple approaches to analysis – represent the possibility for large-scale implementation of authentic performance assessment embedded in engaging learning worlds. National testing giants like ACT and ETS have been increasingly involved in game-based assessment research (e.g. Institute of Play, 2013; Encarnacao, 2014).

Top research and learning game labs like MIT Learnlab, Pearson, TERC, Filament games, Vanderbilt, and the Columbia Teachers' College Educational Data Mining lab are currently collaborating with GLS on future iterations and wide-scale implementations of the ADAGE framework (supported by NSF Award SMA-1338508). Even at a federal government level, there has been recent advocacy for national use of digital learning worlds in education, including serious games – and funded development of corresponding digital assessment methods. The President's Council of Advisors on Science and Technology, linked with the government Office of Science and Technology Policy (OSTP), recently advised in an official presidential report:

[The Department of Education]...should provide robust and diversified support for...R&D that will lay the foundation for educational technologies such as personalized electronic tutors, serious games and interactive environments for education. (President's Council of Advisors on Science and Technology, 2013, p.28)

As part of this recommendation, the Council also advised developing “assessment programs for those technologies that use advanced techniques from ‘big data’ R&D and from the learning sciences” (PCAST, 2013, p.13). Concurrently, the placement of game-specialized digital media advisors in the White House OSTP (including Dr. Constance Steinkuehler in 2011-2012, and Mark DeLoura presently) supports this high-level trend towards understanding and leveraging digital games for learning and assessment. As implied by the direction of nationally-impacting assessment companies, tier one academic consortiums, and government-level advocacy, this forward movement in the field could signal the beginning of a paradigm shift in digitally-based assessment practices on a national scale.

Summary and Content Structure

In respecting educational games as a medium that sets roles and goals (Squire, 2011) in a narrative-based, endogenously motivating context (Costikyan, 2002) – and thus encourages exploration and transgressive play (Salen & Zimmerman, 2004) – it's vital to understand naturalistic learner interaction with the gamespace that can represent authentic play experience. The three analyses of this research, grounded in each lens of game microworlds as designed experience (Figure 2), explore this intersection of play and learning through data furnished by the ADAGE assessment framework.

The following pages of this dissertation will detail the literature base and conceptual framework of ADAGE, as well as describe in detail the empirical methods, results, and findings of all three interlinked ADAGE-based *Progenitor* analyses. Chapter Two provides a broad ADAGE literature base in games for assessment, Evidence Centered Design, and Educational Data Mining. Chapter Three describes the ADAGE framework itself, its application to *Progenitor X*, and the corresponding data collection for this analysis trio. Chapters 4, 5, and 6 are detailed accounts of analysis I, II, and III (respectively). Chapter Seven is the last chapter of the dissertation, providing summary of findings and conclusions about the work as a study of game-based assessment in the digital age.

Chapter Two: Theoretical Framework

Chapter Overview

In creating a framework for assessment of learning through play in educational games, work like ADAGE represents a movement advancing alternative forms of assessment – which may lead to paradigm shifts in the way the education system thinks about measurement of learning. ADAGE supports this innovation in providing a framework to collect mass clickstream data in games, affords a standardized way to organize these data that makes sense across games; and connects these context-rich assessment features with game-tailored methods designed to handle large, unsupervised log file data. One of the reasons this work is so important – and so messy – is that games can provide extremely rich, engaging learning environments in which both learning and assessment are seamlessly integrated into the fabric of the gameworld. This section will review current literature on the value of games for learning and assessment, and follow with a review of two prominent – and very different – approaches of assessing digital streams of learning data: Evidence Centered Design, and Educational Data Mining.

Games: Rich Microworlds for Learning and Assessment

Videogames have distinctive characteristics that make them rich, complex vehicles for learning and assessment. Kurt Squire asserts that “games differ from simulations in that they give roles, goals, and agency”, and enable “transgressive play” (Squire, 2011), p. 29; (Salen & Zimmerman, 2004). Val Shute adds that games are comprised of “conflict or challenge,” “rules of engagement,” and “compelling story and representations” (2011, p. 507). These crucial design elements merge to create a dynamic of endogenous, engaging interaction (Costikyan, 2002).

Well-designed games are examples of situated learning environments in which learning is inseparable from environment or context (c.f. Brown et al., 1989; Greeno, 1997). Just-in-time information (scaffolding) in the well-ordered problems of the gameworld provide formative feedback within cycles of expertise (Gee, 2003). Indeed, good games effectively harness formative assessment to foster ongoing feedback cycles and customized player difficulty levels (Shute, 2011). In order to maintain this immersive context for learning, good games consist of ongoing assessment balanced with engaging mechanics and narrative (Squire, 2006). In multiplayer games, social interactions can provide their own definitive feedback cycles. They can provide a powerful environment for collaborative learning, supporting apprenticeship and collective higher-order thinking skills (Steinkuehler, 2004, 2008). Communities of practice (Lave & Wenger, 1991) often emerge around games (Steinkuehler, 2006a), fostering collective intelligence and an information-sharing participatory culture (Jenkins, 2006).

From Learning to Assessment

Many of these qualities are what makes good games rich learning context – and good assessment environments. Gee describes the systems thinking of digital games as ideal contexts for assessment, because they “allow us to track progress on multiple variable to gauge growth across time” and discover “different trajectories towards mastery and innovation” (Gee, 2012, p. 1). But this assessment is not only for the researcher, it enables formative feedback for the players themselves. *Civilization V*³ infographics are an example of the way players can get “beautifully designed representations of how they are going across time” on “many connected variables” and in comparison to other players (Gee, 2012, p. 1). Computer game-based assessment offers the capability of instantly adaptive embedded assessment within immersive,

³ <http://www.civilization5.com/>

agency-filled learning worlds (Halverson et al., in press). Gameworlds as learning contexts can provide a seamless “match between instruction and assessment,” an essential quality of authentic assessment (Scalise & Wilson, 2012, p. 290; Wiggins, 1990). Multi-player games can also capture an important collaborative learning element (c.f. Steinkuehler, 2004; 2006), what Squire calls “Participatory Assessment” (2012), in which participatory culture (Jenkins, 2006) sets and reinforces important indigenous performance standards. Indeed, standards of knowledge have been considered as central to domain representation in “the Best and Future Uses of Assessment in Games” (Baker, Chung & Delacruz, 2012). The Games for Learning Institute maintains that games can be fundamentally good assessment within four game-based learning functions: measuring preparations for future learning (priming for history lessons, for example), assessing new knowledge or skills (e.g. STEM games on cutting-edge science topics), to capture mastery of existing knowledge and skills (like multiplication tables and second language practice), and to evaluate life skills, including 21st century skills like critical thinking (Plass et al., 2012).

If games are, then, promising vehicles for assessment, what kind of measurements are appropriate to the medium? As we have seen above, games are very different from multiple choice tests, cognitive tutors, or even simulations in the sense that they offer a rich interactive world of roles, goals, and endogenously motivating agency (Costikyan, 2002; Gee, 2005; Squire, 2011). Statisticians and measurement specialists Kathleen Scalise and Mark Wilson tackle exactly this question in “Measurement Principles for Gaming” (2012). Core principles of good game-based assessment are that assessment should align with instructional goals, be able to measure student trajectories of growth over time (not just at a “final or supposedly significant time point”) and produce valid and reliable evidence of what learners know and can do (Scalise & Wilson, 2012, p. 290; National Research Council, 2001).

One way of implementing these principles is using Evidence Centered Design (ECD), which originally proposed to measure “knowledge and skills we want to develop in students, and the kinds of observations we need to evidence them” in contexts like “simulation-based assessment” (Mislevy et al., 2003, p.1). Explored in-depth in the next section, ECD is a multi-step method for aligning teaching content with tasks and evidence, and has had a range of applications in the digital world – including educational videogames.

Evidence Centered Design: Framework and Application to Virtual Learning Spaces

Introduction and Framework

In the world of simulation-based assessment, one approach to measurement of complex, process-oriented learning is Evidence Centered Design (ECD). ECD is a hypothesis-driven assessment method capable of measuring “behavior that bears evidence about key skills and knowledge” (Shute, 2011, p. 510; c.f. Mislevy et al., 2003). In other words, ECD aligns important learning content with tasks and resulting evidence for performance-based assessment.

What is Evidence Centered Design?

Evidence Centered Design is an assessment framework which “enables the estimation of students’ competency levels and further provides evidence supporting claims” about the knowledge and skills being assessed (Shute, 2011, p. 508; Mislevy et al., 2003). In other words, ECD aligns important learning content with tasks and resulting evidence for performance-based assessment.

The whole process (Figure 5) consists of three main chunks: research on what to teach (domain analysis and modeling), the design of the tasks (Conceptual Assessment Framework:

CAF), and implementation (Assessment Implementation and Delivery). While the outer two (first and last) describe parts of a temporal process, the middle layer – the CAF – is more focused on design. This is the center of ECD’s integration of content, evidence, and designed tasks (Mislevy, 2011; Mislevy & Haertel, 2006).

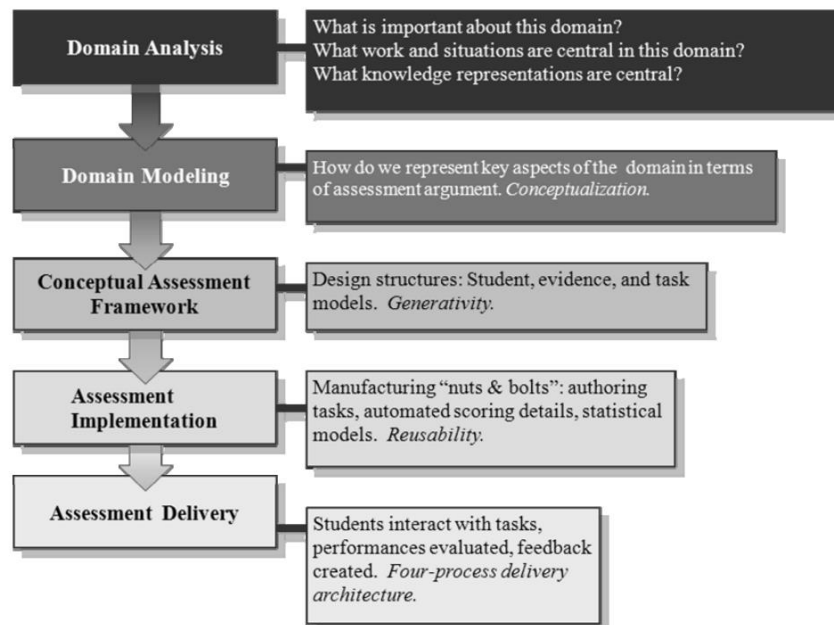


Figure 5. A full-scale ECD model (Mislevy, 2011).

Focusing on the CAF, it’s three main pieces are (Shute, 2011; Mislevy, 2003):

- 1) A competency model (CM) – alternately called the student model – defining key knowledge and skills to be assessed. Sometimes the CM is broken up into knowledge, skills, and attributes (KSAs). (*What are we measuring?*)
- 2) An evidence model detailing what behaviors or performances should reveal the CM’s constructs. (*How are we measuring it?*)

3) A task model, creating specific tasks that should elicit the behaviors that comprise the evidence. (*Where do we measure it?*)

It's worth noting that ECD lays out the design order as: competency model, desired evidence, *then* tasks (or core mechanics). In a simulation, the environment for the tasks would then be constructed in the last steps (implementation) – *after* the assessment task model in the CAF. Literally, the simulated environment is put last priority, and serves only as an auxiliary context in which to embed the tasks. The design of videogames, with richer elements than just a task model, can look very different; we will explore the challenges of ECD and videogames in upcoming sections.

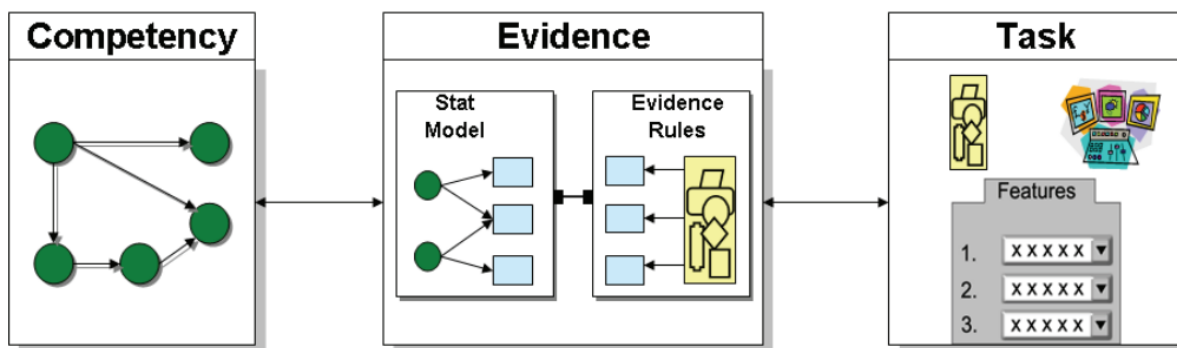


Figure 6. ECD's core – the Conceptual Assessment Framework (Zapata-Rivera, 2009)

Thus, ECD essentially amounts to a CAF sandwich, with this vital center layer capturing the core design and integration of content, tasks and evidence (Mislevy & Haertel, 2006).

Especially for pure simulations or digital performance assessments, ECD can serve beautifully as a streamlined framework articulating content, evidence, and tasks. Behrens says of this model: “It is flexible enough to accommodate the affordances of new technologies and the demand to measure new domains while providing a united framework to describe current practice across a wide range of assessment activities” (Behrens et al., 2012, p. 47). This broad

applicability gives ECD a universal appeal, while leaving open the opportunity to develop other ECD-inspired assessment models specifically tailored to individual learning technologies. Adaptable both in platform and in content, ECD is very useful in tackling abstract or very broad content like complex competencies and 21st century skills – in part because of its first two steps of systematic domain analysis and modeling. It is, in fact, the central method in the recent book Technology Based Assessments for 21st Century Skills (Mayrath et al., 2012), and the darling of several CRESST reports lauding ECD for assessment in the digital world (e.g. Behrens et al., 2010; Mislevy, 2011). In the following sections, we will explore applications of ECD to various contexts of digital assessments.

ECD in Digital Assessment

Recent ECD-based applications in the digital world include simulations and cognitive tutors. For example, Feng, Hansen, and Zapata-Rivera (2009) adapted ECD to an “Evidence Centered Design for Learning” framework to examine the ASSISTments intelligent tutoring system; one feature of this adaptation was the differentiation of assessment and instruction measures in ASSISTments by subdivision of tasks. ECD-related structures have been created for open-world digital learning environments, such as task-based performance metrics (Shelton & Parlin, 2012) auto-scoring in military simulations (Iseli et al., 2010), and engineering network simulation software (Frezza, Behrens & Mislevy, 2009). Other work, such as the “Evidenced Centered Activity Model” (Annetta et al., 2010, p. 24), Activity Centered Design (Gifford & Enyedy, 1999), and Gordon commission technology-adapted assessment structures (Behrens & DiCerbo, 2013) blend ECD constructs with activity-based models (Nardi, 1996). Competency-aligned task evaluation has also been considered central to digital scenario-based inquiry assessments, with particular attention on assessment characteristics (like time duration – pre,

post, or follow-up, proctored face-to-face or over distance, and response types of text or action) (e.g. deJong, Wilhelm & Anjewierden, 2012; Songer, 2012). A fundamental ECD-based framework in this area is “PADI”, which provides “Principled Assessment Designs For Inquiry” for simulation-based science assessments (Mislevy & Riconscente, 2005).

ECD in Videogames: A Primer

Although related to applications in simulations, using ECD in gameworlds can be more complex. Val Shute also notes that “making valid inferences about what the student knows, believes, and can do without disrupting the flow of the game” is a “main challenge” of educators in using games to support learning (Shute, 2011, p. 508). Assessing in-game performance is a “complex process that needs to take into account not only the engaging or motivational aspects of the activity but also the quality criteria that are needed according to the type of assessment that is being developed” (Zapata-Rivera & Bauer, 2012, p. 149). To help meet these challenges, many videogames & learning researchers have recognized the importance of aligned content, task, and evidence models in game-based assessment.

Mislevy, Behrens and team declare that key “Things Game Designers Need to Know About Assessment” are: 1) that “game design is compatible with assessment design,” because 2) assessment is “not really about numbers,” but the structure of reasoning, and 3) “Evidence-Centered Assessment Design” is a key means to bridging the two (Mislevy et al., 2012, p.59, 61, 66). Because ECD aligns content, tasks, and evidence, and can be structured to measure performance over a series of steps rather than in a single point of performance, the authors argue it is an optimal application to an educational gamespace (Mislevy et al., 2012).

Certainly, their support for game-based use of ECD is not an anomaly; many researchers have proposed elements of ECD for use in educational games. For example, qualitative student

evaluation in Quest Atlantis has incorporated some ECD structures in its Designing for Participation assessment model (e.g. Hickey & Jameson, 2012; Shute & Ke, 2012). In application to game-based professional training, Gaydos & Bauman (2012) have recommended ECD to help build in assessment for potential nursing simulation games. ECD can also be useful in vocabulary-building “game-inspired” software; in one example, the virtual learning environment “BELLA,” designed to help teach math vocabulary, experimental assessment was guided by “ECD principles” and combined with Bayesian reasoning to test the game in beta-level pilot studies (Zapata-Rivera, 2009, p.1; Zapata-Rivera & Hansen, 2009). In augmented reality games like the *River City* project, ECD has been used to help align locational data, tasks, and overall performance (e.g. Dede, 2012). The Games for Learning Institute references ECD as a core influence in their discussion of general assessment-informed design mechanics (Plass et al., 2012).

ECD in Videogames: Deep Practice

Various digital game-based assessment methodologies, based firmly in ECD, have been developed and researched extensively in the last decade. Three leading examples of these are Virtual Performance Assessment, Epistemic Network Analysis, and Stealth Assessment. Each of these methods will be described in detail, and briefly evaluated, in the section to follow.

Virtual Performance Assessment

The Virtual Performance Assessment project at Harvard is “developing and studying... immersive virtual performance assessment to assess scientific inquiry of middle school students” (Clarke-Midura et al., 2012, p. 134). The virtual world of VPA, built specifically for the assessment project, is a game-like simulated ecosystem in which the player picks an avatar and

goes on a mission. The missions are tasks in-game that are related to the “KSA”s (Knowledge, Skills, and Abilities) deemed directly related to science inquiry. These KSAs include making predictions, gathering data, reasoning about evidence-based claims, identifying causal relationships, and evaluating alternate explanations (p. 135). The KSA-based architecture is derived from PADI, an ECD-based structure for creating inquiry-based assessment (Mislevy & Riconscente, 2005). Essentially, the KSAs (or desired content for students to learn) align with the ECD task model in the game; avatars can choose their own path through those tasks, which include things like going to the lab, collecting samples around the land, talking to the lead scientist, or reading the latest research. According to the VPA framework, the progressive tasks all result in evidence which is evaluated in relationship to the KSAs. Ultimately, the player talks to the lead scientist and demonstrates his/her knowledge through using an argument constructor to exhibit causal reasoning about selected problems in the ecosystem (Clarke-Midura et al., 2012).

VPA is an interesting framework which puts refreshing emphasis on subtasks involved in the overall game goal; thus, formative assessment seems to play a large role, capturing the players’ process. The open-world component in player choice is an interesting feature, especially with a conceptual assessment framework attached to it. As the project moves forward, detail on specific scoring models and conclusive predictive analyses results (c.f. Clarke-Midura & Yudelson, 2013) can provide additional insight about VPA’s infrastructure. One potential limiting factor may be the requirement of the virtual environment to be designed around VPA, rather than be applicable to multiple game-based learning environments. Overall, VPA is an innovative example of how ECD-based assessment can inform design of game-like open worlds.



Figure 7. The world of Virtual Performance Assessment (Clarke-Midura et al., 2012)

Epistemic Network Analysis

Another example of ECD-related structures dictating design of game-like simulations is Epistemic Network Analysis. Epistemic Network Analysis (ENA) is used by David Shaffer’s research group to analyze data from “Epistemic Games,” which are simulations of professional STEM environments. They are designed based on the epistemic frame hypothesis, a theory of learning that analyzes thinking in terms of connections among frame elements: skills, knowledge, values, and justification or decision-making (otherwise known as epistemology) of a STEM profession (Shaffer et al., 2009). ENA maps to ECD starting with these frame elements (Sweet & Rupp, 2012), each as pieces of a competency model (or things the simulation wants to teach). These profession-based simulations include *Land Science* (a fictional internship with an urban planning company), *Journalism.net* (a fictional internship with a newspaper), and *Nephrotex* (a fictional internship with an engineering company).⁴ ENA is “a form of network analysis for assessing epistemic frames” in each simulation (Shaffer et al., 2009, p. 38). Data is collected as students interact with the simulations via chat and fictitious email (part of the ECD task model); the dialogue input is then coded as a skill, knowledge, identity indicator, value, or

⁴ <http://edgaps.org/gaps/projects/>

epistemology statement (part of the ECD evidence and scoring model). Each of these epistemic frame elements become nodes (or circles) on an SNA-like graph, and their co-frequency within certain excerpt chunks of chat become connectors. Over time, ENA maps player trajectories in each of the epistemic frame elements, and compares them to an “expert model” of participation in the game (Shaffer, 2009; Sweet & Rupp, 2012).

ENA is an interesting application of ECD to visualization of discourse over time. The element of quantifying qualitative input over the course of a simulation is methodologically very useful. However, it seems ENA would not have applicability to learning worlds not designed exactly under the constraints and scripted textual responses of an epistemic frame scenario. Working at an email-simulated job may, well, feel like work to some students, thus limiting the element of engagement and agency. It would be inspiring to see broader applicability and significance of this tool beyond that of a narrowly-defined epistemic simulation – however, the quantified visualization of a discourse-based ECD model is a novel method with potential in future game applications.

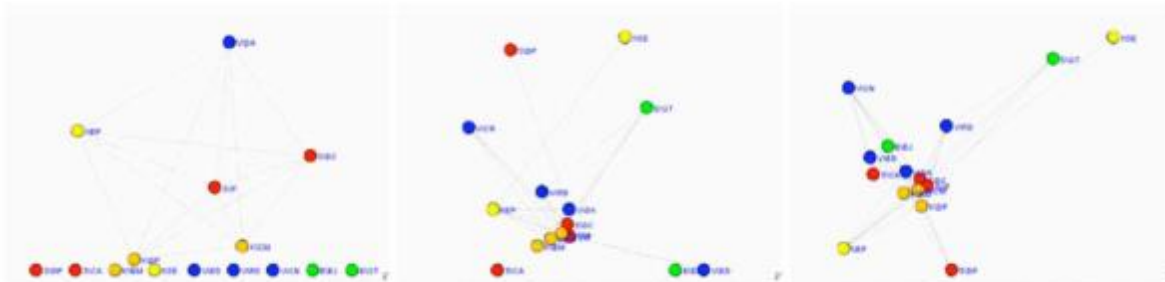


Figure 8. Epistemic Network Analysis

(http://www.wcer.wisc.edu/news/coverStories/2009/assessing_learning.php).

Stealth Assessment

Val Shute and her team of researchers at FSU have been working on Stealth Assessment (SA), the mission of which is to “identify key competencies and use games as instructional learning vehicles” (Shute, 2011, p. 505). SA, in essence, is an ECD-based model that is focused on connecting 21st century skill competency models to existing video games. Of the three methods talked about in detail in this section, SA sustains the most meticulous use of ECD. In line with ECD, SA maps out three pieces: 1) the competency model (CM) (skills and knowledge), 2) the evidence model (behaviors or performances evidencing the CM) and 3) the task model (also called an action model). The CM is usually broad 21st century skill (like Creative Problem Solving), which can then be broken down with a Bayesian network (Shute, 2011). An evidence model based on the CM is created, and then aligned with a task model (which defines player action within existing game mechanics). Thus, through gameplay, “learner performance data are continuously gathered during the course of playing/learning and inferences are made about the level of relevant competencies” (Shute, 2011, p. 504).

Application to the game *Oblivion* is given as an example. SA begins with a CM of “Creative Problem Solving”, a content base clearly different from that of the original game design (Shute, 2011). Where Bethesda studios likely had core game content goals revolving around immersive gameplay (e.g. combat affordances, economic and social interaction with NPCs, professions and customization opportunities), SA’s core content is an academically-defined “21st century skill” (Mayrath et al., 2012, p.3). The two base content models are quite different, and only directly overlap where the task model utilizes selected areas of the game mechanics (see diagram above for visual representation). One specific SA example in *Oblivion* considers possible player action in response to the in-game challenge of crossing a river full of

dangerous fish (choices were ultimately scored for levels of “Creative Problem Solving” ex-post-facto by two Oblivion experts) (Shute, 2011, p.517). Based on expert evaluation, certain gameplay paths were deemed more or less creative than others, then used to train a Bayesian network for ongoing evaluation of players. Herein lies the “stealth” component – SA uses existing core game mechanics to evaluate player actions connected to competency constructs.

Stealth Assessment is a clearly structured and well-supported model for applying very broad competencies (like 21st century skills) to existing video games. SA, admirably, respects player engagement and non-task-model game content, since it uses a “quiet but powerful process” which is “intended to support learning and maintain flow” (Shute, 2011, p. 504). SA in possible application to all kinds of engaging games (commercial games included) make it appealingly flexible (e.g. Shute & Kim, 2011). Currently, work is being done with SA and *Newton’s Playground* – a basic physics sandbox game – to demonstrate 21st century skills and implicit physics understanding (Ventura, Shute & Kim, 2013). However, SA’s complexity may not be necessary for games created from scratch in which the content model (e.g., addition) is clearly aligned with game mechanics (e.g. doing addition) and assessment goals (e.g. doing addition right). Additionally, SA may not be as transparent as methods like VPA for clear impact of assessment mechanics on core game design.

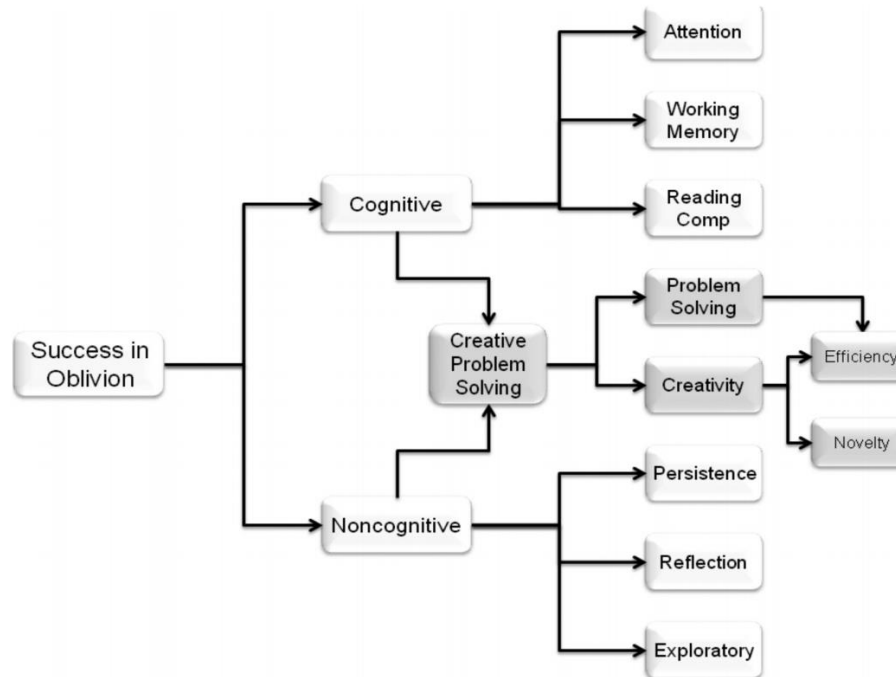


Figure 9. A Bayesian network example of mapping game success to creative problem solving in SA (Shute, 2011, p. 516).

Implications for Game-Based Assessment

As many researchers have recognized, ECD vitally informs our understanding of game-based assessment by emphasizing the importance of aligned content, tasks, and evidence. This has implications for educational game design, including clear articulation of content before the game design process, and integration of assessment-specific mechanics into the fabric of the game. As mentioned earlier, however, a good game is made up of more than just a task model – and in designing too much around specific performance assessment tasks one ends up with a simulation rather than a game. However, if joined with an assessment data framework more comprehensive than just a task model, ECD has the potential to tap into the power of total

integration: underlying game content, with evidence model, with total experience mechanics, with corresponding click-stream data stream and learning evidence.

Very recent work with ECD and games has followed this ADAGE-themed line of hybrid research, moving beyond ECD to leverage data mining techniques for an exploratory-confirmatory approach to identifying significant game events. One example is Jody Clarke-Midura's follow-up to her 2011 ECD work, teaming up with Ryan Baker to explore a hybrid ECD-data mining approach to VPA (Baker & Clarke-Midura, 2013). In another recent collaboration, GlassLab has formed as a recent partnership between the Institute of Play, EA Games and ETS to explore learning patterns in the simulation SimCity.EDU. Using techniques directly aligned with the ADAGE, the project has combined domain modeling of 21st skills for Evidence Centered Design with data mining for an "exploratory-confirmatory" approach not commonly employed with ECD-based research (Institute of Play, 2013).

ADAGE is informed by, and readily supports, multi-directional assessment of this nature. One key advantage in acknowledging all kinds of interactions, and connecting them fully with click-stream data structures, is the potential for maximizing player just-in-time feedback. Assessment shouldn't be just for the researcher or the teacher – if joined with the right underlying click-stream data framework, ECD could help leverage assessment data in ongoing support of the most important stakeholder of all: the player and learner.

Big Data and Assessment: Educational Data Mining

Introduction and Framework

Educational Data Mining (EDM), “is concerned with developing methods for exploring” large educational data streams (Baker & Yacef, 2009, p. 324), and “using those methods to better understand students and the settings in which they learn” (Romero & Ventura, 2010, p. 601; “International Educational Data Mining Society,” n.d.). Representing a host of education-specific machine learning tools, EDM can provide excellent groundwork for defining data mining methods readily applicable to educational gameworlds. This chapter outlines a schema of machine learning methods applied across EDM that can enable exploration of the potentially rich telemetry streams of digital learning environments (including educational videogames).

EDM data streams are typically massive, and sourced in continuously developing digital contexts; thus, it is an emergent, multi-disciplinary field in constant evolution (c.f. Romero & Ventura, 2010). Even EDM experts (Baker & Yacef, 2009; Romero & Ventura, 2010; U.S. Department of Education [DoE], 2012) survey the field from different perspectives. The state-of-the-field reviews from Baker & Yacef (2009) as well as the U.S. Department of Education (2012) focus on data modeling goals and methods, while Romero & Ventura (2010) organize around the human subjects of study (students, teachers, etc.) and context (classrooms, e-learning spaces, etc.). However, upon deeper analysis, a common schema can be derived from the expert reviews by extracting four underlying mutual components: 1) base educational contexts, 2) data types, 3) broad analysis goals, and 4) specific methods. All four of these focus on methods (rather than context, subject, or broad modeling); the first two describe the nature of the data, and the last two focus on the analysis of the data (through the lens of machine learning methods).

EDM: A Machine Learning Approach

The first two common pillars are base educational contexts and data types. First, EDM's main educational contexts include "offline education" sites like schools or tutoring centers (Romero & Ventura, 2010, p. 601); Learning Management Systems like e-learning sites and digital libraries (DoE, 2012; Baker & Yacef, 2009); and computer adaptive software – e.g. Intelligent Tutoring Systems (DoE, 2012; Romero & Ventura, 2010) and computer adaptive testing (Baker & Yacef, 2009). (Although not specifically mentioned in these expert EDM reviews, educational gameworlds can be categorized as computer software responsive to the user.) Secondly, the kinds of data derived from any of these settings can be both quantitative and qualitative. They can range from remote click-stream and text-based log file data (Romero & Ventura, 2010) to psychometric testing data (Baker & Yacef, 2009) to observational student interaction data (e.g. Baker et al., 2004).

The third and fourth EDM core components – broad analytic goals and specific methods – revolve around data analysis. Broad analysis goals (or "metagoals") common to the expert EDM synopses are visualization, relationship mining, and prediction (c.f. Baker & Yacef, 2009; Romero & Ventura, 2010; DoE, 2012). Visualization involves graphic representations of data to elucidate patterns; relationship mining looks specifically at associative patterns in the data; and prediction can project outcomes via algorithms of sequence, probability, and regression. The last core EDM component, specific method types, are subunits of these metagoals. These method types are finer-grained analysis categories, which include: descriptive visualization, social networks, clustering, association, classification/regression, and pattern mining (Figure 11). A loose mapping of these specific methods to the broad metagoals is visualized in Figure 11, complemented by a chart connecting common categories to all three synopses (Figure 10).

	Baker & Yacef (2009)	Romero & Ventura (2010)	Dept. of Education (2012)
Metagoals			
Visualization	<i>"Distillation of data for human judgment"</i>	<i>"Analysis and Visualization of Data"</i>	<i>"Distillation of data for human judgment"</i>
Relationship Mining	<i>"Relationship mining"</i>	<i>"Classification and Association Rule Mining"</i>	<i>"Relationship mining"</i>
Prediction	<i>"Prediction"</i>	<i>"Predicting Student Performance"</i>	<i>"Prediction"</i>
Method Types			
Descriptive Visualization	<i>"Distillation of data for human judgment"</i>	<i>"Analysis and Visualization of Data"</i>	<i>"Distillation of data for human judgment"</i>
Social Networks	<i>("Distillation of data for human judgment")</i>	<i>"Social Network Analysis"</i>	<i>("Distillation of data for human judgment")</i>
Clustering	<i>"Clustering"</i>	<i>"Clustering"</i>	<i>"Clustering"</i>
Association & Correlation	<i>"Association Rule Mining"</i> <i>"Correlation Mining"</i>	<i>"Association Rule Mining"</i> <i>"Correlation Analysis"</i>	<i>"Association Rule Mining"</i>
Classification & Regression	<i>"Classification"</i> <i>"Regression"</i>	<i>"Classification"</i> <i>"Regression"</i>	<i>"Classification"</i>
Pattern Mining	<i>"Sequential Pattern Mining"</i>	<i>"Sequential Pattern Mining"</i>	<i>"Sequential Pattern Mining"</i>

Figure 10. Metagoals and Method Types Common To All Three Expert Synopses

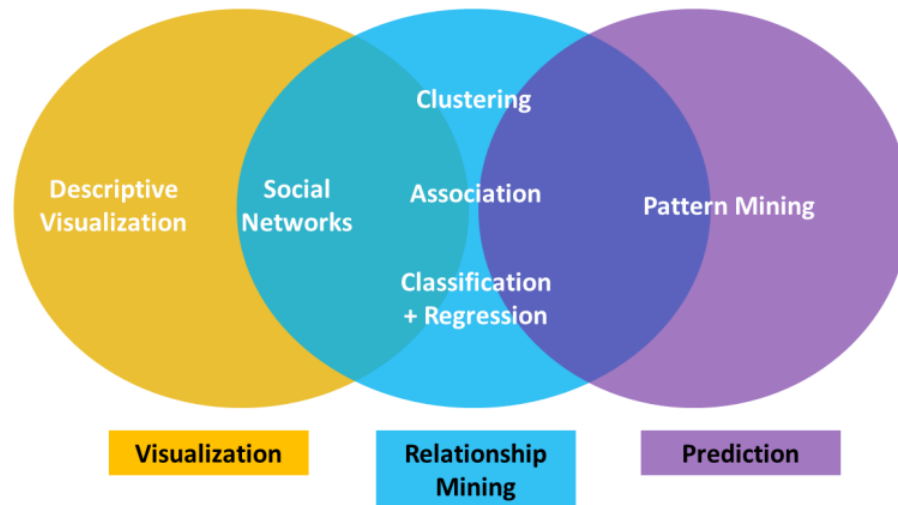


Figure 11. EDM analytics – specific methods loosely mapped to three base metagoals

Based on the core EDM content arc derived here, this literature review will move from visualization to relationship mining to prediction. In this trajectory, it will review the method types (across the metagoals above) in the following order: descriptive visualization, social networks, clustering, association/correlation & classification/regression, and pattern mining (see Figure 11). Under each method type, core analysis techniques will be discussed with examples from current EDM literature. It should be noted that the mapping in Figure 11 is intended to be “fuzzy”, in the sense that some analyses or method types belong to more than one category. Placement on the map above is in no way intended to be an absolute or mutually exclusive characterization.

Visualization, Cluster Analysis, and Social Network Analysis

Moving first into the broad category of Visualization (e.g. Tufte & Graves-Morris, 1983), we have what Romero and Ventura (2010, p. 4) call “analysis and visualization of data.” Descriptive statistics and visualization techniques are the two main vehicles for this met

category, which exists mainly to easily display global data characteristics like summaries of learner behavior (e.g. Wu & Leung, 2002). Indeed, this umbrella of techniques exists in the realm of what Baker and the DoE report as “distillation of data for human judgment” (Baker & Yacef, 2009, p. 5; DoE, 2012, p.11). As such, it entails graphic representations of statistics or information visualization techniques to give information applied to digital learning spaces. Visualization tools can range from graphs and charts in the Excel menu and SPSS data “explore” functions, to more elaborate tools like maps and integrated text graphics from software like visual.ly, to advanced code-based tools like Tableau, Google Chart API, Flot, D3, and Raphael⁵. In online course environments, for example, visualizations can include graphical displays of information about student entry/exit, popular pages used, and use over time (e.g. Ingram, 1999). From any LMS or student interface, it can show the most popular resources used by students (e.g. Sheard et al., 2003), time on site (e.g. Cohen & Beal, 2009), or small-scale performance measures like number of problems/assignments complete for a given time period (e.g. Feng & Heffernan, 2006). Baker, Corbett & Wagner’s (2006) text replay displays focus on the production of a textual pop-up summary of ITS student usage per problem for researchers, including click-stream interpretation categories like time, input, context/level in tutorial, and evaluation of performance. Hershkovitz and Nachmias (2008) represent student performance over time in “learnograms”, and visualizations showing *learning curves* are considered a very important tool in EDM (e.g. Baker, 2013; Ritter, Anderson, Koedinger & Corbett, 2007).

Social networks, another category of visualization, map the connections between individuals (nodes) in a web-like network (c.f. Srivastava, 2008). Recently, machine learning experts Baker and Siemens (2014) have supported this category as common to both EDM and learning analytics. Social network analysis and related techniques have been used to make

⁵ See this site for more aggregate information: <http://www.netmagazine.com/features/top-20-data-visualisation-tools>

teacher tools to visualize learner trajectories for optimized student grouping (Berland et al., 2013) and convey hierarchical changes in social structures over time (Carley, 2003). In text-based interactions, it's been used for connecting users with sources of online phrases; Simmons, Adamic, and Adar (Simmons, Adamic, & Adar, 2011). recently studied the use of memes in social media and their mapping to certain online sources through SNA. Network analysis has been leveraged to study patterns in online social interaction, indeed, since the early days of public internet use (e.g. Garton, Haythorntwaite, & Wellman, 1997).

Another visualization-related method is clustering. Very connected with descriptive graphic tools, it visualizes relationships by clustering similar data points together. Cluster analysis has been used in several forms (including k-means, k-nearest neighbor, and hierarchical) in EDM, especially to identify organically similar groups of students. At the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST), Kerr and Chung (2013) use clustering in an analysis of student performance in an educational game. Primarily a methods piece, their research found that fuzzy cluster analysis is more suited to the gameplay data, and more effectively identified unexpected player strategies in gameplay which helped explain student performance errors. Martinez and team (Martinez et al., 2011) used clustering to study patterns of interactivity between students engaged in collaborative learning around an interactive tabletop. The movements along the tabletop, which entailed reading and organizing slips of text, were coded in sequence and mined for patterns. Results included successful characterization of reasoning patterns in high- and low- achieving groups. Clustering has also been useful in characterizing LMS users. For example, Xu and Recker (Xu & Recker, 2012) employ this technique in creating profiles of digital library users. Through cluster analysis, they were able to

characterize three main groups of teachers (at varying levels of usage frequency) based on their habits in visiting and interacting with the digital learning space.

Association, Regression, and Classification

Romero and Ventura group them together as well, naming “regression, ...classification, and association rule mining” among “the most commonly applied” EDM tasks (2010, p. 603). Association rules – expressions that describe conditional relationships between variables (Zhang & Zhang, 2002; Sasikala et al., 2011) – are a simple and useful tool in educational data mining (c.f. Merceron & Yacef, 2011). Frequently used in the domain of providing feedback for supporting instructors (Romero & Ventura, 2010), association rules have been leveraged to provide automated information for appraise online course effectiveness (e.g. Retalis et al., 2006), and to help improve education quality in the academic community in conjunction with cluster analysis (e.g. Vranic et al., 2007). Similarly, association rules been employed to help improve virtual educational environments (Zaïane & Luo, 2001; Zheng et al., 2008).

Correlation mining is in a similar family, since its essential function is to find patterns of association in the data (Baker, 2010; Romero & Ventura, 2010). This kind of mining is so “hot” that it is the engine of Google Analytics’ new brainchild: Google Correlate⁶. This tool is a classic example of applied data mining; it finds search patterns which correspond with real-world trends. Specific to EDM, correlation-based analyses have been used to predict e-learners’ performance in online courses (Wang & Newlin, 2002) and exam scores in online tutoring (Pritchard & Warnakulasooriya, 2005). It has also been used in conjunction with other analyses. For example, Nkambou and team use correlation and association rule discovery methods with sequential

⁶ <http://www.google.com/trends/correlate>

pattern mining to help better guide learners in ITS problem-solving scenarios (e.g. Nkambou et al., 2007). Correlation mining procedures, combined with other EDM techniques, are mapped for use in creating auto-assessment of e-learning Moodle course structures by Romero, Ventura, and Garcia (2008). Merceron and Yacef combine correlation strategies with association rule mining to better understand co-occurrence of mistake types in a cognitive tutor environment (2011).

Generally, classification and regression are predictive methods that can include classic linear models like logistic and step regression (e.g. Baker, Gowda, Corbett & Ocumpaugh, 2012), as well as tree-based predictive models. In the fuzzy mapping in Figure 11, this category is considered mainly predictive, but also can provide visualization (in tree and linear forms) and explores relationships between variables; thus, it is placed in at the intersection of the three metagoals. Prominent techniques include Classification and Regression Trees (dubbed “CART” – e.g. Breiman et al., 1984), analysis techniques that use a tree-like branching schema. CART can be used to describe a broad category of analyses, and is also the name of a discrete predictive algorithm used in data mining software like WEKA (Hall et al., 2009). In this review, it is mainly referred to as the umbrella category of classification and regression trees (e.g. Breiman et al., 1984). Generally, CART is designed to explain a chosen outcome variable through the mapping of different associated conditions. For example, CART was employed to create profiles of students based on propensity to take online classes (Yu et al., 2008). In another example, classification trees enabled automatic detection of students’ learning style with LMS log data (Lee et al., 2009); additionally, one study used it as an illustration of learning behavior to better categorize learners into different cognitive style groups (Lee, Chen & Liu, 2007). CART can also be leveraged in conjunction with other analyses. In one analysis on a virtual educational game, CART was used in conjunction with a first-order Markov model to show characteristics of

gameplay which predicted completion of learning tasks (Owen, Ramirez, Salmon & Halverson, 2014). In a more traditional EDM study, Anaya and Boticario (2011) triangulated cluster analysis with a REPTree classifier to help power a teacher-friendly collaborative learning tool. Kelly & Tangney (2005) use probability-based “naïve” Bayes classifier (also a classification algorithm) to characterize learning style according to digital content interaction and learning behavior in “First Aid for You,” a novel adaptive educational program. In a more recent study, a Bayes classifier was used in tandem with logistic regression and rule-based mining in finding predictors of student attrition at the university level (Dekker et al., 2009).

Core to early EDM development was the use of classifiers used to develop detectors of student strategies or affect in cognitive tutor environments, an application called developing “student models” (Baker, 2010, p. 326). A blueprint for this method was defined in “Developing a Generalizable Detector of When Students Game the System” (Baker et al., 2008), with foundational work started several years before (Baker et al., 2004). Other similar classification work done with detectors includes Shih and team’s research on distinguishing helpful and unhelpful kinds of hint retrieval behaviors in ITS (Shih, Koedinger, & Scheines, 2011); identification of off-task behavior (Cetintas et al., 2009) that indicates gaming the system (Walonoski & Heffernan, 2006) and impact on learning (Cocea et al., 2009); using text-based interaction for detecting learning affect (D’Mello et al., 2008); and measuring affect around agent-based instruction for students with learning disabilities (Woolf et al., 2010). It’s heartening to note that for low-performing students, working with a virtual “pedagogical” character had a positive effect on self-reported frustration and anxiety in Woolf’s study. Other studies around positive affect and detectors include the work of Chaffar (Chaffar, Derbali & Frasson, 2009) and

McQuiggan (McQuiggan, Mott & Lester, 2008), who leveraged classification techniques to detect positive emotional state and self-efficacy in the context of cognitive tutor use.

Continued research on student modeling has merged detector work (classification and correlation mining, (e.g. Baker et al., 2004) with the power of statistical regression techniques (e.g. Baker et al., 2010). Logistic, linear, and multivariate regression, like CART, generate predictive information about an outcome variable. Used in applications like providing feedback to teachers, traditional statistical regression has been instrumental in assessing the effect of different educational interventions on students (e.g. Feng, Beck & Heffernan, 2009), predicting end student performance from web-based log and test scores (e.g. Yu et al., 1999; Ibrahim & Rusli, 2007), and anticipating future time spent on an LMS web page (Arnold et al., 2005).

Pattern Mining

Another major EDM method type is sequential pattern mining. Sequential pattern mining entails techniques that “capture sequential events” (DoE, 2012, p. 11). “Sequential pattern mining,” as Romero and Ventura explain (2010, p.606), “aims to discover the relationships between occurrences of sequential events, to find if there exists any specific order in the occurrences.” This category can include text pattern mining techniques, models of temporal sequence, and Bayesian probability methods.

One kind of raw data highlighted by Romero and Ventura is text or written language based (Romero & Ventura, 2007). This is especially relevant in connecting with educational research, where student writing and reflection can be a large component of assessment (e.g. Hickey & Jameson, 2012). Computational linguistics and related machine learning studies have used Natural Language Processing (NLP – c.f. Manning & Schütze, 1999) to mine text-based data for linguistic patterns. NLP and related text mining techniques have contributed to many

EDM areas of application (Romero & Ventura, 2010). Providing instructor feedback and constructing concept maps to support curriculum are two such areas. For example, in providing instructor feedback, text mining algorithms have been key in valence-based auto-evaluation of user opinions (i.e. categorized as positive or negative) in e-learning forums (e.g. Song, Lin, & Yang, 2007). Other automated text-based analysis have measured various forms of student verbage, including evaluating spoken responses to tutors (Zhang et al., 2008) and mining students' writing differences to explore divergent cognition styles (Huang et al., 2006). Auto-creation of domain concept maps from academic articles are another pedagogical support outcome of NLP (e.g. Chen et al., 2008).

Methods expressing sequential probabilities have been used with NLP. Markov modeling is an example of this, used in Rabiner's 1989 article on "Hidden Markov Models and Selected Applications in Speech Recognition." Other applications include methods research on language patterns in a collaborative writing process (Southavilay, Yacef & Calvo, 2010), contrasting two types of Markov models in mining for core trends. Markov modeling and related temporal-sequence methods are also used in non-textual data for EDM purposes. In LMS research, for example, Markov models have been used to monitor system information in support of teachers, namely providing notification of student errors and technical issues (Heathcote & Prakash, 2007). They've also been of use in student classification; one LMS study created user characterizations according to HMM output on navigation and content-access patterns (Fok et al., 2005), while an educational game study used an MM for tracking navigation in high- and low- performing user groups (Owen et al., 2014). Doug Clark and team also detail an experimental design using HMM with Computer Adaptive Testing principles to uncover

students' strategic moves and explanatory responses in scientific modeling games (Clark, Martinez-Garza, Biswas, Luecht & Sengupta, 2012).

In related probability-based methods, Bayesian models have been used with EDM research. In another case, “outlier detection” of extreme data points (e.g. Vee et al., 2006; Romero & Ventura, 2007) was achieved in an e-learning context through Bayesian predictive distribution (Ueno & Nagaoka, 2002). In predicting the need for help in an e-learning environment, Mavrikis (2008) utilized Bayesian networks; he formalized his process in a follow-up piece on “Modelling student interactions in intelligent learning environments: constructive Bayesian networks from data” (Mavrikis, 2010). Supporting collaborative learning, Bayes' nets have been combined with clustering methods to create informed skill-based student groups in distance learning courses (e.g. Hämäläinen et al., 2004). In a Bayesian methods piece, Millan and team (2010) provide a framework for using Bayes' nets to engineer student models in intelligent tutoring systems. Also focused on student models is Bayesian Knowledge Tracing (a kind of Bayes network), used to help increase accuracy with classification of learner behaviors into informed, guessing, or slipping performance behaviors (e.g. Corbett & Anderson, 1994; Baker, Corbett & Aleven, 2008).

EDM: Conclusion

In the applied machine learning schema, main EDM methods include the metagoals of visualization, relationship mining, and prediction. These metagoals can organically extend to the related click-stream data pools of educational gameworlds – and can greatly inform game telemetry assessment structures. For example, visualizing paths through the gameworld, exploring relationships between gameplay patterns and learning, and modeling predictive significance of player actions at specific points in gameplay can provide valuable insight into

play and learning. Through specific method types like descriptive visualization, clustering, and social network analysis, graphic representations can reveal important connections in large game data sets. Association rule mining & correlation can define clear connections between game data and learning. The predictive power of classification and regression can provide deep insight about the gameplay factors which impact learning outcomes. Sequential pattern mining and Bayesian networks can uncover vital likelihoods and sequential connections between gameplay and assessment elements, with temporal- and probability-based mappings Overall, this schema of data-mining methods for educational games – based in EDM-applied machine learning – can help inform our game-based assessment structures to better enable adaptive, engaging learning experiences in play.

Chapter Three: ADAGE and Progenitor X

Introduction

Evidence Centered Design (ECD) and Educational Data Mining (EDM) both hold important contributions to understanding assessment in digital learning environments. ECD promotes content-driven design choices, and aligns data collection with selected performance tasks hypothesized to constitute evidence of learning (c.f. Mislevy & Haertel, 2006). Application of ECD to computer-based learning realms (e.g. Mayrath et al., 2012) means that such evidence often resides in data-rich log-files. While rigorous research has been done on the conceptual frameworks of ECD in digital worlds (e.g. Mislevy, 2011) there is little mention of the specific alignment between user action and click-stream data structures from which evidence is obtained. EDM, on the other hand, focuses heavily on masses of educational log-file data (Baker & Yacef, 2009; Romero & Ventura, 2010). In data mining, assigning semantic meaning to data “often need[s] to be determined by properties in the data itself, rather than in advance;” unfiltered data sets with contextual information like “time, sequence, and context” play “important roles in the study of educational data” (International Educational Data Mining Society, n.d.).

In educational videogames, the idea that design should align with evidence of learning (a la ECD) need not be mutually exclusive from the idea that unfiltered, richly-structured data is vital to forming meaning (a la EDM). In merging these two perspectives, core game design frameworks can be synthesized with distinct pedagogical task models to capture a wide range of context-rich interactions. An optimized game-based assessment model, then, articulates a click-stream data framework aligned with educational game mechanics for broad, context-rich assessment data output – that can be used with both hypothesis testing and machine learning

techniques. At the Games+Learning+Society group, this optimized model is ADAGE: Assessment Data Aggregator for Game Environments (Owen & Halverson, 2013).

ADAGE (Assessment Data Aggregator for Game Environments)

ADAGE was designed to transform game-based log file data into evidence of learning. Essentially, it integrates core game design structures into a click-stream data (telemetry) schema, which is then seeded with context vital to informing learning studies. These data can be used to identify patterns in play within and across players (using data mining and learning analytic techniques) as well as statistical methods for testing hypotheses that compare play to content models (cf. Loh, 2012; Halverson & Owen, 2014). Overall, ADAGE provides a standardized game telemetry framework with a rich, method-agnostic data yield, efficient enough to have scalability, and flexible enough to use across games.

Currently, ADAGE is both a conceptual frame for capturing assessment data for games, as well as an API and data output engine. The following paragraphs will overview the assessment mechanics and telemetry schema of ADAGE, using the game of *Progenitor X* as an example.

Assessment Mechanics

Assessment mechanics are ADAGE structures built into the game that allow for research on play and learning. Understanding game-based learning requires two levels of assessment mechanics: one to trace the paths players take through a game, and the other to access the player experience of game play (Schell, 2008). Squire asserts that games as designed experiences (2006) provide endogenous engagement (Costikyan, 2002) for the player through “roles, goals, and agency” (Squire, 2011, p. 29). Thus, in learning games, there can two core kinds of designed

mechanics: one set related to progression through the gameworld, as an engaging learning context (Gee, 2005; Salen & Zimmerman, 2004); another may be designed as more direct measures of the content the game is trying to teach (e.g. Clarke-Midura et al., 2012). Ideally, these also overlap; good educational games meld learning mechanisms with the core mechanics of the game, where gameplay itself is the only necessary assessment (Gee, 2012; Shute, 2011).

The ADAGE framework identifies underlying game mechanics for which serve as core occasions for player interaction. There are three base types of Assessment Mechanics: *Game Units* (capturing basic play progression), *Critical Achievements* (formative assessment of content), and *Boss Level* (naturalistic summative assessment). As “Assessment Mechanics”, they serve as data-collection (or assessment) anchor points, which yield data informed by core educational game design structures. This terminology also parallels concepts of formative and summative assessment in formal learning environments (Harlen & James, 1997), and formalizes them as powerful elements of game design (c.f. Gee, 2012).

Through Assessment Mechanics (AMs), ADAGE operationalizes player interaction (Salen and Zimmerman, 2004) as the vital link between experience and game design (Schell, 2008). These three core AM types can easily overlap within a gameworld; they are not mutually exclusive, though they have distinct categories. Additionally, every game does not have to have all AMs in order to use ADAGE. In this section, we will describe each mechanic, and connect it to ADAGE’s underlying telemetry structure.

Game Units. The game Units represent the core progress mechanic of the game. For example, in a game like *World of Warcraft (WoW)*, the core unit is quests. By definition, game units have the property of being a repeating, consistent vehicle for making progress through the gameworld. Units can also be part of a hierarchy – for example, one set of quests may make up a particular

map area, and completing all the maps means finishing the game. Thus, from broadest to smallest, game Unit hierarchy might be: game-map-quest. The idea behind Units is that they are flexible enough to work across genres; Currently, ADAGE Unit structure is applied to five different GLS games (*Progenitor X*, *Fair Play*, *Anatomy Pro Am*, *Tenacity*, and *Crystals of Kaydor*)⁷ each with different genres and Unit types.

Critical Achievements. Critical Achievements (CAs) in ADAGE are direct formative assessment slices of the content model. They are moments of direct content measurement within the context of normal gameplay. Seamlessly woven into the fabric of the game, CAs use naturalistic game mechanics to measure underlying educational content. For example, *Fair Play* is a GLS game which teaches about implicit bias in graduate education settings. In one *Fair Play* CA, the player needs to correctly identify a given bias to another character in order to progress. This is a direct demonstration of bias knowledge (as opposed to indirect movement through the learning context, like in game Units). The CA data structure aligns very well with ECD task analyses. CAs (analogous to the “task model” in ECD) are intended to be one kind of direct content assessment embedded in gameplay, looking at selected moments of performance as learning measures. Ultimately, CAs are a unique feature of educational games, and capture both learning AND play dynamics in the user experience.

Boss Level. The Boss Level is a final stage of a game that is a culmination of skills learned in gameplay. It is a naturalistic summative assessment, and can include both learning and progress mechanics (like CAs and Units). Gee notes that powerful embedded assessment occurs in “boss battles, which require players to integrate many of the separate skills they have picked up” throughout the game (2008, p. 23). Games are an ideal medium for this summative assessment,

⁷ <http://www.gameslearningsociety.org/projects/>

he asserts, since they can provide just-in-time performance feedback with low cost of failure (Gee, 2005). By formalizing the Boss Level as an Assessment Mechanic in ADAGE, we encourage deliberate inclusion of summative assessment in game design, and provide corresponding telemetry API structures for implementation.

Telemetry Framework

The Assessment Mechanics, informed by game design and assessment research, create a conceptual framework for identifying interaction data. The next ADAGE step moves us from concept (AMs) to implementation (telemetry). The telemetry framework hinges on the AMs to create a schema of context-rich data tags for implementation in the game code. Interpretation of student interaction often hinges on the context of the learning environment (in this case, the designed gameworld). The telemetry schema addresses this need by seeding the AM interaction data with vital contextual information.

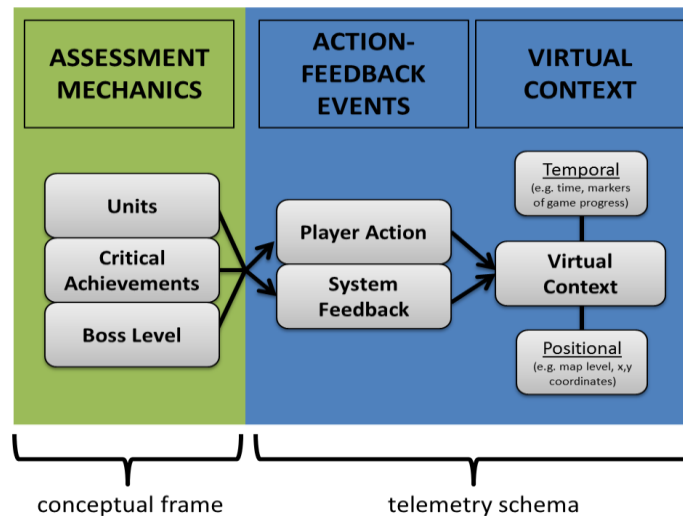


Figure 12. ADAGE Assessment Mechanics and telemetry schema.

The telemetry schema has two layers: an action-feedback layer, and a Virtual Context layer. First, for each Assessment Mechanic, it identifies two sources of direct interaction: user action, and system feedback. It articulates the vital action-feedback loop (c.f. Salen & Zimmerman, 2008) that comprises interaction between the player and the game. The second layer, called the Virtual Context, attaches important contextual information to each action-feedback event. The Virtual Context can include things like timestamp, map level, and screen x,y coordinates. These two layers work in tandem to provide context-rich telemetry data on AM-based gameplay trajectories (Figure 12).

Feature Engineering & Analysis Lenses

ADAGE's context-rich data make ideal building blocks for feature engineering. Features are essentially variables of interest in the data, which can range from simple click locations to complex measures like accuracy over time. The features constructed, in turn, can be used across a broad range of analysis techniques. Data lenses can include descriptive statistics, hypothesis-driven applied statistics, and machine learning techniques. Methodologies for hypothesis testing (like ECD) can use ADAGE data as dependent variables, independent variables, and covariates for use in associative or predictive modeling. Lastly, ADAGE data lends itself to learning analytic techniques often used with big data sets.

ADAGE: Application to *Progenitor X*

This dissertation's empirical exploration of ADAGE begins with its application to the GLS game *Progenitor X*. *Progenitor* is a puzzle game set in an apocalyptic world overrun by ravenous zombies, providing the player with the agency and motivation to become the sole regenerative biologist who can save the planet, one zombie at a time. The regenerative biology

content model is manifested in the main mission of the player: to cultivate and differentiate stem cells, assemble tissues and replace organs that have been contaminated with a zombie virus. Despite its supernatural storyline, *Progenitor* is designed to teach cutting-edge knowledge and processes in stem-cell science, and is rooted in serious collaboration with top regenerative medicine scientists at the Wisconsin Institute for Discovery (WID). Collaborators include Dr. Jamie Thomson, Director of Regenerative Medicine at the Morgridge Institute for Research (MIR); Dr. Rupa Shevde, Director of Outreach Experiences at MIR; and Dr. Gary Lyons, an esteemed UW professor of regenerative biology.

Completing game play requires players to solve 15 cell, tissue and organ puzzle cycles within a series of eight sequenced “Objectives”. In early objectives, players encounter a cell cycle that involves a sequence of treatment and collection tools that transform pluripotent stem cells into particular cell types (Figure 13). Next, tissue cycles require players to layer successfully transformed cells into segments of tissue; in later game play, the organ cycle requires the assembly of tissue segments into organ shapes (Figure 13, last two graphics). While players learn the cell cycle first, subsequent play requires players to repeat cell-tissue, then cell-cell-tissue cycles in order to move through the game. In the final level of the game (Objective 8), the organ-building phase functioned as a boss-level that required players to use all the skills learned in the cell and tissue cycle (e.g. an organ-cell-cell-tissue cycle sequence) to complete the game. Table 1 lays out the basic sequence of play. An average play-through of all the cycles in the game has taken middle school players an average of 25 minutes (with 40 minutes defining an upper limit of $+2\sigma$).



Figure 13. Progenitor X Cell, Tissue, and Organ Cycles

Table 1

Structure of Progenitor X Puzzle Gameplay

Cycle Type	Objective Number	Objective Name
cell	0	(cell tutorial)
cell	1	"collect 10 red mesoderm cells"
cell	2	"collect 10 blue ectoderm cells"
cell	2	"collect 10 blue ectoderm cells"
tissue	3	"create a tissue"
tissue	4	"create a second tissue"
cell	5	"create green endoderm cells & build the final tissue"
cell	5	"create green endoderm cells & build the final tissue"
tissue	5	"create green endoderm cells & build the final tissue"
organ	6	"locate necrotic zombie tissue"
tissue	7	"create a replacement heart tissue"
organ	8	"find and replicate remaining Necrotic Zombie Tissue"
cell	8	"find and replicate remaining Necrotic Zombie Tissue"
cell	8	"find and replicate remaining Necrotic Zombie Tissue"
tissue	8	"find and replicate remaining Necrotic Zombie Tissue"

Progenitor Data Collection and Telemetry

Data collection and early analysis of *Progenitor* telemetry revealed interesting data features and opened up new lines of research inquiry. Originally, GLS invited 110 middle school students to play the game as a part of a summer enrichment program at the Wisconsin Institute for Discovery in 2012. As part of the IRB-approved protocol, students completed a pre- and

post- content assessment, which included a series of questions about the stem-cell content model based on consultation with UW-Madison regenerative biologists. This pre-post protocol incorporated interview, multiple choice, and open-ended questions, and resulted from collaboration between WID content experts (including Dr. Gary Lyons and Daryl Nelson), secondary school science teachers, psychometricians, and game-based assessment researchers. (See Appendix for more detail.) As part of the process, we also collected demographic information on the 110 players who completed *Progenitor X* to enable connection of player background with in-game learning trajectories.

Players' improvement on this biology assessment (from the pre- to the post- test) is used as a learning measure in this dissertation. Percent improvement (delta) from pre- to post- was used to sort players into highest and lowest learning groups. The pre- and post- test data were ideally distributed for this measurement (Figure 14), with pre- scores averaging 46% (with 97% of players scoring below the maximum). Post- scores averaged 67%, and had a graduated increase in score (right, Figure 14), with only 6% of players maxing out scale at 100%. The upper quartile of learners consisted of players with the highest percent *improvement* in score (n=33), while the lower quartile of learners consisted of players with the lowest improvement (n=41).

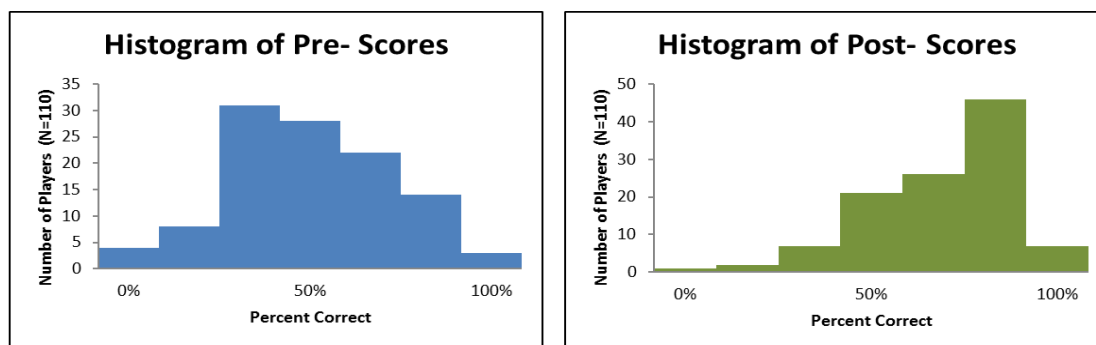


Figure 14. Histograms of scores on *Progenitor* pre- and post- biology assessment.

In triangulation with this pre-post protocol, telemetry feature engineering started early on in the ADAGE process with *Progenitor X*, initially in investigation of possible player paths in the gamespace. Examination started with mapping performance outcomes for each cycle in *Progenitor X*. When overall failure (as one broad umbrella) demonstrated no relationship to learning in pilot research (Owen et al., 2012), nuances of failure became the next natural exploration. What resulted was a differentiation of ways to succeed and fail, and the derivation of a new data key feature: “far failure” (Owen et al., 2013).

Far failure, essentially, was a kind of failure that occurred as a result of student performing actions directly contrary to game cues. An example of this would be loading the wrong cell onto the screen, or collecting the wrong cell at the end of a cycle. Near failure, on the other hand, occurred when players followed all instructional cues, but failed while operating within the suggested parameters of the task (i.e. running out of health while working to generate the right cells). The upcoming analysis, Chapter Four, delves into these nuances of *Progenitor* performance in detail.

Success, far failure, and near failure became central performance data features for *Progenitor X*, along with learning outcomes from the pre-/post-, and cycle-based ADAGE virtual context data. Exploring these data in connection with one another can provide insights into in-game performance as it relates to learning, as the analyses below examine. Each of the following analyses described uses the data from this section (n=110), collected with the methods detailed above, and based in ADAGE for the game *Progenitor X*.

Chapter Four:

Success and Shades of Failure – Feature Engineering and Applied Statistics

Introduction

Exploring player performance within the game's core biology laboratory mechanics, this chapter investigates the intersection between the microworld elements of game as content delivery (Lens B, Figure 1) and player choice (Lens C). In doing so, this analysis section focuses on ADAGE-based, iterative feature engineering and applied statistics around the concepts of failure and success in *Progenitor X*. It explores the research question: How are in-game success and kinds of failure related to learning outcomes? A vital part of analysis for big data, feature selection processes (e.g. Guyon & Elisseeff, 2003; Romero et al., 2011) emphasize iteration for this reason – early research can inform more nuanced, ongoing feature engineering (e.g. Fogarty, 2006; Arnold, Nallapati, & Cohen, 2008). Accordingly, this analysis builds on the existing research to systematically engineer more sequentially nuanced, refined indices of failure and success. These indices align with three analysis strategies, which use descriptive and nonparametric statistics to newly enrich understanding of success and failure in detailed gameplay progression and relationship to learning outcomes. Thus, the research question for this analysis is: how do fine-grained patterns in performance data connect to learning outcomes?

Overall, the feature engineering and analysis sections are aligned with the research question in theme. Using six computational lenses for feature generation, these new telemetry indices provide greater sequential resolution, more sophisticated cross-group comparative features, and more nuanced performance differentiation (down to each cycle level). These align with the goals for analysis, which include mapping previously unexplored patterns over time,

differences in these patterns between learning groups, and relationships of in-game performance to learning outcomes. To explore these goals, three analysis strategies were employed to help visualize and quantify important trends in the newly enriched data. Based on these analyses, results yielded three trends of performance findings connected with learning: overall game progression and success, tissue failure in mid-game synthesis levels, and the changing meaning of cell failure (specifically in early vs. late gameplay levels).

The study's feature engineering process is outlined first, including the definition of *Progenitor* performance features, computational lenses, and final output categories. The second part of the chapter discusses analyses, starting with methods, then followed by results along the three main trends mentioned above.

Feature Engineering

Overview

Iterative feature engineering with ADAGE starts with the basic telemetry schema (ADA Base Tags) for raw categories of game output. This comes first in individual student logs, then can be organized into more aggregate multi-student data. Third, new indices can then be made, informed by the research question. In this section, this three-step process will be described, and then applied to this study's specific goals.

After data collection, step one of the ADAGE-based feature engineering process is producing individual logs with all base data tags intact. Individual student data sheets have one user action per line, whose columns would give detail about the meaning of the action. (Student logs have had up to 15,000 event lines each.) For example, one line of one student's data from a GLS game would contain the following kinds of information:

	player	timestamp	objective added count	objective completed count	total failure	total successes	total populates
1							
2	c_10	1329	2	1	39	15	54
3	c_7	1497	4	3	9	7	16
4	c_8	1433	7	6	7	19	25
5	d_5	977	15	13	6	25	29
6	e_1	1357	6	5	6	8	14
7	e_3	1492	5	4	8	25	33
8	e_4	908	8	7	4	13	16
9	e_6	1034	5	4	5	8	13
10	f_10	1643	11	10	10	18	25
11	f_2	1144	5	5	6	8	14

Figure 16. Example of telemetry totals csv.

After basic multi-player data is aggregated, a third step is creating more sophisticated data indices, called feature engineering. (This often occurs after an intermediary round of exploratory analysis on the step two data.) Like step two, the data focus will depend on the research question. Any of the base data types (see Figure 15 and Table 2) can be combined to engineer a new data feature. As seen in a use-case functionality mockup of the ADAGE analytic interface (Figure 17), a range of mathematical operations can be applied to two base data types to create a new telemetry feature. For example, one can take total time and divide it by the number of objectives completed; this yields a new feature of time per objective completed.

ADAGE FEATURE ENGINEERING OPTIONS:

Which computational lens would you like to use?

Totals

Average

Ratio

First/Last Unit Played

Time series

Isomorphic series

Figure 17. ADAGE analytic interface: functionality mockup.

Progenitor X: Feature Process and Computational Lenses

Specifically for *Progenitor X*, a main study goal was to systematically engineer fine-grained, failure-specific features which take into account cycle type, specific objectives, sequential patterns, isomorphic play cycles and ratio-based performance. Using the steps outlined above, this round of feature engineering identified base data types of interest for an aggregation of multi-student base data (steps 1+2), and then engineered new indices based on these data with a finite series of mathematical operations (step 3).

Steps one and two required identification of base data types of interest for both individual and aggregate logs. The core features of interest were several types of failure (far failure, near failure, and total failure) and success – all parsed by individual cycle, the core unit of the game. As an additional layer, each of these data were also identified by cycle type and objective number. A description of each kind of failure is given in detail in the “Definitions” section following Table 3.

Third came combining these base data types, using mathematical operations, to create new *Progenitor* data features. Delving deeper into the significant constructs of far failure and success, the new telemetry indices explored greater context-based failure differentiation, more sophisticated (compound) indices for player comparison, and increased temporal resolution. These became “data themes” in alignment with analysis goals. To create new features along these themes, six computational lenses were used. The table below illustrates I) computational processes to be used in creating the new features (plugging in any one or two base data types), II) examples of resulting indices, and III) the corresponding data theme.

Table 3

Feature Engineering Computational Lenses: Operation, Example, and Theme

I) Computational Lenses	II) Example Features	III) Data Theme (A, B, C)
<ul style="list-style-type: none"> Totals (overall, for each objective, and for each cycle type) 	<ul style="list-style-type: none"> - total objective 1 far failure - total tissue near failure - total objective 3 successes - total failures (whole game) 	A) Context-specific performance data
<ul style="list-style-type: none"> Ratios (proportions of performance** features) 	<ul style="list-style-type: none"> - near failures : far failures - successes: failures - far failures: total failures 	B) Compound indices for comparison
<ul style="list-style-type: none"> Averages (per objective & per cycle type) 	<ul style="list-style-type: none"> - average far failure per objective completed - average success per objective added 	B) Compound indices for comparison
<ul style="list-style-type: none"> Performance data for last objective played (customized per student) 	<ul style="list-style-type: none"> - successes in last played objective - near failures in last played objective 	B) Compound indices for comparison
<ul style="list-style-type: none"> Time series (taken as a sequence of data points, by objective & by cycle type) 	<ul style="list-style-type: none"> - list of near failures: objectives 1, 2, and 3 - list of successes: objectives 6, 7, & 8 - list of far failure in tutorial levels only: objectives 1, 2, and 4 	C) Temporal sequence data
<ul style="list-style-type: none"> Isomorphic sequence (for identical cycles only) 	<ul style="list-style-type: none"> - identical cell cycle successes (from objectives 2, 5, and 8) - identical tissue cycle failures (from objectives 3, 4, and 7) 	C) Temporal sequence data

Note. **“Performance” data refers to in-game success, near failure, far failure, and total failure

The totals, ratios, averages, last objective played, time series, and isomorphic sequence (Table 3) are computational lenses applicable to each of the base performance data types (success, far failure, near failure, and total failure). These provided data along three themes: context-specific totals (theme A), compound indices that are standardized for comparison between groups (theme B), and temporal sequence data (theme C). To clarify these data, definitions of the base performance types of success and shades of failure follow.

“Base Performance” Feature Definitions

For practical engineering of these features, “nuanced performance” was defined with one success type and three base failure categories: success, near failure, far failure, and total failure. Success was defined as the collection of the correct biological material at the end of a cycle. Near failure and far failure were definitions of failures which resulted from a detailed mapping of potential play actions and outcomes in a given cycle of play. These types of failure were distinguished after overall failure (as one broad umbrella) demonstrated no relationship to learning in pilot research (Owen et al., 2012); thus, nuances of failure became the next natural exploration. (The discussion and conclusion chapter dives more deeply into theories of play, failure, and learning in consideration of analysis results.)

To begin this examination, all outcomes of the start-treat-collect cycles of the game were mapped. The *Progenitor X* cycle involves populating an initial grid with the right kinds of cells (*start*), transforming those cells into a target cell/tissue (*treat*), and *collecting* the correct cells for the next cycle of the game. The cycles can unfold in several ways. First, players are guided to populate the grid with the right kind of cell (green check, Figure 18). After this population, the cycle can end in three results: collecting the right cell (success), collecting the wrong cell

(failure), or over-manipulating the cells so they die (the Ph of the culture becomes toxic, and the cycle results in failure).

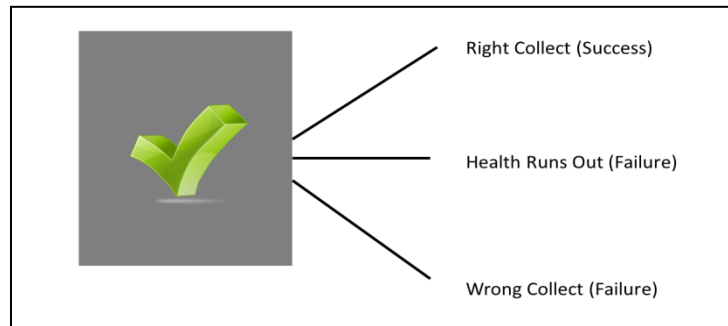


Figure 18. Progenitor gamespace - correct initial grid population.

Second, a player could have also initially populated the grid with the wrong cell (red X, Figure 19). In this case, there are two options for ending the cycle: collecting the wrong cell (fail), or overmanipulating the cells until the Ph levels (health) becomes toxic (fail).

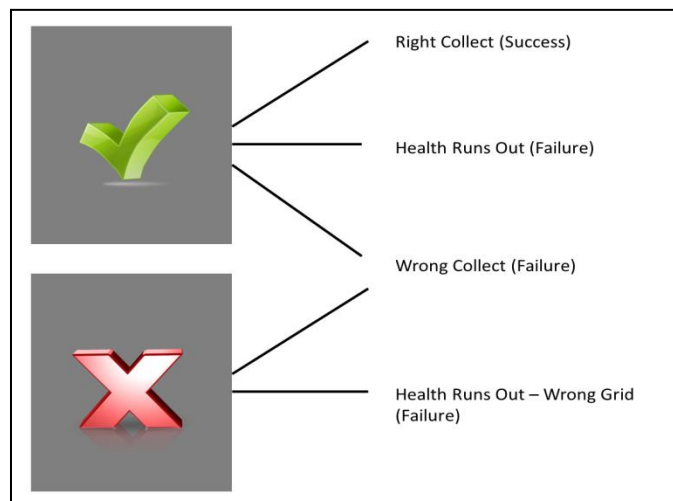


Figure 19. Progenitor gamespace - incorrect AND correct initial grid population.

With reflection on the possible outcomes of play cycles, distinction of the varying degrees of player compliance with instructional cues (e.g. flashing buttons & in-game narration) guided operational definitions of different kinds of failure. The concepts of “near” and “far failure” were then developed (Figure 20) to describe failed play in accordance with the suggested play path (near) and at odds with the suggested path (far). Three possible player outcomes for *Progenitor X* cycles were grouped as 1) correct collection (successful); 2) correct set-up but health runs out (“near failure”); and 3) incorrect setup and/or incorrect collection (“far failure”).

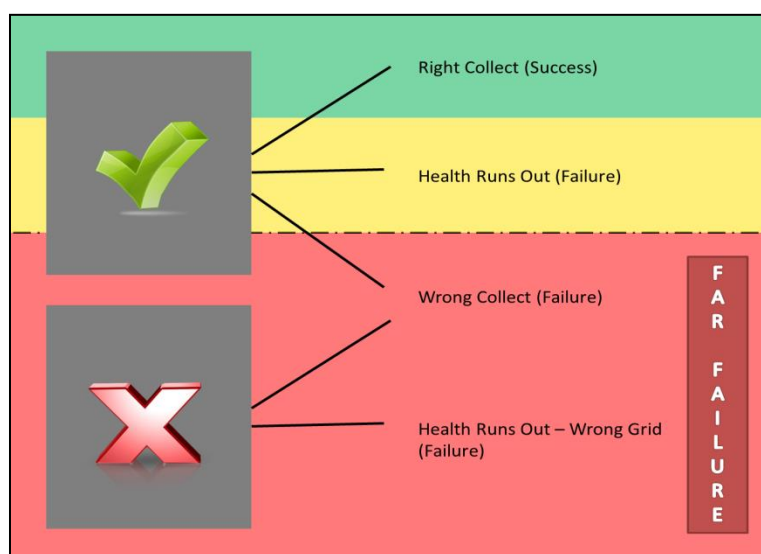


Figure 20. “Far failure” in the *Progenitor* gamespace.

Thus, three nuanced performance features were defined: success, near failure, and far failure. For the purposes of this investigation, it should be noted that these “success” and “failure” labels are simply operational definitions of actions in the gamespace. All feature definitions here are made according to the original learning game design, which complied with content experts’ vision of an expert pathway through the game’s procedural laboratory mechanics. This was reflected in the model-scaffold-fade (Collins, Brown & Newman, 1990)

structure of cell and tissue play progression, which assumed success through this sequence to be an optimal pathway. This “expert pathway” definition is solely used in this study for clearly labeling data features of success and failure – this research, however, does not assume that this intended play pathway is the optimal one. It merely uses these embedded criteria of game performance to label data features.

Progenitor X Data Feature Output

In using the computational lenses (Table 3) of feature distillation, the first step was to determine the basic “totals” information. Each of these four base performance types (success, far failure, near failure, and total failure) were aggregated across whole game session. Then, each of these performance labels were also identified based on the base cycle type (leveraging ADAGE “virtual context”). Table 4 shows the resultant performance categories. In tissue cycles, the only way to fail is through a grid destroy (near failure), so far failure is not an option. Also, organ cycles were extremely simple in the game, and were more for narrative sake than for demonstrating a skill. Since they were essentially impossible to fail, and only one interaction (success) was required per cycle, organ cycles were not included here as relevant for nuanced failure information. Thus, the next kind of “totals” amassed were for cell success, cell near failure, cell far failure, tissue success, and tissue near failure (Table 4).

Table 4

Data Feature Labels – Performance Types Merged with Cycle Types

	Success	Near failure	Far failure
Cell	Cell success	Cell near failure	Cell far failure
Tissue	Tissue success	Tissue (near) failure	

As a last step in gathering totals for each feature performance type, the Table 4 categories also could be aggregated by objective. This was important in giving context-specific performance information. Table 4 below shows cycle-type performance indices mapped to each objective. Objective 6 is omitted from the chart below, because it was the only organ-exclusive level, and existed for narrative cohesion rather than skill demonstration. Only cycle types present for each objective were identified; for example, Objective 1 only contained cell cycles, so tissue performance was not relevant. Therefore, this next “totals” category is represented in Table 5, listing each kind of cycle performance per objective (Objective 1 cell success, Objective 1 cell near failure, Objective 1 cell far failure, Objective 2 cell success, etc.).

Table 5

Data Feature Labels – Cycle-Based Performance for Each Objective

	Cell success	Cell near failure	Cell far failure	Tissue success	Tissue (near) failure
Objective 1	✓	✓	✓		
Objective 2	✓	✓	✓		
Objective 3				✓	✓
Objective 4				✓	✓
Objective 5	✓	✓	✓	✓	✓
Objective 7				✓	✓
Objective 8	✓	✓	✓	✓	✓

For the final feature output, success, near failure, and far failure were assembled on each level discussed: aggregate game totals, aggregate game totals by cycle type (Table 4), objective-specific totals, and objective-specific totals by cycle type (Table 5). This constituted the “totals” category of computational lens in Table 3. These totals then became a basis for application with

all other computational lenses in the chart: averages, ratios, last objective played, and time series (Table 3).

Ultimately, based on the computational lenses presented (Table 3), 10 core categories of features were created (Figure 21). Each feature was taken for success and nuanced failure types for each *Progenitor* objective, as well as for core play information like time elapsed, number of cycles played, number of objectives completed, and game completion (see Table 3 for examples of each). An average of 19 features per color-coded category type (Figure 21) was generated, creating a feature count of 194. Taken per student (n=110), this came to an aggregate matrix of 21,340 telemetry data cells (Figure 22). These final indices resulting from the feature engineering of this chapter served as a foundation for the entire dissertation arc of analyses.

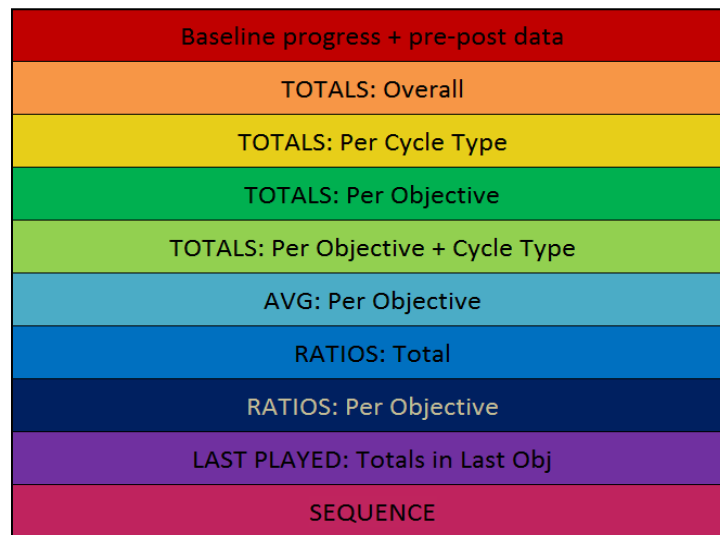


Figure 21. *Progenitor* feature distillation categories by color-coded label.

ID	last objective played	subobjectives (last obj played)	rounded objective number (last obj played)	seconds played (last)	total ONLY	total cycles	total success	total failure	total time (sec)	total failure	total success	total time (min)	total time (hr)	total time (day)	total time (week)	total time (month)	total time (year)
c_1	1	1	1	100	1	100	0	0	100	0	100	1.67	0.00	0.00	0.00	0.00	0.00
c_1	2B	4	2	191	2	278,92269	5	2	279	37	34	4.65	0.00	0.00	0.00	0.00	0.00
c_1	3C	3	5	199	4	265,05714	14	6	4	2	13	7	2	5	14	0.7777778	0.26
c_2	4	4	4	191	2	192,09895	5	3	2	7	5	3	3	3	0.75	0.25	0.18
c_3	5C	5	5	193	1	151,95995	5	2	1	1	15	13	1	12	2.5	144444444	0.66
c_4	3	5	3	157	3	125,33333	3	3	3	7	5	2	1	1	0.6666667	0.4	0.11
c_5	finished	5	5	192	0	0.222	36	5	0	5	5	0	0	0	0	0.00	0.52
c_6	5C	5	5	193	1	78,93333	9	1	0	3	2	0	2	0.4	0.2222222	0.11	0.20
c_7	4	5	4	191	0	74,96667	7	8	0	8	9	1	0	1	0.25	0.9666667	0.04
c_8	7	11	7	137	0	93,93334	13	9	5	7	2	0	2	0.287403	0.839338	0.09	
c_9	3	3	3	121	5	193,64668	14	5	7	27	22	2	20	7.3333333	4.4	1.00	
c_10	5	10	6	166	1	172,57436	8	4	1	3	5	1	0	1	0.6666667	0.1	0.06
c_11	5C	5	5	193	0	105.5	15	6	0	6	8	2	0	2	0.4	0.2222222	0.12
c_12	5C	5	5	193	0	91,66667	10	0	0	0	0	0	0	0	0.00	0.00	0.00
c_13	10	10	8	195	0	96,99999	16	7	0	7	8	1	0	1	0.125	0.0666667	0.06
c_14	5C	5	5	195	1	186,44422	8	9	1	4	6	1	0	1	0.5	0.1000000	0.06
c_15	finished	10	8	199	1	101,95228	14	8	1	7	9	1	0	1	0.125	0.0625	0.05
c_16	4	6	4	164	0	121,66667	12	0	0	0	11	11	0	11	2.75	0.8333333	0.76
c_17	5C	5	5	192	4	105,48774	5	7	4	3	9	2	0	2	0.4	0.25	0.13
c_18	7	11	7	165	0	72,363473	10	2	0	2	2	0	0	0	0.00	0.25	0.18
c_19	5A	7	5	194	0	49,66667	7	2	0	2	2	0	0	0	0.00	0.40	0.29
c_20	5B	8	5	191	0	49,33333	8	0	0	2	2	0	2	0.4	0.25	0.15	
c_21	5B	8	5	171	0	79,75	12	3	0	3	7	4	0	4	0.8	0.5	0.39
c_22	5C	5	5	195	0	137,73768	25	3	3	0	8	5	0	5	1	0.25	0.21
c_23	10	10	8	197	0	125,23078	13	3	3	4	1	0	1	1	0.125	0.0666667	0.08
c_24	5C	5	5	198	1	78,66667	8	2	1	1	5	3	1	2	0.6	0.3333333	0.20
c_25	2B	4	2	160	2	201,36667	2	2	2	7	6	3	2	2	0.5	0.25	0.63
c_26	5C	5	5	191	0	165,49475	17	11	2	9	25	14	0	10	2.9	1.9999999	0.60
c_27	11	10	8	143	0	112,93475	12	2	0	2	4	2	0	2	0.25	0.1333333	0.08
c_28	5A	7	5	199	0	78,96667	7	8	0	5	6	1	0	1	0.2	0.1428571	0.06
c_29	10	7	7	157	0	77,53333	10	1	0	10	10	0	0	0	0.00	0.143	0.01
c_30	finished	10	8	128	0	73,96667	14	0	0	1	1	0	1	1	0.125	0.0625	0.05
c_31	5C	5	5	192	1	123,43333	4	1	2	0	11	11	0	11	2.8	1.4285714	0.44
c_32	5C	5	5	195	0	87.25	8	5	0	5	6	1	0	1	0.2	0.1111111	0.06
c_33	5B	8	5	184	1	103,04228	10	5	1	4	10	8	2	6	1.6	1	0.41
c_34	5C	5	5	192	2	133,33669	10	11	2	9	11	0	0	0	0.00	0.20	1.22
c_35	5C	5	5	195	0	122,88254	15	3	0	3	5	2	0	2	0.4	0.2222222	0.10
c_36	finished	10	8	165	0	124,70742	20	3	0	3	3	0	0	0	0.00	0.38	0.15
c_37	finished	10	8	131	0	101,95	9	8	11	11	10	0	0	10	2.8	1.39	0.61
c_38	finished	10	8	163	2	122,625	23	4	2	2	4	0	0	0	0.00	0.50	0.15
c_39	finished	10	8	132	0	120,33895	8	10	7	7	15	5	0	5	0.625	0.3125	0.18
c_40	finished	10	8	242	1	227,45074	24	9	1	8	19	10	1	2	1.25	0.6999999	0.25
c_41	finished	10	8	140	3	131,45074	18	7	3	4	8	1	0	1	0.125	0.0625	0.04
c_42	finished	7	11	7	242	1	240,0568	63	10	3	7	36	25	25	3.7428571	2.3083333	0.63

Figure 22. Progenitor full feature distillation per color-coded feature category (partial view)

New Progenitor Feature Analysis: Methods and Results

These new measures of performance were designed to help give more nuanced insight into play patterns connected with learning. By creating more points of sequential comparison between learners, these new indices of success, near failure, far failure, and total failure afforded insight more detailed divergence and convergence in learner play patterns. The three data themes in Table 3 – “A” (context-specific performance), “B” (compound indices), and “C” (temporal sequence) – connected directly with analysis goals of visualizing performance trajectories, measuring association and comparing features in learner groups, and understanding performance across identical play cycles.

To achieve these analysis goals, three statistical analysis strategies were used. For the first analysis goal of visualizing performance trajectories, descriptive statistics were employed to visualize new feature trends (success and failure over time relative to learning) through basic scatterplots, comparative graphs, and time-series charts. Features from data theme “A” and “C”,

were visually mapped to show success and shades of failure in different sequential objectives. These were partitioned by high- and low- achieving learner groups for contrast, and took into account information like game completion.

The second and third analysis strategies were comparison and association of the new data features, though nonparametric correlation (Spearman's Rho) and mean contrasts (ranked mean Wilcoxon tests). Multiple comparisons were accounted for through controlling of False Discovery Rates (Storey, 2002); the p-values shown here were evaluated for significance using the QVALUE statistical package in R (Dabney & Storey, 2004). All adjusted p-values are thus called q-values, or "*q*", in the results below. These statistics worked largely with data themes "A" and "B", connecting new compound indices and context-specific performance features with learning outcomes. Used in Wilcoxon contrasts, the basis for quartile learning groups (i.e. data collection methods and full player pool) are described in greater detail in Chapter Three. Essentially, based on a pre-post assessment on regenerative biology (developed with content experts), they are made up of two groups: *Progenitor* players with the greatest positive change in score, and players with the lowest change in score. The upper quartile consists of 33 players, and the lower quartile consists of 41 players. "UQ" is an abbreviation used throughout the dissertation for the upper quartile of learners, and "LQ" stands for lower quartile of learners. These only refer to learner groups (as determined by pre-post gains) – no other kinds of quartile groups are discussed in this dissertation. For all correlation and non-quartile analyses in this chapter, N=110.

The implementation of these analyses had several possibilities for sequence; one order flow of analyses is represented in Figure 23, connected with the corresponding data themes. This flow started with time series graphs along data theme C (upper left, Figure 23). The

representation of these data in descriptive sequential line graphs launched deeper investigation of sequential trends, and inform the use of bar graphs, scatterplots, and two-sample Wilcoxon tests to contrast learner groups (data themes A & B). To corroborate these quartile-based inferences, Spearman's correlation was conducted on relevant features in relationship to learning outcomes (also A & B). With any data applicable for comparing identical, repeated cycles (e.g. cell success in objectives 3, 5, and 8), deviation between the cycles was visually mapped (Figure 23).

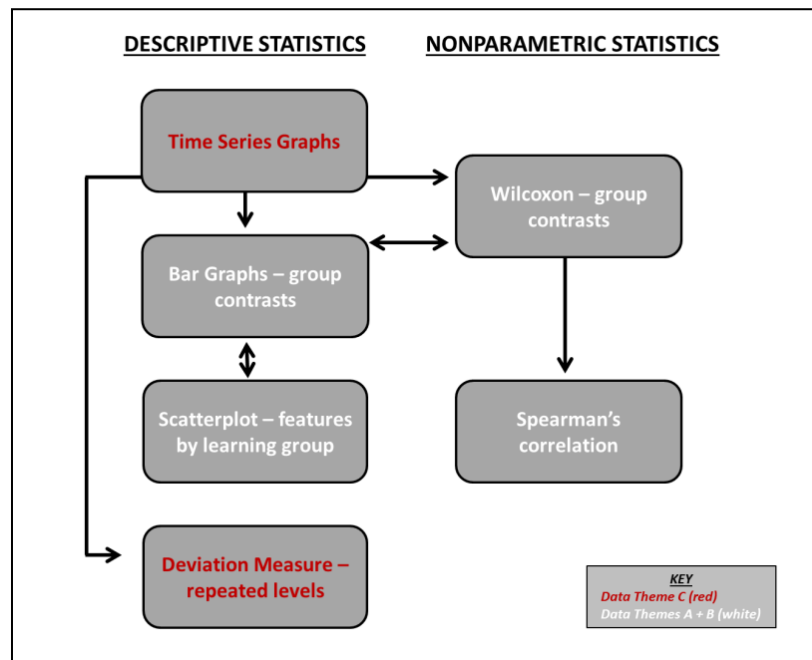


Figure 23. A sample flow sequence of analyses, beginning with time series graphs.

Discussion of Analysis Results

When analyzed with this methods schema, results yielded findings along three main trends of play. Trend One described overall play progress metrics and success in relationship to learning, while Trends Two and Three broke open the construct of “monolithic” failure into meaningful subtypes whose relationship to learning evolved with the game context.

Trend One: Game Progress Connected with Learning

The first trend of findings characterized forward movement in the game, including overall success, time, and progress patterns. Broadly, objectives completed and success in the game were connected with learning gains. A correlation between objectives completed and learning gains (Table 6) revealed a positive relationship ($r=.272$; $q=.018$), with a significant difference ($q=.044$) between the quartiles in number of objectives completed (Figure 20). Because it is possible to quit and restart the game, number of “objectives added” was considered in addition to objectives completed, as was the objective number of the “last played objective”. Both of these progress measures also demonstrated positive correlation with learning gains (Table 6), as well as significant differences between the quartiles (Table 7). Time elapsed, notably, was not correlated with learning gains nor showed any significant difference between the quartiles (Table 7). It seems the upper quartile was making comparatively efficient progress in the same time frame as the lower. Success as well as game progression mattered; the number of successful cycles in gameplay was positively correlated with learning outcomes ($r=.216$). Boss level success was also measured (since it is a clear summative assessment level in the ADAGE infrastructure), and also found to be positively correlated with learning ($r=.223$; $q=.033$). Thus, game success (overall and in the boss level) and overall progress (e.g. number of objectives completed) were positively associated with learning.

Table 6

Game Progress and Success in Progenitor X: Correlation

Trend	Feature	Correlation (vs. learning outcomes)
Progression	Objectives added	$r=.272; q=.018$
	Objectives completed	$r=.269; q=.018$
	Last played objective	$r=.257; q=.018$
Success	Total success	$r=.221; q=.033$
	Boss level success	$r=.223; q=.033$

Table 7

Game Progress and Success in Progenitor X: Contrast Between Quartiles

Feature	Upper Learner Quartile Average (n=33)	Lower Learner Quartile Average (n=41)	Significance
Objectives Added	7.0	6.0	$q=.044$
Objectives Completed	6.4	5.3	$q=.044$
Total Seconds Played	1453	1428	<i>none</i> ($q=.379$)

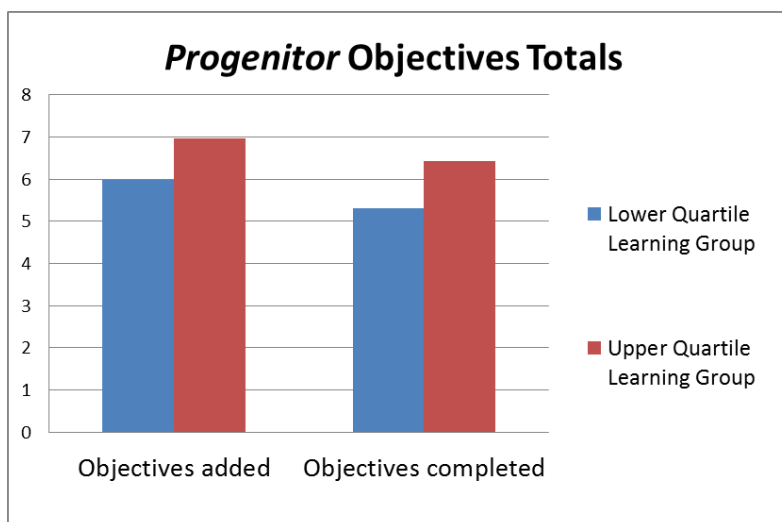


Figure 24. Progenitor average objective progress for each quartile.

Final Trends: Breaking the Monolith of Failure

The next results trends, branching into two currents of findings, supports two inferences about failure in gameplay: first, that differentiating kinds failure is important; and second, the relationship of each kind to learning changes with gameplay context. In short, kinds of failure – and, in turn, the changing context of failure types – matter for learning impact. A primary finding opened deep insight into this idea: failure, taken as an overall game total, had no statistical significance ($q=.359$) in connection with learning gains (either in correlation or in upper/lower quartile contrast). Thus, “monolithic” failure had no relationship with learning in *Progenitor X* play. This opened deeper investigation into kinds of failure and learning in the gamespace.

Far failure and near failure were two kinds of failure that could occur in the *Progenitor* cell cycles. In tissue stages of the game, it was also possible to fail through running out of health (also termed a “grid destroy”, because the grid of the in-game petri dish implodes when the cells run out of health and die). Tissue failure was thus another kind of failure that was distinguished

in these features, along with cell far failure and near failure. Looking at each kind of failure during each chunk of gameplay uncovered two main failure trends.

Trend Two: Tissue and Mid-level Failure

Breaking failure down into tissue failure provided insight into play patterns. The tissue mechanic, a Tetris-like shape manipulation, was different from the cell skill and represented an important step in the game's regenerative medicine procedures. This was an important mechanic both for keeping the player engaged in play well through mid-game, as well as for exposure to biology content (in illustrating more of an organ regeneration process). Therefore, understanding student performance in tissue cycles gives insight into sustaining play progression and optimizing experience with the academic content.

In particular, tissue cycles in mid-game levels (like Objective 5 and 7) turned out to be critical points for both learning and game completion. Objective 5 was the first level presenting advanced cell and tissue cycles together (instructional cues having been faded out in earlier levels). Tissue failure was connected with the lower learning group; in Objective 5, it was twice as high for the lower quartile of learners than it was for the upper quartile (Figure 26). Tissue mastery in Objective 5 was correlated positively with game completion ($r=.247$; $q=.030$), and positively associated with learning in Objective 7 ($r=.241$; $q=.028$). (Interestingly, tissue *failure* in Objective 7 had no relationship to learning.) Tissue performance in this mid-level especially influenced game quit points, as the histogram below shows (Figure 25). Displaying last objective played, the chart shows that most students either finished the game (Objective 8), or dropped out in Objective 5 – the compound cell-cell-tissue level. Interestingly, just taking chunked data by category in each objective, it's clear that tissue failure in Objective 5 sets the two learning quartiles apart (Figure 26). Identified with this baseline failure data, this phenomena of tissue

attrition opens up more nuanced questions about types of failure in this critical middle level, explored further in the detailed sequential investigation of Chapter Five's Markov modeling.

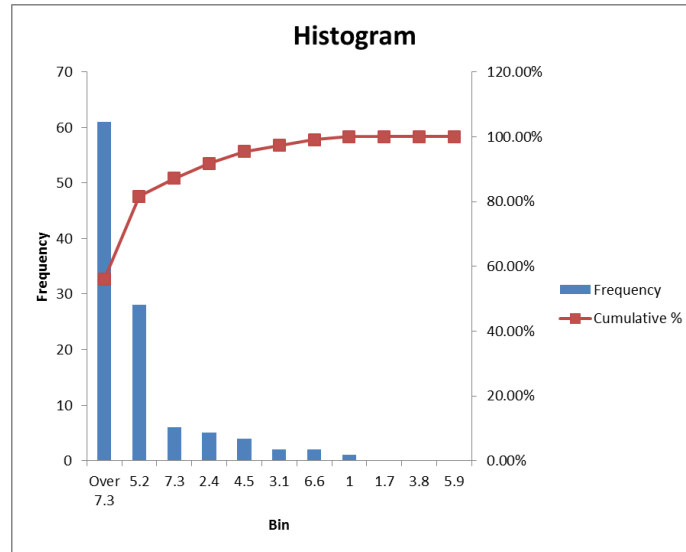


Figure 25. Histogram of transition from cell to tissue in mid was critical drop-off point.

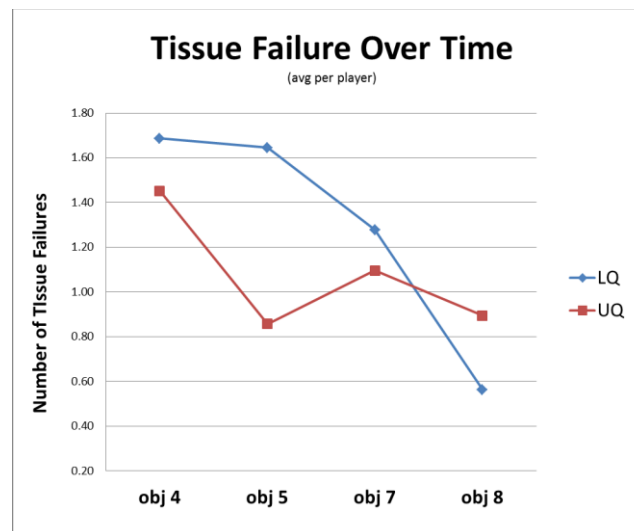


Figure 26. Tissue failures over time with the upper and lower quartile.

Trend Three: Far Failure, Learning, and Evolution Throughout Play

In addition to tissue failure, far failure (specific to cell cycles) mattered for learning as well. In overall play patterns, far failure proved to be a construct that showed significance for learning and game progress. Generally speaking, far failure had negative connection with learning and play progression, with finished players only experiencing 37% far failure (out of total failure) versus non-finished players with 63% (Figure 27). Another representation of far failure's relationship with play and learning can be seen in Figure 28, which shows a similar trend for both non-finished and lower learning quartile players.

Far Failure and Game Progress

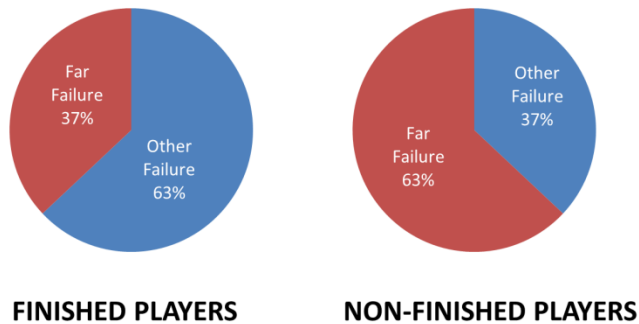


Figure 27. Far failure and game progress.



Figure 28. Far failure per objective, averaged for low learning quartile and nonfinished players.

Looking at a more nuanced narrative, however, far failure had critical variations in relationship to learning depending on the specific game level. Early game far failures appeared to negatively impact learning and game progress; astoundingly, this relationship reversed completely by late game levels, where far failure actually showed a *positive* relationship to learning.

Table 8

Far Failure Trends in Progenitor X: Contrast Between Quartiles

Trend	Feature	Upper Learner Quartile Average (n=33)	Lower Learner Quartile Average (n=41)	Significance
Early Game	Early far failure (Obj 1,2,5)	0.1	0.9	$q=.044$
Late Game	Objective 8 far failure	0.3	0.0	$q=.044$

Table 9

Late Game Far Failure in Progenitor X: Correlation

Trend	Feature	Correlation (vs. learning outcomes)
Late Game	Objective 8 far failure	$r=.217; q=.035$

Evidence of this trend emerged with visualization, correlation, and quartile contrast. In early levels of the game, far failure was negatively connected with learning (Table 8). This trend throughout objectives show that the upper and lower quartile differed significantly ($q=.044$) in early game far failure averages. The lower group of learners had nearly ten times the far failure (on average) than the UQ in early game (Figure 8). The descriptive trend in Figure 29 also shows that far failure was higher for the lower quartile, particularly in Objective 1.

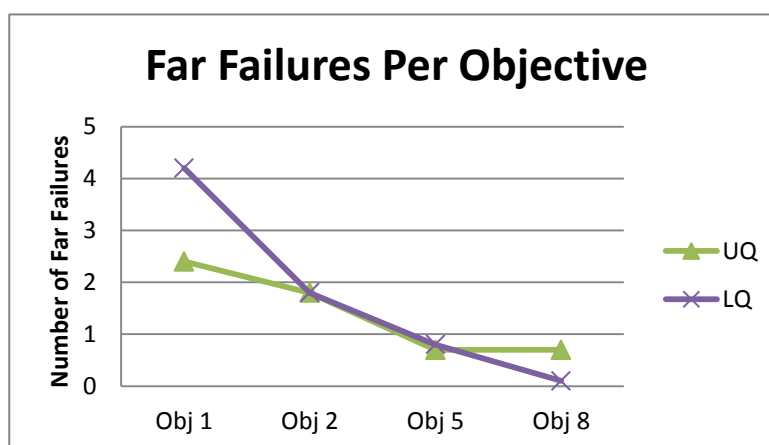


Figure 29. Far failure per objective, averaged for upper and lower quartile learners.

Looking at far failure's role beyond this point, however, paints a very different picture. Interestingly, the lower quartile far failure steadily decreases as play moves forward, but upper quartile far failure actually increases in frequency from mid to late game (Figure 30). Thus, far failure in *later* game objectives actually becomes characteristic of higher learning gains. Correlation supports this pattern (Table 9), showing that far failure in Objective 8 actually shows a *positive* relationship to learning ($r=.217$; $q=.035$), as opposed to the negative relationship in early game (Table 8). This movement implies that the meaning of far failure shifts from early training levels to later levels of mastery – and possible experimentation or strategy.

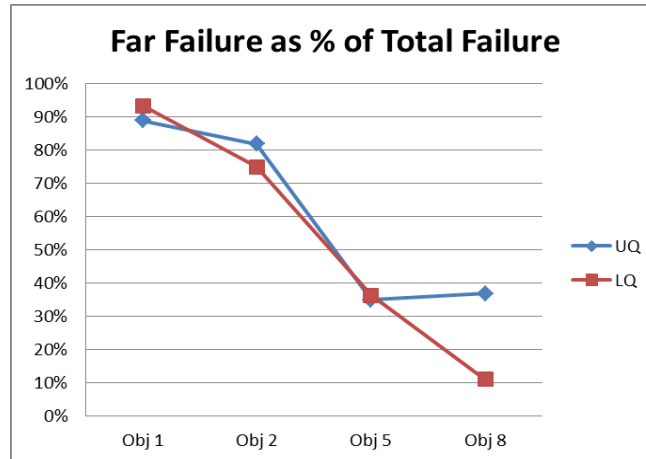


Figure 30. The ratio of (far failure / total failure) in each objective, calculated for lower and upper quartile players.

Summary and Conclusions

The findings of this chapter – based in differentiated failure in specific contexts – shows evolving trends of early, mid, and late game performance in relationship to learning. These trends shed light on this study’s driving research question on the relationship between in-game success and kinds of failure with learning outcomes. While overall success and game progression were positively related to learning, failure trends were much more nuanced. First, tissue failure in critical levels of scaffold-and-fade during mid-game proved critical in sustaining learning. Secondly, cell-based far failure showed interesting significance in early gameplay (negative relationship to learning), yet a positive relationship with learning later on during the boss level. Thus, in addition to providing a solid base of systematically engineered features for the entire dissertation analysis arc, this chapter sets the groundwork for the upcoming nuanced inquiry into fine-grained sequential patterns and far failure transgressive play behaviors of the following two analyses. These analyses build on these base findings to explore further the evolving role of kinds of failure in transgressive play and learning throughout the game.

Chapter Five:

Markov Modeling Of Learner Progression Through the *Progenitor* Gamespace

Introduction and Framework

In the overarching exploration of player learning through the lenses of game experience, this dissertation's second analysis studies learner choices (Lens C, Figure 1) in navigation through the game as a narrative progression (Lens A). Game design experts Jesse Schell and Matt LeBlanc identify game narrative as a fundamental dynamic of engaging game experience (Schell, 2008; Hunicke, LeBlanc, & Zubek, 2004). The term "narrative" here is not defined as "the telling of a prescribed, linear story", but more generally a "dramatic unfolding of a series of events" that is meant to support player roles, goals and agency (Schell, 2008, p. 109; Norton, 2008; Squire, 2011). In investigating sequential learner decisions around the designed milestones of play progression, this study seeks to understand patterns of forward movement, stagnation, and attrition within pivotal segments of the game as a cohesive arc of play. Specifically, its research question asks: how does organic play progression differ between groups of learners?

While Chapter Four's analysis looked at player actions chunked by types of cycles, and counted in terms of totals within each chunk, this study structures the interaction data differently. In order to effectively study real-time play progression, player actions needed to be itemized in real-time sequence at the cycle level. For example, if a player named Troy was to finish the game, we would look at every core cycle Troy completed from beginning to end -- with the specific lens of *sequence* from first cycle to last cycle (*not* aggregate grouping by cycle type or objective) to order the data. Troy's data for this analysis would, instead of a list of totals, look like an ordered sequence of events: 1) cell cycle, 2) cell cycle, 3) tissue cycle, 4) cell cycle, 5)

organ cycle, 6) cell cycle...and so on. This is wonderful on an individual student level – we could visualize a single player’s game progress with this information – but one challenge of this research lens was to find a method that was able to simply and clearly describe movement through the space for multiple players simultaneously. One such method, Markov chain modeling, can create visualizations of user movements from one cycle to the next for multiple players at once (c.f. Rabiner, 1989; Clark et al., 2012). For this reason, Markov modeling was the method of choice for investigating this research question in the *Progenitor X* gamespace. To understand play progression in relationship to learning outcomes, the research design builds and compares two Markov models: one for the students with the highest biology learning gains⁸ (upper quartile), and one for students with the least improvement (lower quartile).

Made possible by ADAGE data, Markov modeling works in this study by taking different points of gameplay progression and identifying them as states. In this case, the Markov states are anchoring points of progress that help identify where the player was in the context of game completion. Example states in *Progenitor* are Objective 1 cell cycle, Objective 2 cell cycle, Objective 3 tissue cycle, etc. A Markov model then shows the probability that players will move from one state to the next. For example, a model might show a 75% probability that students would start in Objective 1 cell cycle, and move directly to an Objective 2 cell cycle. It might also say there is a 20% probability that students starting from that same Objective 1 cycle would end up repeating it for their next move. Through this temporal probability modeling, Markov chains produce a whole matrix of probabilities of moving from one given state to the other (a transition matrix). Using first-order Markov modeling in this study provided detailed probabilities of each movement in the gamespace from one state to another. Thus, this analysis deemed to understand play progression by examining probabilities of movement from one state to the next, and

⁸ As measured by pre-post biology survey data – see Chapter 3 for details.

contrasting these transition matrix probabilities between the two learner groups. Examining these differences carefully enabled the identification of nuanced sequential play trajectories most characteristic of learning.

Overall, this Markov analysis creates new, nuanced telemetry indices for consideration in relationship to learning, reveals organic findings consistent with the themes of Chapter Four, and enriches understanding of these trends for final investigation in the third and last analysis. To be clear, the Markov model is a descriptive analysis which allows indices from Analysis I to be visualized in higher temporal resolution. The following pages of this chapter will discuss Markov findings along three main trends, deepening insight into three dissertation-wide play themes: early game tutorial attention, mid-game scaffold-and-fade performance, and endgame strategic navigation.

Methods and Output

To most effectively mine the ADAGE data for contrasts between the learner groups, three sets of models were built with three different levels of resolution. Each first-order Markov model was built using the “Markov” algorithm (Berland, 2012) in NetLogo, a multi-agent modeling environment (Wilensky, 1999).

Each set of models was built based on the data from two groups: an upper quartile and a lower quartile of learners. This designation is based on a pre-post assessment on regenerative biology (developed with content experts, and described in greater detail in the ADAGE/*Progenitor* methods Chapter Three). Relative to this performance, the quartiles are made up of two groups: *Progenitor* players with the greatest positive change in score, and players with the lowest change in score. The upper quartile consists of 33 players, and the lower

quartile consists of 41 players. “UQ” is an abbreviation used throughout the dissertation for the upper quartile of learners, and “LQ” stands for lower quartile of learners. These only refer to learner groups (as determined by pre-post gains) – no other kinds of quartile groups are discussed in this dissertation. For all correlation and non-quartile analyses in this chapter, N=110.

Markov Model Set One: Base Resolution

The first set of Markov models were built with base progression data, creating a simple set of game states which designated progress through each objective. The cycle was chosen to represent this progress, because it is the smallest consistent unit of *Progenitor* gameplay. To create these states, each cycle type (cell, tissue, or organ) was listed in order of occurrence and corresponding objective (Figure 31). This information was then synthesized into simplified Markov labels for each cycle (right column, Figure 31). These simple features become the “states” for the Markov model (i.e. possible positions in the gamespace). From a given state (e.g. obj1_cell), a player could make one of three moves: repeat the cycle, move on to the next level cycle, or quit the game. The Markov model, then, maps the probability of each group in repeating, moving forward, or quitting immediately after a given cycle.

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

Objective	Cycle Type	Markov State Label
1	cell	obj1_cell
2	cell	obj2_cell
3	tissue	obj3_tissue
4	tissue	obj4_tissue
5	cell	obj5_cell
5	tissue	obj5_tissue
6	organ	obj6_organ
7	tissue	obj7_tissue
8	organ	obj8_organ
8	cell	obj8_cell
8	tissue	obj8_tissue

Figure 31. Detailed cycle sequence & simplified conversion to Markov state labels.

In this first model set, two Markov chains were built in NetLogo. The first took all sequential cycle activity of the upper quartile of students and built a probabilistic model of play progression. (The upper quartile and lower quartile groups are often referred to here as “UQ” and “LQ,” respectively. It should also be noted that all players were given generous time to finish at 60 minutes per session; 25 minutes was the average playthrough duration, with 40 minutes defining an upper limit of $+2\sigma$.)

When the data for each quartile were put into the Markov algorithm in NetLogo, two visual maps of transitions to and from each state were generated. Each shows a clear trajectory of play characterizing each group, shown side by side below. (Specific Markov results are visualized in greater detail in the findings section of this chapter.)

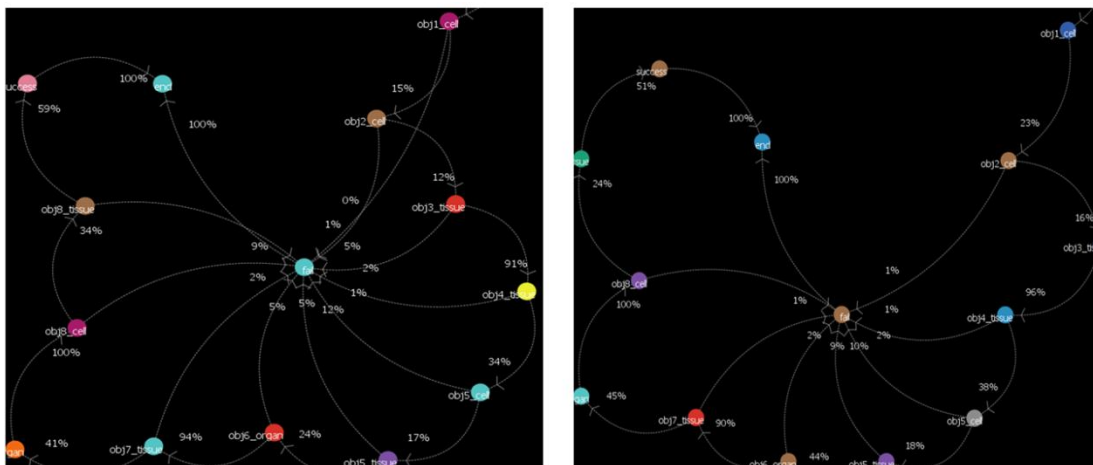


Figure 32. Base Markov model of the upper quartile (left) and the lower quartile (right) of learners.

Markov Model Set Two: Moderate Resolution

The first pair of Markov models gave clear mapping of play trajectory, with simple demarcation of repetition, progression, or quitting. With cycle repetition in particular, however, there are several different reasons players may repeat. Only in certain cases does it mean failure and re-trying; in particular objectives, players can move freely between cell, tissue, and organ cycles and repeat successful cell cycles as many times as they wish before choosing to move on. In order to better understand player choices during cycle repetition, then, the data resolution was intensified to include success and failure at the individual cycle level.

The second set of Markov models reflects this increased resolution, intended to give deeper insight into the simple, powerful results of the first model. The data for this moderate-level model takes state labels for each objective, just as the first models did, but this time incorporates success and kinds of failure for each cycle. For deep definitions of failure types,

please refer to Chapter Four, where the base features for analysis are described in detail. A cursory set of definitions include:

- “far failure” as the kind of failure that happens when the player is acting in direct opposition to instructional cues,
- “near failure” as a “softer” failure (see Figure 20) in which the player has started off a cycle correctly but has simply run out of health, and
- “tissue failure” as failure during a tissue cycle (there is only one way to fail in this case).

In this second, more detailed Markov, rather than just denoting objective number, the cycle labels include performance data. For example, if a player successfully completed a cycle in Objective 1, the label might be “objective 1 success”; if they experienced far failure, then the label would be “objective 1 far failure”; if near failure, then the cycle would be identified as “objective 1 near failure”. (Each of these is abbreviated in the actual model labels, shown below.) The states for this model start with an objective number, and combine it with every performance outcome possible for that objective⁹. Possible outcomes for cycles include: success, near failure, far failure, tissue failure, and failure via unfinished cycle (e.g. quit fail). The table below shows all combinations of information into state labels.

⁹ As detailed in Chapter Three, some objectives only have one kind of cycle available, and thus only include labels for the appropriate failure type. “Far failure” and “near failure” are specific to cell cycles, and “tissue failure” is specific to tissue cycles; Objectives 0, 1, and 2 are exclusively cell levels, and Objectives 3, 4, and 7 are exclusively tissue levels.

Table 10

Moderate-level Markov State Labels (Model Set Two)

	success	far failure	near failure	tissue failure	quit failure
Objective 1	1_S	1_FF	1_NF		1_QF
Objective 2	2_S	2_FF	2_NF		2_QF
Objective 3	3_S			3_TF	3_QF
Objective 4	4_S			4_TF	4_QF
Objective 5	5_S	5_FF	5_NF	5_TF	5_QF
Objective 6	6_S				6_QF
Objective 7	7_S			7_TF	7_QF
Objective 8	8_S	8_FF	8_NF	8_TF	8_QF

This new resolution of data was assembled and sequenced for the upper and lower quartile of learners, then put into two separate Markov models for contrast in NetLogo. The model output for each group is shown in Figure 33. Please note that findings from each model are visualized in greater detail in the results section below.

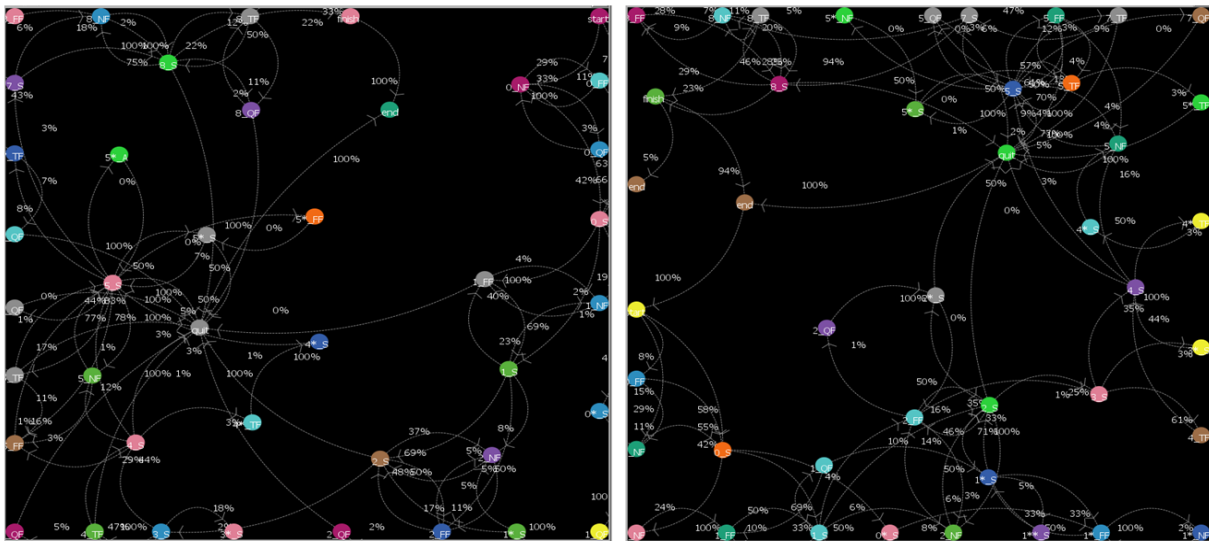


Figure 33. Moderate-resolution Markov model, with the upper quartile of learners on the left, and the lower quartile on the right.

Markov Model Set Three: Highest Resolution

Overall, the moderate-level Markov (model set two) was an excellent balance of preserving n-size per cycle and getting a sense of broad-stroke trends, while gaining informative resolution into repetition cycles. This is especially true for the *Progenitor* objectives which contain only one kind of cycle (either cell, tissue, or organ). However, some objectives contain all three. These “compound” levels are Objectives 5 and 8, which carry a cell-cell-tissue-organ sequence, but allow non-linear play (free range to move between these cycle types and repeat as often as desired). While making for great synthesis of gameplay skill, these two objectives present a challenge in clearly parsing the data for analysis. For example, with the moderate-level Markov data, an “objective 5 success” could mean a cell, tissue, or organ success (and at any non-linear point of the player’s choosing); an “objective 5 quit fail” could have happened during any of these cycles as well, and does not give information about how much of the objective content the player has completed. Therefore, specifically for more nuanced examination of these compound objectives, one last model set of higher resolution was created.

In model set three, compound objectives were examined more closely, and failure types were even further differentiated for maximum resolution. The state labels for these models consisted of objective, plus the kind of success or failure experienced in that cycle. Cycle outcomes for this model set were expanded to include: success, far failure type 1 (wrong cycle start), far failure type 2 (wrong cell collect), near failure, and incomplete cycle (quit fail). (Please, again, refer to Chapter Four for deep explanations of failure types.) For example, an Objective 1 cycle ending in far failure because of a wrong start would take the label “objective 1 far failure type 1”. (This is abbreviated in the final label notation, shown below.) To clarify different sections of compound objectives 5 and 8, the different cycle phases of each were

divided into subobjectives, and the data painstakingly labeled as such. A table of these subdivisions is shown in Table 11, with corresponding cycle type and new subobjective abbreviation.

Table 11

Compound Objectives Broken into Subobjectives

Objective	Cycle Type	Subobjective
5	Cell (Phase 1)	5A
	Cell (Phase 2)	5B
	Tissue	5C
8	Organ	8A
	Cell (Phase 1)	8B
	Cell (Phase 2)	8C
	Tissue	8D

State labels were then created using these new subobjectives and each cycle's outcome.

The final Markov state labels for each play cycle are shown in the table below.

Table 12

Detailed-level Markov State Labels (Model Set Three)

	success	far failure 1	far failure 2	near failure	tissue failure	quit failure
Objective 1	1_S	1_FF1	1_FF2	1_NF		1_QF
Objective 2	2_S	2_FF1	2_FF2	2_NF		2_QF
Objective 3	3_S				3_TF	3_QF
Objective 4	4_S				4_TF	4_QF
Objective 5 - A	5A_S	5A_FF1	5A_FF2	5A_NF		5A_QF
Objective 5 - B	5B_S	5B_FF1	5B_FF2	5B_NF		5B_QF
Objective 5 - C	5C_S	5C_FF1	5C_FF2	5C_NF	5C_TF	5C_QF
Objective 6	6_S					6_QF
Objective 7	7_S				7_TF	7_QF
Objective 8 - A	8A_S					8A_QF
Objective 8 - B	8B_S	8B_FF1	8B_FF2	8B_NF		8B_QF
Objective 8 - C	8C_S	8C_FF1	8C_FF2	8C_NF		8C_QF
Objective 8 - D	8D_S	8D_FF1	8D_FF2	8D_NF	8D_TF	8D_QF

Using these state labels, sequenced play data for the upper and lower quartile groups were put into NetLogo, and a Markov model created for each. These are shown in Figure___, and discussed in further detail in the “Results and Findings” section below. Please note that findings from each model are visualized in greater detail in the results section of this chapter.

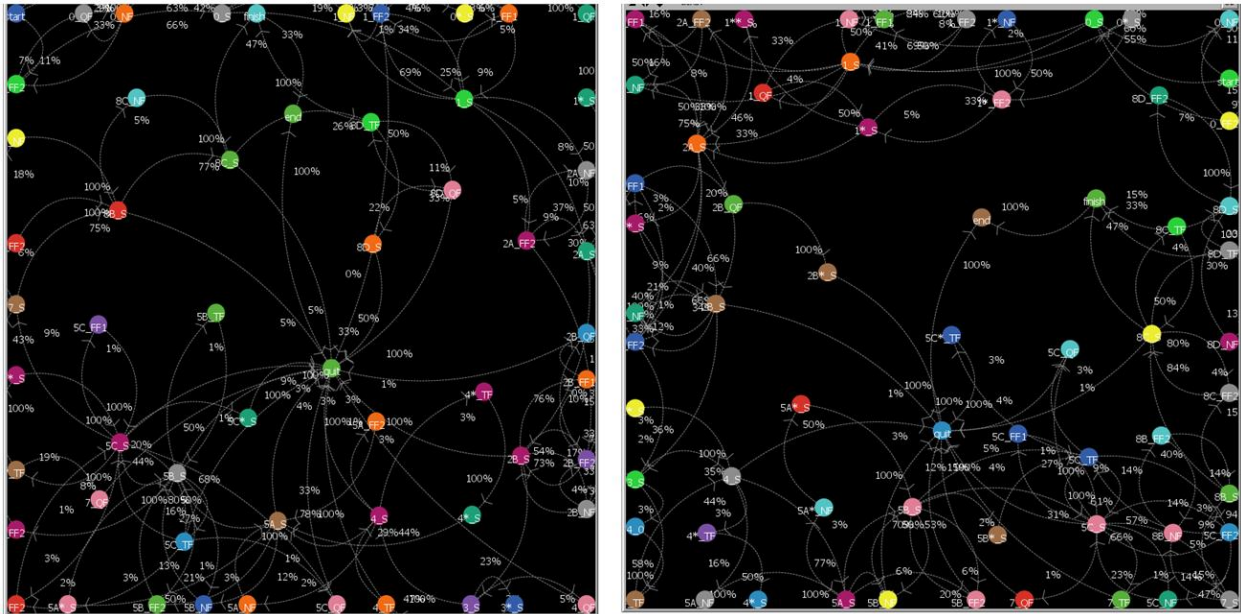


Figure 34. Highest-resolution Markov model output from NetLogo, with upper quartile on the left and lower quartile on the right.

Together, these three Markov sets support an informed understanding of play progression patterns characteristic of learning in *Progenitor X*. The base model shows simple, clear patterns of progress, stagnation, and attrition in the gamespace, while the two high-resolution models elucidate varieties of repetition and forward movement in nuanced phases of play. The second and third model sets convey essentially the same information, with the exception of compound Objectives 5 and 8, when the highest resolution Markov is used for nuanced insight. Thus, this

third model surfaces mainly in findings with Objectives 5 and 8, since it provided little new information for the other levels of play.

Results and Findings

This section will discuss the results and findings from the three Markov models, corroborated with descriptive and nonparametric statistical analyses of the models' state data. Results will be presented in themed groups, constituted by visualization and discussion of that section's findings. Together, the probability visualizations (Figures 34 through 40) report all results from the Markov models which meet the n-size threshold and cross-validation criteria detailed below. (Full model output can be seen in the transition matrices given for each set in the Appendix.) Reinforcing statistical analyses will be also discussed with each trend of findings in the upcoming section.

In interpreting the models, state transition matrices of each Markov pair (see Appendix) were contrasted to understand both broad and nuanced trajectory differences between the learning groups. It should be noted that this analysis focuses on the areas of *contrast* between the upper and lower learning quartile groups in order to differentiate play trajectories characteristic of the highest-achieving learners. To help distinguish a meaningful contrast between the upper and lower learner groups, a 95% confidence interval was performed on the probability *differences* between the two quartiles. Any difference in probabilities (UQ minus LQ) over 5% was considered in results, since this was the lower limit of the confidence interval.

A few heuristics were developed in the consideration of results, including a minimum n-size threshold per state and the existence of cross-validating evidence per trend. The need for interpretation thresholds has been recognized in similar mathematics, econometrics, and NLP Markov-based research (e.g. Zhao, 2010; Hansen, 2000; Lee & Kim, 1999). For results

consideration in this study, a player n-size minimum per state was necessary, because not all states of the models were required to play the game all the way through. For example, a player could move through the entirety of the game and *never* have a failure. That player's track might look like: obj 1-success, obj 2-success, obj 3-success...obj 8-success. This means that his/her play would not contribute to the count of students which had failures, and therefore would not contribute to the data used to calculate the model probability of actions starting from a failure state. Thus, in interpreting the Markov models, number of students represented in each state transition probability had to be considered (especially in a study of this modest n-size). Therefore, standards for minimum player n per state and cross validation were developed. All Markov results reported as findings have a minimum user n size (10 players) per source action, a cutoff point determined using the lower limit of a 95% confidence interval for the number of players contributing to each Markov state. Additionally, cross-validating evidence was essential to supporting each results trend, performed with statistical analysis (descriptive, correlation, or mean comparison) on the newly created nuanced Markov indices. Mean comparison of these new telemetry features was performed using a two-sample Wilcoxon ranked test, and ranked correlation was calculated with Spearman's Rho (both in SPSS). The resultant p-values have been evaluated for significance with the R Studio QVALUE package (Dabney & Storey, 2004), controlling for multiple comparison based on False Discovery Rates (Benjamini & Hochberg, 1995; Storey, 2002). All adjusted p-values are thus called q-values, or "*q*", in the results below.

These criteria were established to substantiate trends in that represent the broadest set of players, to help identify patterns most applicable to a larger *Progenitor X* audience. While not the emphasis of this study, examining of the individuals in the lower outliers could be valuable in other research; in conjunction with qualitative interviews and study, for example, it could

support an experimental design focused on singular students and more individual ethnographic patterns. Near failure could be a germane construct to investigate in specific case studies, since individual patterns in this kind of failure vary greatly and have been difficult to clearly capture on a collective level. The current study, however, has placed an emphasis on mining broader patterns in telemetry, and has been able to reveal strong aggregate far failure, success, and learning themes. In the results of this group Markov analysis, this chapter uncovers clear findings (corroborated by nonparametric analyses) in relationship to learning in three trends: early-phase failure ruts, mid-level tissue performance, and learning-supportive far failure in late game.

Results Trend I: Stuck in a Tutorial Rut

Trend I is comprised of results in early tutorial levels of the game, specifically Objective 1, that reveal contrasting patterns of repetition, failure, and success between the two quartiles. Together, these results support thematic findings that the lower quartile of learners were stuck in tutorial cycles, and experienced more frequent far failure at the tutorial game level. To show this, first the Markov findings relevant to the trend will be visualized and explained, and then corroborating statistics discussed.

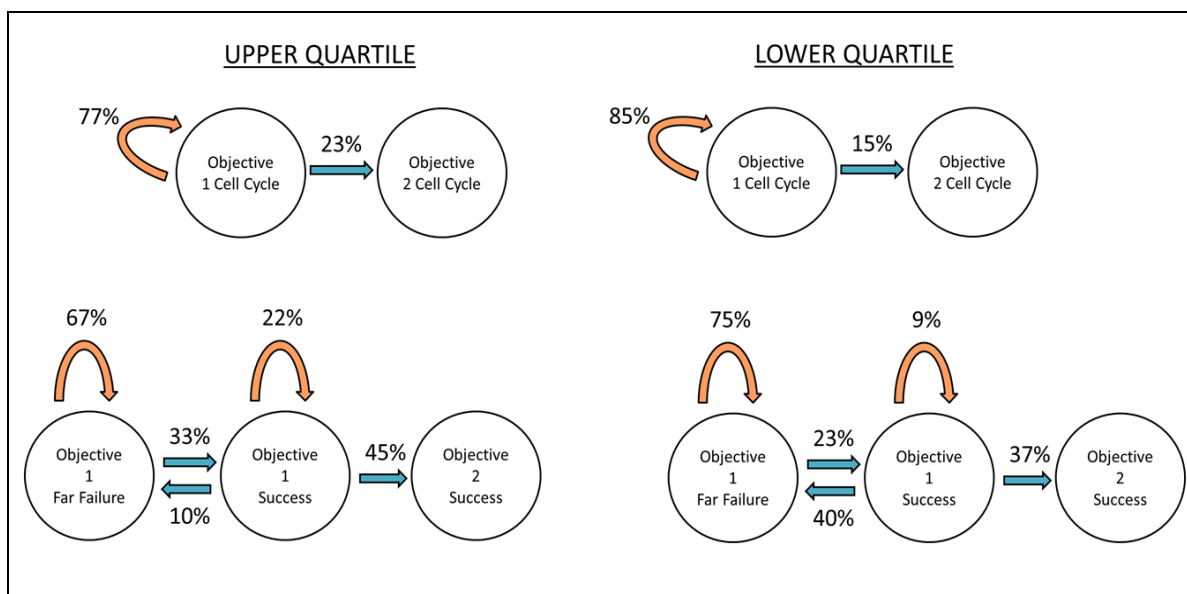


Figure 35. Visualized Trend I Markov findings for early-game failure rut; base model on top, and detailed model on bottom.

Trend I: Detailed Results and Visualization Key

Figure 35 above shows the core Markov findings in this trend. (For clarity, only Markov findings relevant to the trend and corresponding game objective are visualized in this section. Full model output is shown in the early sections of this chapter, and complete transition matrices can be referenced in the Appendix.) States are represented in the circle nodes, showing objective number and performance type. Probabilities of transition between nodes are represented by arrows with corresponding percents. Finally, each row of visuals shows results from a different set of Markov models. For example, the top level in Figure 35 shows results from the simplest Markov model. These display a 77% likelihood of repetition of Objective 1 cell cycles for the upper quartile, and a 85% likelihood in the lower. The probability that a given player will move from an Objective 1 cell cycle to an Objective 2 cell cycle is 23% for the upper quartile, and 15% for the lower quartile. The bottom row of results in the figure shows the more detailed Markov

results, starting with a 67% probability of repeating Objective 1 far failure for the upper quartile, with 75% for the lower. Moving right, the upper quartile (“UQ”) players have a 33% likelihood of transitioning to an Objective 1 success, which is 23% in the lower quartile (“LQ”). From this Objective 1 success state, players tended to go three ways: they either fell back into far failure (left-pointing arrow), repeated the success (orange arrow), or transitioned to an Objective 2 success. Thus, from a starting point of Objective 1 success, any given player in the UQ had a 10% likelihood of falling back to an Objective 1 far failure, a 22% chance of repeating the Objective 1 success, and a 45% chance of moving to an Objective 2 success. Lower quartile likelihood of these were 40%, 9%, and 37%, respectively. The following paragraphs discuss what these contrasting numbers imply for failure and success patterns most characteristic of learning.

Discussion of Trend I Results

The broadest patterns of the Markov models showed that lower quartile players had a higher chance of repeating and failing at Objective 1 cycles. The base Markov model, for example, showed that LQ (lower quartile) players repeated Objective 1 cycles more frequently than UQ players. The upper quartile had a 77% probability to repeat cycles in this level, while the lower had an 85% probability of repetition (8% higher). The more detailed Markov model revealed this repetition was mainly due to failure, showing that LQ players were more likely to have repeated far failure at this level (75% versus 67% in the upper group). Conversely, with repeated successes, the upper quartile was more than twice as likely to have one success followed by another in Objective 1 (22%), while the lower quartile only had a 9% probability of this success-success transition. The upper learning group also had a greater chance of recovery from far failure, having a 33% likelihood of moving from a far failure to a success (compared

with only 23% in the LQ). In contrast, the lower quartile was *four* times more likely to slide back into failure after a success, with a “Success → Far Failure” transition likelihood of 40% (versus only 10% in the upper group). Not surprisingly, the UQ also had a greater chance of succeeding at an Objective 1 cycle and moving on to Objective 2 (a 23% likelihood, versus only 15% in the lower quartile). Specifically, the upper group had a higher chance of moving from a success in Objective 1 to a consecutive success in Objective 2 (46%, contrasted by 37% in the LQ).

Strong evidence on early failure from Chapter Four supports these trends. First, far failure (wrong cell collects, specifically) in Objective 1,2, and 5 had were significantly different between the upper and lower quartiles (Table 8). The lower quartile had an average of 1, while the upper quartile had an average of 0. Descriptive trends in Figure 29 also show that high far failure early on is more characteristic of the lower quartile of learning, and is also a trend which forecasts poor game completion rates (Figure 28).

In summary, this trend of findings shows greater tutorial level stagnation and failure for the lower quartile. The lower group of learners here is characterized by repeated failure, and falling back into failure even after a success. By contrast, the group of greater learning gains had more frequent consecutive success, recovery from far failure with an immediately following success, and carryover of success from Objective 1 straight to Objective 2. Corroborated with statistical results, these patterns imply that far failure in the first objective had negative impact on game completion and learning.

Results Trend II: Synthesis Levels and Tissue Cycle Performance

Results Trend II highlights pivotal tissue performance differences in the learner groups throughout the game – particularly during their scaffold-and-fade, mid-game synthesis, and endgame boss level objectives. During the tissue cycle introduction in Objective 3, Markov data

showed learning curve trends emphasizing recovery from initial failure as a feature differentiating the two quartiles. By Objective 5, when all tissue “help” scaffolding had been faded out, and the tissue skill layered in sequence with the cell cycle skill in a synthesis of play mechanics, mastery of the tissue cycle characterized the higher learning group. In Objective 8 (the boss level), when a more difficult synthesized tissue cycle was embedded, learning curve behavior paralleled early objectives with better recovery from failure (not repetition of failure) characterizing the upper quartile group. Throughout play, quitting of the game immediately following a tissue cycle was a chronic pattern in the lower quartile, recurring in Objectives 5, 7, and 8. In the following paragraphs, these Markov findings will be visualized and discussed, and connected with cross-validating statistical analysis.

Detailed Sequence of Trend II Results

Detailed learner group performance on tissue-building phases of *Progenitor* differs consistently throughout the sequence of gameplay. The tissue cycle, which employs a Tetris-like puzzle mechanic using the building blocks of cells harvested in earlier objectives, first appears in Objective 3 of *Progenitor X*. In this heavily scaffolded objective, it is impossible to fail, as all parts of the UI are locked except for those enabling the correct action. Objective 4 is the first tissue objective in which it’s possible to fail, but an easy puzzle and ongoing text instructions serve as player support.

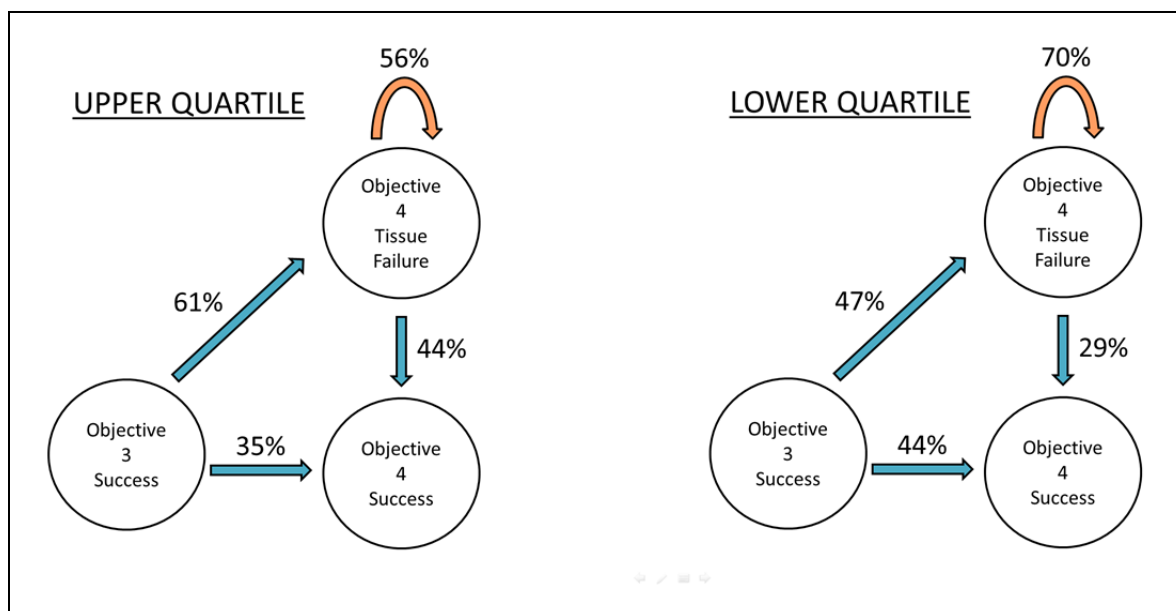


Figure 36. Trend II Markov results visualized for early-game tissue cycles (Objectives 3 and 4).

Figure 36 above shows the core Markov findings early in this trend, specifically introductory Objectives 3 and 4. Similar to the previous diagram, the objectives and performance are the nodes, and the transition probabilities are shown as numbers corresponding to the arrows. (Note once again that – for the purposes of clarity – only findings relevant to this trend and corresponding objectives are visualized; full model output is shown in the beginning of this chapter, and can also be seen in the full transition matrices supplied in the Appendix).

As illustrated in Figure 36, the first independent tissue cycle (Objective 4) reveals interesting learning curve patterns that differed between the quartiles. The upper quartile of pre-post learners actually tended to fail their first Objective 4 tissue cycle (61% likelihood of initial failure, versus only 47% in the LQ), but bounced back from it quickly with a consecutive tissue success (15% higher transition rate from tissue failure to tissue success than the LQ). The lower quartile, on the other hand, tended to have consecutive Objective 4 tissue failures (70% probability of repeating failure, versus only 56% in the UQ). Once again, the lower quartile of

learners seem to be stuck in an early failure rut, while their upper quartile counterparts recover more immediately from failure.

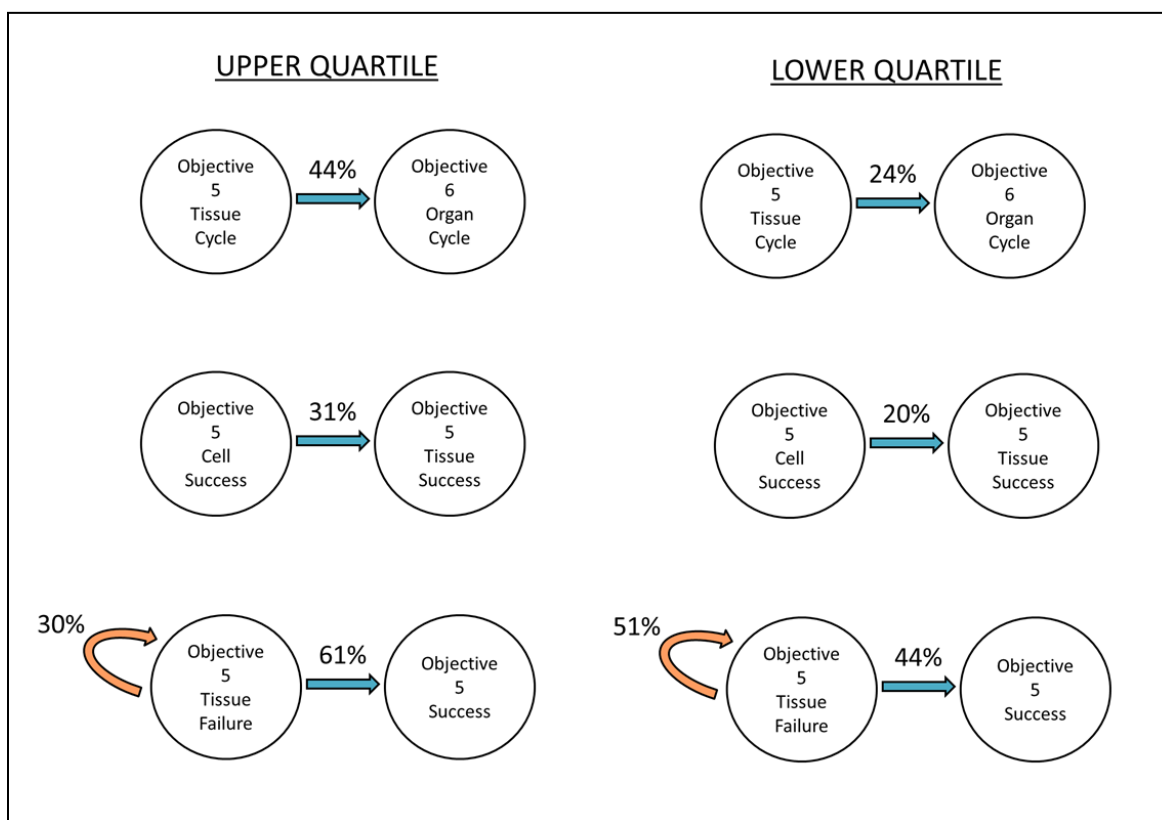


Figure 37. Mid-game tissue cycles visualized from Markov models for Trend II.

Moving to the next phase of tissue play, mid-level findings are shown above in Figure 37. This diagram again shows the objectives and performance as nodes, and transition probabilities as arrows. The first level of the diagram shows the simplest Markov information, the middle tier shows the second Markov set results, and the bottom row shows the most detailed model findings.

This upper quartile pattern of strong recovery from failure continues in Objective 5, where transition to success was again a hallmark of the group. Objective 5 contains wholly unscaffolded tissue puzzles alongside cell cycles which are equally unguided, thus fully

synthesizing the two main mechanics of the game into a pivotal mid-game level. Upper quartile players had the same resilience as in earlier tissue performance, showing a 17% higher rate than the LQ of going from tissue failure to consecutive success. Conversely, the lower quartile players once again tended to be stuck in a repetitive rut of tissue failure (51% chance of repeated failure, versus only 30% in the UQ). Descriptive statistics supports the trend of the lower quartile's tissue failure rut, showing that the lower quartile had twice as many Objective 5 tissue failures as the upper quartile (Figure 26). Significantly, tissue failure in this pivotal mid-game objective also had a negative correlation with game progression ($r=-.512$; $q=.015$). Success rates in Objective 5, however, were positively correlated with learning gains ($r=.205$; $q=.03$).

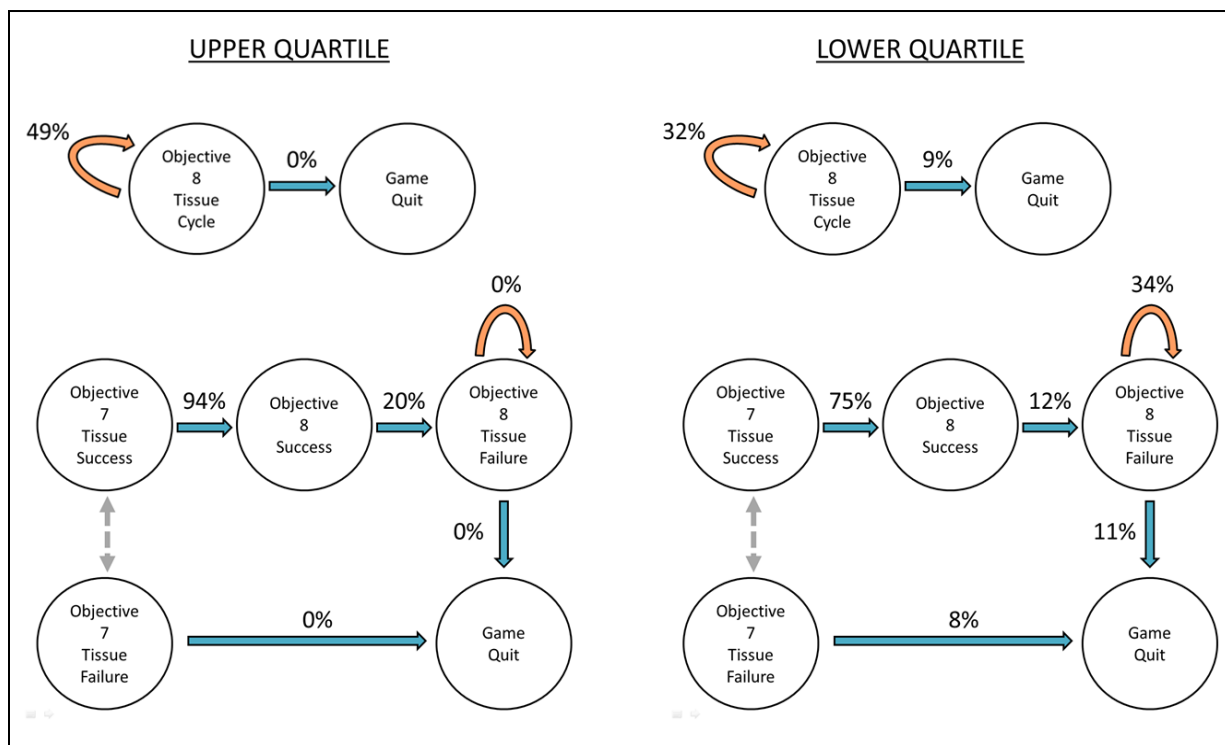


Figure 38. Late-game tissue progression, visualized via Markov for Trend II.

The last Markov diagram in Trend II, Figure 38 shows late-game tissue progression patterns (the top row is a simple, base-model version; the bottom diagram is from the more

detailed Markov). In demonstrating synthesis of cell and tissue skills, and advancing to new gameplay objectives, tissue mastery in Objectives 5-7 connects positively with learning gains. Upper quartile players who succeeded at Objective 5 cell cycles were also likely to move on to tissue success (31% UQ transition rate from cell to tissue success, contrasted with only 20% in the LQ). In turn, upper quartile players who did well with tissue cycles in Objective 5 tended to move smoothly to the next objective (a transition from Objective 5 tissue → Objective 6 success being 44% likely; versus only 24% in the lower quartile). Lower quartile players, on the other hand, had high quitting rates during Objective 5 tissue cycles: 31% of the lower quartile dropped out of the game during the Objective 5 tissue cycle, while only 18% of the UQ quit the game at this point (Figure 39). Similarly, dropout rates after Objective 7 tissue cycles were 8% more likely in the lower quartile group, with *zero* likelihood in the upper quartile. It follows that the upper quartile were more likely to go on to smashing success after their Objective 7 completion, with a whopping 94% chance of going on to an Objective 8 cell success immediately following a successful Objective 7 tissue cycle. (This contrasts with only 75% probability in the lower quartile group.)

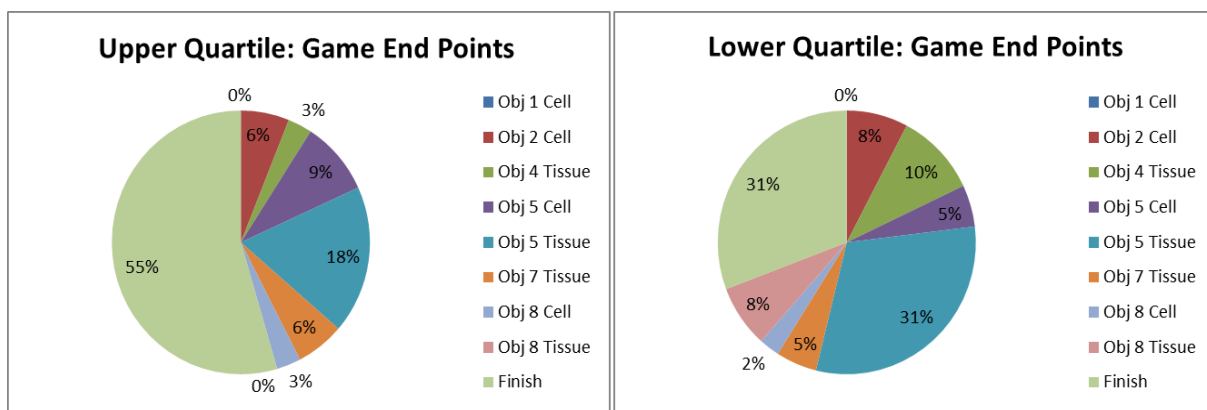


Figure 39. Charts of final play levels for each quartile of learners.

During the boss level, learner groups demonstrated clear differences in tissue cycle performance. Indeed, Markov data reveals that the higher learning group continued the trend of resilience, bouncing from success to failure and back, but not getting stuck in a consecutive tissue failure rut – nor quitting. Broadly, the upper quartile had higher repetition of Objective 8 tissue cycles (49% probability, versus 32% for the LQ); in this base Markov data, Objective 8 repetition included both consecutive successes and failures, not specifying performance but implying tenacity on the part of the upper learning group. The performance-detailed Markov provided deeper insight; in it, the probability of moving from success to tissue failure was higher in the upper quartile (transition probability of 20% versus 12% in the LQ), but unlike their counterparts, the upper quartile did not get stuck in this failure. The UQ had a *zero* percent probability of repeating tissue failure or quitting the game after failing in Objective 8, while the lower quartile were 34% and 11% likely get stuck in a tissue fail rut and quit (respectively). Chapter Four insights show that Objective 8 is an important level, since its completion is positively associated with learning gains (Table 6). In regards to tissue, however, the data also show that the tissue cycle in Objective 8 is a large quitting point (Figure 39), with zero “rage quitting” from Objective 8 tissue cycles happening in the upper quartile of learners. Thus, extended tissue failure in Objective 8 (also compounded with resultant cell failure) seems to be unfavorable for learning outcomes.

Discussion of Trend II Results

In tissue levels, a recurring pattern characteristic of the higher learning group was *not* aversion to failure, but increased *recovery* from failure. The upper quartile had plenty of instances of failure in tissue cycles, but from that failure point had notably higher transition rates to a consecutive tissue success, and from there tended to move forward in game progression –

especially in scaffold-and-fade tissue objectives and synthesis levels combining unguided cell and tissue cycles. Conversely, getting stuck in a tissue rut with repeated failure was characteristic of the lower learner group. In a related pattern, quitting the game immediately after a tissue failure occurred in the lower quartile throughout the last half of the game. The upper quartile, interestingly, had *zero* occurrence of this tissue-fail-to-quit behavior. This insight into the pivotal role of tissue performance on game completion and learning can help inform iterative design of *Progenitor*. Building in player-adaptive layers of instructional cues and help resources in critical scaffold-and-fade tissue levels, for example, can help optimize the game experience for both learning and play progression.

Results Trend III: Turning the Tables on Failure

Trend III is comprised specifically of results around far failure, which demonstrate fascinating changes in relationship to learning throughout the game. Similar to Chapter Four, far failure seems to evolve from having a negative to a positive relationship with learning over the course of play. Specifically, patterns of far failure emergent in Objective 5 and continuing to Objective 8 contrast sharply with the negative impact of early-game far failure in Trend I. Certain far failure patterns in these later game phases shows a positive connection with learning, implying its possibly deliberate use as part of engaged experimentation or strategic play. Just as in Trend I and II, a visualization and discussion of the relevant Markov findings will follow, in connection with corroborating statistical analyses.

Detailed Sequence of Trend III Results

Analysis of Objective 5, the first cell objective since tutorial levels 1 and 2, revealed patterns in upper quartile far failure that continued through the Objective 8 boss level. In this

context, three main action sequences characterized the greatest learning group: repeated far failure, multiple failures before a success, and an immediate success-success sequence progressing to the next game level.

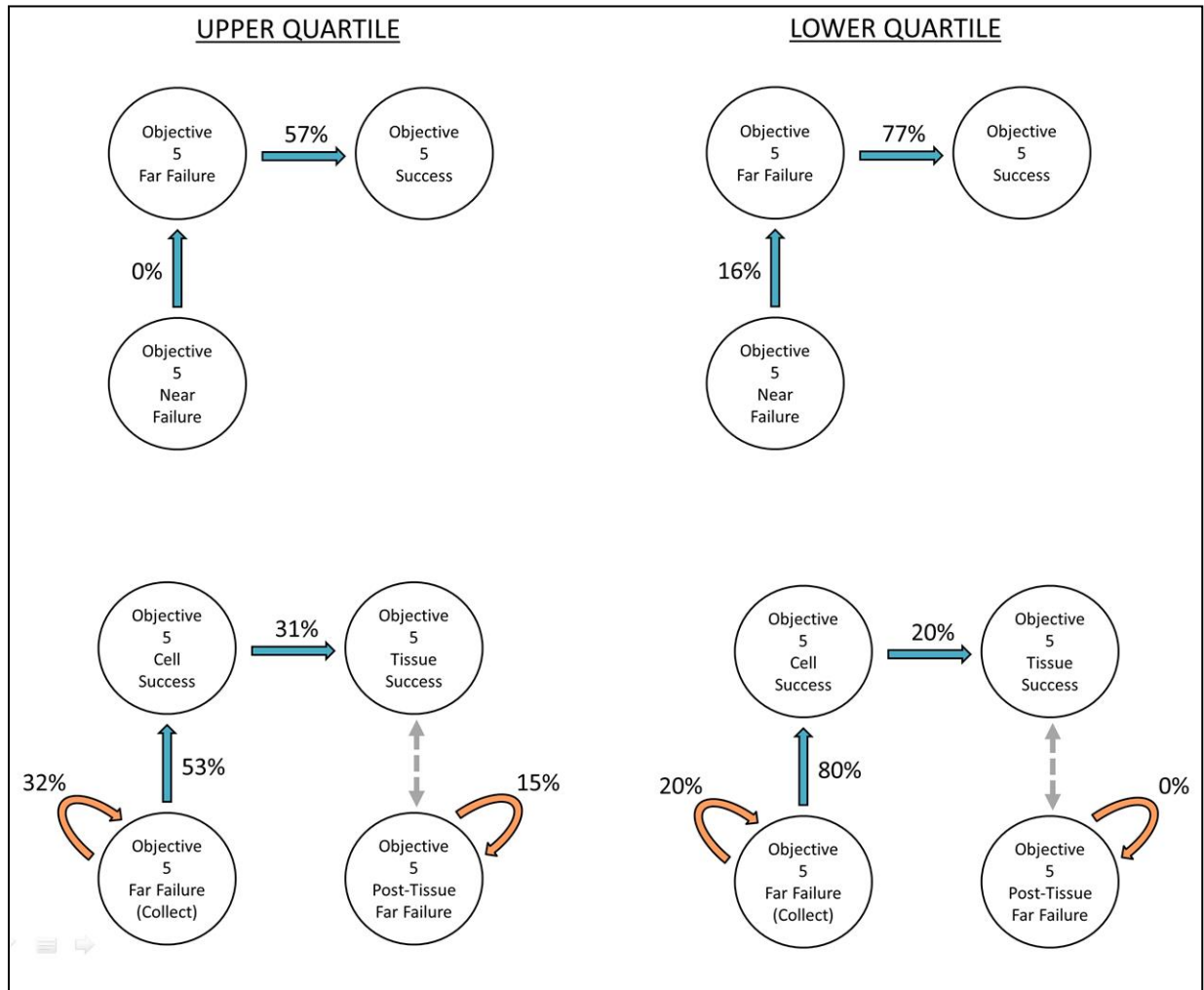


Figure 40. Objective 5 far failure patterns, visualized from Markov findings for Trend III.

This (far failure)-(far failure)-(success)-(advance) pattern recurred throughout Objective 5 and Objective 8 in the upper quartile, with cross-validating evidence – and exists in clear, stark contrast to the negative role far failure played for learning in Trend I. Figure 40 above shows basic findings relevant to this pattern in Objective 5, with a moderately-detailed Markov diagram

at the top, and the highest-res Markov diagram on the bottom. These results showed that the UQ was more likely to have consecutive far failures in both cell phases of Objective 5, with a 12% and 15% higher repetition rate (respectively). It follows that in repeating far failure, the LQ was less likely to go directly from a far failure to a success (57% in the UQ versus 77% in the LQ). However, once the higher gains group reached a success, they were more likely to have a consecutive success in the next game level (in transitioning from cell phase 1 to cell phase 2 in Objective 5, the UQ had a 31% probability to go directly from success to success, as opposed to only 20% in the LQ). Near failure did not appear to be a part of this progression, as the upper quartile had *zero* Objective 5 near failure to far failure transitions – while the lower quartile had a 16% probability of this movement. This implies that the recurrence of far failure for the higher learners may be deliberate, and not a result of haphazard oscillation between failure types in careless play. Overall, for the upper quartile, this sequence shows a series of repeated far failures, and then a success, which tended to lead directly to another success in the next game level.

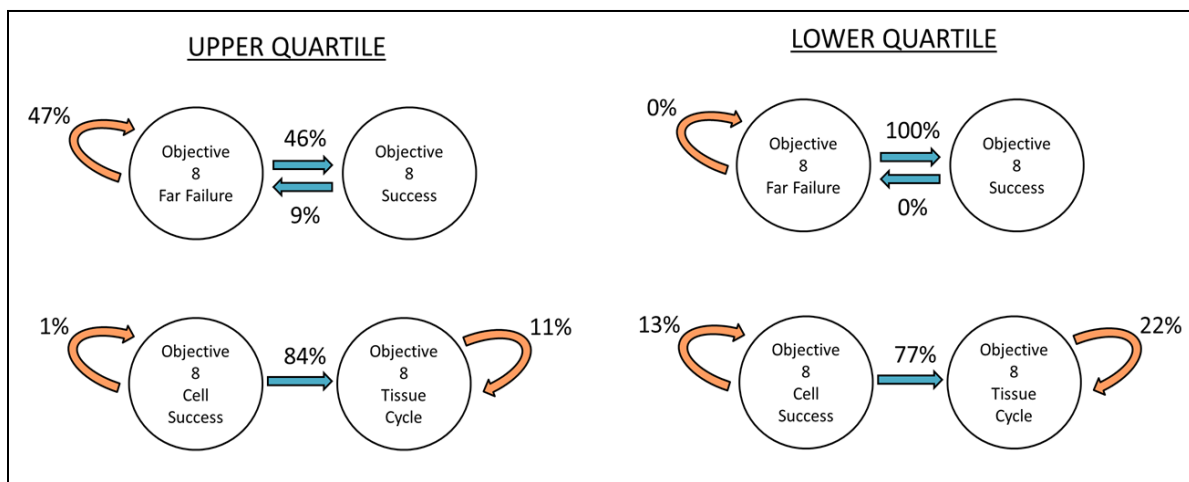


Figure 41. Markov findings of Objective 8 far failure, visualized for Trend III.

Figure 41 visualizes late-game evidence of this far failure-success-progression pattern, which even more strongly characterized learners in Objective 8. In this boss level, the lower quartile of learners had a *zero* percent likelihood of repeating cell far failure; the upper learning group, however, had a 47% probability of far failure repetition. Just as in Objective 5, it then follows that UQ far failure repetition was also more likely than an immediate transition to success (the upper quartile had a 54% smaller probability of a single far failure to success transition). Once a success was achieved, however, the upper learners tended to start the far failure cycle over, or progress onward to the next level. (UQ probability of an Objective 8 success to far failure was 9%, and a cell phase 1 success to cell phase 2 success was 84% – while the LQ had 0% and 77% chance, respectively.) Starting in a given cell phase, the upper quartile transitioned most frequently to far failure or progression in the next phase; thus, same-phase consecutive successes were less likely (12% lower in the UQ as compared with the LQ). As the chart of last objective played shows (Figure 39), a key characteristic of the upper quartile was transitioning to success and game completion after far failures throughout Objective 8. Cross-validating results from Chapter Four also include the positive relationship between these boss level success and learning gains, as well as the significant positive relationship between game progression and pre-post performance (Table 6). Another corroborating pattern from Chapter Four is the interesting increase in far failure of the upper quartile between Objectives 2 and 8, while the lower quartile had steadily decreasing far failure throughout (Figure 27). Objective 8 far failure was also shown to be positively correlated with learning gains (Table 9).

Discussion of Trend III Results

In the last two cell objectives of the game (5 & 8), far failure positively characterized learners in a recurring pattern of play: repeated far failure, multiple failures before a success, and

an immediate success-success sequence progressing to the next game level. This trend seemed to strengthen from Objectives 5 through 8, and is supported by cross-validating evidence from statistical and descriptive analyses. The clarity of the pattern, recurring throughout four cell phases within Objectives 5 & 8, suggests there may be something deliberate about this behavior specific to learners with the highest gains. This kind of creative exploration of failure, or deliberately “transgressive” game behavior, may very well be a part of engaged experimentation or strategic play.

In summary, these three trends show interesting patterns of failure in relationship to learning. Trend I reveals a negative relationship between Objective 1 far failure and learning gains, showing greater tutorial-level stagnation via far failure in the lower quartile. The upper quartile also experienced far failure, but tended to recover quickly from a given instance of failure and transitioned more frequently into subsequent success. Trend II centers on tissue levels, where a similar recurring pattern characteristic of the higher learning group was *not* aversion to failure, but increased *recovery* from failure. Trend III focuses on the last two cell objectives of the game (5 & 8), where far failure had a clear, marked shift in relationship to learning. In these cell mastery levels, it *positively* characterized learners in a recurring pattern of play: repeated far failure, multiple failures before a success, and an immediate success-success sequence progressing to the next game level. Cross-validating analyses corroborate these trends, which enrich our understanding of failure as a non-monolithic, sequence-sensitive contextual construct in play and learning.

Discussion and Conclusion

This analysis goes beyond binned, frequency-based counts of failure, instead studying high resolution, context-specific play sequence for new insights into the evolving relationship of

nuanced failure and learning. These findings reveal that it is not blanket existence of failure - rather, its context-specific relationship to actions before and after – that matters for learning. If Chapter Four’s analysis showed that failure cannot be monolithically defined in relationship to learning, this sister study shows that the contextual positioning of those failure types within a play sequence can unlock deeper patterns of learning.

In early objectives and tissue levels, the role of recovery from far failure and tissue failure was key for learning. In these specific levels, repeated failure showed negative correlation with learning and often resulted in game quitting. In later cell levels, the role of far failure evolved, as it became clearly connected to a series of consistent play actions characteristic of the upper learning group – an integral segment of a series of play actions positively correlated with learning and game completion. Thus, specific kinds of failure actually start to play a positive role in learning gains as play progresses. The increasingly positive impact of failure in more elaborate, successful play sequences could signal the evolution of reactive play (emphasizing recovery from failure as a learning characteristic) into more proactive strategic thinking (with a sense of mastery and agency, deliberately leveraging the game’s failure mechanisms for forward movement). As analysis I opened up inquiry into kinds of failure, this analysis reveals an evolving meaning of failure types in relationship to learning throughout gameplay, and leads naturally into the next chapter’s investigation of possible transgressive play patterns emergent in these findings.

Chapter Six:

Experimentation and Learning – Predictive Modeling with Detectors

Introduction and Framework

This analysis chapter builds on the investigation of performance and play trajectories with educational data mining of player exploration trends related to learning. The data features from previous analyses – specifically, in-game performance measures and base play progression – are leveraged in this study to make inferences about player experimentation in the gameworld of *Progenitor X*. The investigation of experimentation in play and learning represents a merging of all three lenses of game microworlds as designed experiences (Figure 1), melding player-specific goals (Lens C), game as educational content (Lens A) and the game as a play-driven medium (Lens B). This intersection is explored with the method of detector building – a data mining technique used to mine log data for indicators of behavior (e.g. Baker & De Carvalho, 2008; Cheng & Vassileva, 2006; San Pedro, Baker & Rodrigo, 2011). Selected features of event-stream gameplay are used as input variables in a predictive modeling of thoughtful exploration in the gamespace, and the exploration codes are then descriptively and statistically investigated in relationship to learning. The core research question is: What play data features characterize experimentation in *Progenitor X*, and how does this behavior connect with learning outcomes?

The study of play experimentation and learning in *Progenitor X* is based in past data mining work and educational games research. Previous data mining research has used detectors to categorize student behavior within digital learning spaces, such as learners “gaming the system” in cognitive tutors (e.g. Baker, Corbett & Koedinger, 2004) and measuring user “goal seriousness” (e.g. DiCerbo & Kidwai, 2013) in completing tasks. Along the same lines of

characterizing student objectives, this *Progenitor X* study seeks to mine patterns of user interaction that indicate experimentation. In games, mining this designed “system of interaction” is vital to understanding “player experience” (Salen & Zimmerman, 2004, p. 61; Schell, 2008), especially around experimental behavior. In contrast to task-driven tutoring systems and simulations, gameworlds are set up to provide roles and goals (Squire, 2011) in an narrative-based, endogenously motivating context (Costikyan, 2002). As such, they invite a kind of transgressive play (Salen & Zimmerman, 2004), in which players navigate the game in unanticipated ways, guided by their own goals and interest. Indeed, educational games are a complex medium which involves the intersection of at least three very different sets of goals – discussed in the unifying games as microworld lenses of Chapter One. Game genres generally invite exploration and testing of game constraints (one kind of goal), content designers often impose another (e.g. a goal of learning biology), and users come in with their own individual motivations and curiosities for play (suggesting a whole range of player-specific goals). This makes more traditional binary constructs like “on task” or “off task” very one-dimensional for game study (especially when used relative to a single assumed goal). In interpreting player behavior in gameworlds, the construct of experimentation may better represent the intersection of explorable worlds, academic content models, and interest-driven player paths. In interpreting user action data through the lens of experimentation, the goal is to better understand the complex, evolving roles of interaction data like far failure in exploratory play experience. For example, is far failure a sole characteristic of blind-clicking, or does it also occur with transgressive play? If so, how does it evolve in its relationship to experimentation and learning from objective to objective? This analysis explores these kinds of inquiries, first building a

detector for thoughtful exploration in play, then connecting specific kinds of strategic exploration with learning outcomes.

The following chapter will review methods for detector building, then explain the thoughtful exploration construct as well as a detailed coding schema and process. It will then discuss three main analyses arising from the fully coded exploration data: 1) a predictive model of thoughtful exploration during *Progenitor X* play; 2) descriptive analyses of the relationship between exploration and learning outcomes; and 3) a detector of learning-supportive strategic failure in *Progenitor X*.

Methods: Building a Detector for Experimentation in *Progenitor X*

Frequently used in educational data mining, a detector is an automated model that can infer from log files whether a student is behaving in a certain way. To create, or train, that automated model, it relies on something computers do not have: human judgment. Several steps are summarized below:

- **Decide on a behavior construct.**

Researchers building a detector need to first decide what kind of student behavior they're looking for – for example, “gaming the system” (e.g. Baker et al., 2004). This behavior construct is deliberately general, and qualitative in nature; it does not *require* a hypothesis about specific data features which will predict it, since its purpose is organically mine data patterns connected with the interpreted behavior.

- **Aggregate student interaction data.**

The researchers then gather student interaction data which they believe will give insight into the behavior construct. These data must be synchronous to log-file activity of

interest. Student interaction data varies based on the study, and can include observational data (e.g. or text replays (Baker & de Carvalho, 2008)

- **Code for behavior construct.**

The coding for behavior happens next. For each “stanza” of data, a researcher will qualitatively code it with the behavior construct (for example, “gaming” or “not gaming” the system.)

- **Predictive modeling with coding and log file data.**

Once all the coding is finished, the coded data is synchronized with the log files so that behavior can be connected with click-stream action. Then, this synchronized data is put into a predictive model. The behavior (e.g. “gaming” or “not gaming”) becomes the outcome variable, and the data features connected with the behavior are the predictors. Data features most common to the behavior can be clearly identified. Thus, a model is created that can automatically predict the behavior if given future log-file data. It can then be used to drive interventions or in discovery with models analyses.

Each step for this study (as generally outlined above) will be described in the following section.

Progenitor Behavior Construct: Experimentation

This detector aims to better understand player experience in the multifaceted realm of learning gameworlds through the behavior construct of experimentation. Defined here as “thoughtful exploration,” it centers on attentive exploration of the gamespace, deliberately testing constraints and consequences of in-game action. (Blind clicking in impatience, without

curiosity or attentiveness, would not constitute experimentation in this definition.) The general schema for experimentation can be seen in matrix form (Table 13). Conceptually, the main categories were “thoughtful exploration” or “not thoughtful exploration”. Any behavior which went outside the bounds of strict game instructions, exploring the game UI or mechanics boundaries in a seemingly thoughtful or systematic way, was considered thoughtful exploration. Player actions which seemed to characterized blind, hasty, or redundant clicking around the space were not considered thoughtful exploration. Also in the “not thoughtful exploration” category was behavior which stuck narrowly and strictly to game cues, never deviating from prompted actions.

Table 13

Basic Exploration Code Categories

<i>Core Category</i>	Thoughtful Exploration		Not Thoughtful Exploration	
<i>Description</i>	Basic (feature exploring)	Strategic use of explored mechanics	Careless clicking	Straight and Narrow (no exploration)
<i>Abbreviation</i>	“TE”	“TES”	“C”	“S&N”

As Table 13 describes, thoughtful exploration can be seen as a basic exploring of UI features or game mechanics, which is code “TE” above. Another related variant is strategic use of explored mechanics, which involves using knowledge of game boundaries for clear play strategy. This is abbreviated “TES” for “thoughtful exploration – strategic”. The two basic categories not considered thoughtful exploration are careless clicking (“C”), as well as the behavior of no deviation at all from prompted game instruction, called “S&N” for straight and

narrow. These four basic categories were based on the original thoughtful exploration construct, and then refined based on emergent trends in the data. More detailed explanation and examples of each follow in the “coding” section.

Aggregating Data for Evaluation

The data was presented for the coding of experimentation using a kind of text replay, a data-mining form of “distillation of data for human judgment” (Baker & Yacef, 2009). Text replays are a visual summary of student interaction data, grouped together as a series of actions for evaluation by the researcher. For example, Figure 42 shows a text replay from a series of five problems in a cognitive tutor data set (Baker & de Carvalho, 2008). For each problem, the researcher can see the corresponding tutoring unit, the student’s answer to the problem, and the time it took to answer. For this “clip” of five problems, the researcher would overview the data and then decide if the evidence pointed to “gaming” or “not gaming” the system.

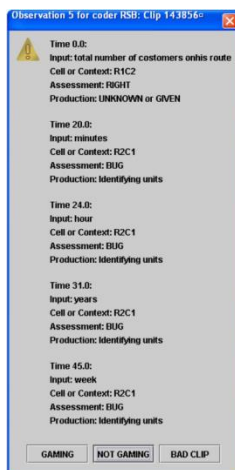


Figure 42. Example of a Text Replay (Baker & de Carvalho, 2008)

Similarly, snapshots of log-file play data were used for the *Progenitor* coding of experimentation. Existing text replay software (e.g. EDM Workbench – Rodrigo et al., 2012)

could handle display of a fixed number of cycles, but in *Progenitor* this did not correspond with the subobjective context (since cycle number varies wildly from objective to objective). This, *Progenitor*-specific data snapshots were created in excel to fit the needs of the experimental design. Each clip was created to display features deemed relevant to the construct of thoughtful exploration (Figure 43). These included metadata about student restarts and game completion, as well as cycle specific data detailing cycle start and end, cell types, cycle duration, outcome, failure/success evaluation, UI buttons used, and tool use inside the grid. Each of these features were chosen because they were considered good indicators of UI affordances and game action parameters (relevant to the construct of exploration). This concise layout worked well, being a pleasantly sparse but efficient way to convey a dense amount of player information.

METADATA					CYCLE-SPECIFIC DATA									
MY CODE	Student ID	Current Obj	Final Obj	# of Restarts	Timestamp	Action Type	Starting cell	Cycle duration	Outcome	Collected Cell	EVALUATION % of turns	Buttons us	# of FALSE	Tool use
student_666	0	5C	0		0	HEADING								
					3307	CYCLE START	Fibro					--> next	5	ShockFALS
					3338	*						<<- back		
					3419	*						*iPS Cells*		
					3462	*						--> next		
					3762	CYCLE END		40	Collect	IPS1	Success	100%		

↑ CODE INPUT

↑ CYCLE START AND END EVENTS

↑ SUCCESS OR FAILURE OF CYCLE

↑ UI BUTTONS USED during cycle (outside of in-grid tool use)

Figure 43. Adapted text replay clip for *Progenitor* TE coding.

Coding the Data

Initially, the base coding schema was binary (Table 13) – coded for Thoughtful Exploration (TE) or Not Thoughtful Exploration (Non-TE). It was soon evident, however, that more subtle behaviors were occurring in the gamespace, and soon emergent subcodes evolved from categorizing more nuanced player action. The final coding scheme (Table 14) shows the main categories of subcodes that were created both for TE and Non-TE base behavior. The four

subcodes of “TE” (base thoughtful exploration), “TES” (strategic use of explored mechanics), “C” (careless clicking – not TE), and “S&N” (straight and narrow – no exploration) remain from Table 13. During coding, more specificity was added for variations in these codes, particularly based on the outcome of the level (Table 14). The syntax for codes came to be made up of two parts: 1) the base exploration code and 2) outcome of the objective.

Take the example of a text replay clip that had a basic exploration code of TE. If the objective was completed with instant success (and no failure), the code would have a suffix of “success” (abbreviated to “Succ”). Thus, the final syntax would be TE-Succ. If the objective eventually ended in success, but only after much failure, its code suffix would be “tenacious” (“Ten” for short, with a final code of “TE-Ten”). If the objective’s clip ended in a quit fail, the code would say “TE-Quit”. This pattern continues across all four TE/Non-TE code types in the table.

Table 14

Detailed Exploration Code Categories

	Thoughtful Exploration		Not Thoughtful Exploration	
	<u>“TE”</u>	<u>“TES”</u>	<u>“C”</u>	<u>“S&N”</u>
Instant success	TE-Succ	TES-Succ	C-Succ	S&N-Succ
Failure, then success (tenacious)	TE-Ten	TES-Ten	C-Ten	S&N-Ten
Quitting the game	TE-Quit	TES-Quit	C-Quit	S&N-Quit

Data was coded across all objectives, which were broken into subobjectives of parallel size. For example, Objective 0 is a training objective with only one cell cycle as a goal, so this was kept as Objective 0. Objective 5, on the other hand, is a compound objective which contains

three separate goals: cell phase 1, cell phase 2, and a tissue phase. Thus, Objective 5 was broken up into 5A, 5B, and 5C. The same was done for Objective 8 (see Table 15). Twelve final subobjectives per player were coded for thoughtful exploration, visualized simply in Figure 44, and broken down into more detail in Table 15. One snapshot per subobjective for each individual player (like Figure 43) was coded at a time. This meant that a player who finished the game would have 12 discrete exploration codes total. The clip level of subobjectives was chosen because it framed player action in a very clear, consistent, and specific context, thus making judgment of exploration behavior more likely to be accurate.

Table 15

Subobjective Labels for Progenitor

Objective	Cycle Type	Subobjective [‡]
0	Cell (type 1)	0
1	Cell (type 2)	1
2	Cell (type 1)	2A
	Cell (type 2)	2B
3	training level – min. player action	--
4	Tissue	4
5	Cell (type 1)	5A
	Cell (type 2)	5B
	Tissue	5C
6	organ level – min. player action	--
7	Tissue	7
8	organ level – min. player action	--
	Cell (type 1)	8B
	Cell (type 2)	8C
	Tissue	8D

[‡] Objectives 3, 6, and 8A had very little player action involved and thus were not coded for thoughtful exploration.



Figure 44. Color-coded visualization of the 12 Progenitor subobjectives.

Total subobjectives coded across all 110 players numbered 1,084 (since not all players finished the game). For construct consistency, multiple coders were used and were measured for interrater reliability using Cohen's Kappa, for a final value of $K = .908$.


Examples of Coding Schema

In order to illustrate the meaning of thoughtful exploration in *Progenitor* play, and thus deepen understanding of analysis findings, this section will give examples of codes most commonly used during the evaluation process. It also serves as a more ethnographic set of example findings – individual manifestations of game-wide exploration patterns.

Exploration Without Failure (“TE-Succ”)

Figure 45 is a coding snapshot of a player's game interaction during Objective 0. This example is given first because it is one of the simplest kinds of codes. The player has experienced no failure, yet has taken the time to explore the UI with almanac vocabulary entries and the back button to review instructions (neither of which are prompted or required interactions). This entry was thus scored “TE-Succ”, because it had no failure (only success), but had elements of thoughtful exploration. For most follow-up analyses, this was simplified to TE”.

Time	Action	Start	Duration	Outcome	Cell	EVALUATION	% of turns	Buttons used
0	HEADING							
199	CYCLE STA	Fibro						
215	*							*Electroporate*
216	*							--> next
243	*							--> next
257	*							<-- back
263	*							--> next
288	*							--> next
310	CYCLE END		111	Collect	IPS1	Success	0.866667	



Objective success
No failure

almanac or "back" use


Figure 45. An example clip of exploration without failure (code: “TE-Succ”).

Exploration With Failure (“TE-Ten”)

Figure 46 shows an example of thoughtful exploration in the gamespace (in early levels) with failure. This instance would have been coded “TE-Ten”, meaning thoughtful exploration with tenacity (several failures before success). This student shows methodical exploration of the UI, going from one almanac word, to finding a different one, to then discovering the instruction perusal button (the “back” button, which reviews the last instruction given). None of these are prompted or required buttons of interaction. The pace of finding these UI elements is unrushed. In addition, the player has two different kinds of failure without seeming to get stuck on either one, or repeating mistakes. Because it is still very early in the game (Objective 0 tutorial), the player seems to be discovering different ways to fail and learning from each one. The third cycle is a success. (Interestingly, this same player went on to earn a “TES” code – or strategic use of

explored parameters – later in the game.) This was simplified to “TE” for most subsequent analyses.

Time	Action	Start	Duration	Outcome	Cell	Eval	% of turns	Buttons used
0	HEADING							
3307	CYCLE STA	Fibro						
3338	*							--> next
3404	*							<-- back
3419	*							*iPS Cells*
3457	*							*Fibroblasts*
3462	*							--> next
3591	CYCLE END		69	Collect	Fibro	FF - collect	0.866667	
3599	CYCLE STA	Fibro						
3611	*							<-- back
3612	*							--> next
3704	CYCLE END		37	Grid Destroy		NF		
3722	CYCLE STA	Fibro						
3762	CYCLE END		40	Collect	IPS1	Success	1	



Eventual success
More than 1 failure

almanac or "back" use

Figure 46. An example clip of exploration with failure (code: “TE-Ten”).

Strategic Use of Thoughtfully Explored Mechanics (“TES”)

Next is an example of the TES code – strategic use of thoughtfully explored mechanics – in which players combine mastery of the core skills with strategic failure to improve their efficiency. Thus, TES is often referred to as the strategic failure code. This is the last example of an exploration code in this set – and arguably the most complex and interesting.

Strategic failure and the TES code often was characterized by a behavior now dubbed “harvesting”. This phenomena was unknown to the researchers before the coding started, and

emerged as a clear recurring behavior during certain cell cycles of the game. To understand harvesting, we must delve a bit deeper into expected game cycle behavior.

Normally, *Progenitor* expects a player to start with one kind of cell, treat it so that it transforms, and then collect the new kind of cell (for a successful cycle). However, “harvesting” is a way around this. Instead of having to perform this start(old)-treat-collect(new) cycle, some players figured out the game would allow them to start with the cell they needed to collect, move it around on the grid so that it replicated, and then simply collect the expanded number of original cells. This avoids the “treat” phase all together and helps keep health up longer, thus looking like a start-move-collect sequence (all using only one kind of cell).

For example, a common “legit” cell cycle starts with fibroblasts (pink cells), then directs the player to treat them with electricity, thus making stem cells (purple cells) for successful collection (Figure 47). By contrast, a harvesting cycle would populate the grid with purple stem cells, move them around a bit, and then collect the expanded batch of purple cells to reach target numbers (Figure 48). Thus, the player had all the purple iPS cells they would need, all without going through the intended shocking-pink-cell process.



Figure 47. “Legitimate” cell cycle of start-treat-collect, which ends with a new kind of cell.




Figure 48. A “harvesting” cell cycle in strategic failure, which starts and ends with the same cell (in a start-move-collect sequence, skipping the treat stage).

A fascinating harvesting fact, however, is that a player needs at least one legitimate cycle *before* they start harvesting, so that they can have enough of the unique new cells to start a harvest cycle. This means that harvesting requires mastery of the base mechanics to use successfully. Hence, this is *not* like “gaming the system” (which by definition avoids the intended skill acquisition); instead, this strategic failure requires mastery of game mechanics, thorough and thoughtful understanding of the game’s boundaries, and the metacognition to put them together in a hybrid strategy to maximize health and cycle efficiency. As Salen and Zimmerman (2004) say in *Rules of Play*: “To skillfully break rules requires an intimate knowledge of the rules themselves” (p. 282).

Time	Action	Start	Duration	Outcome	Cell	EVALUATION	% of turns	Buttons used
0	HEADING							
298	CYCLE START	Fibro						
322	CYCLE END		24	Collect	IPS1	Success	20%	
327	CYCLE START	IPS2						
341	CYCLE END		14	Collect	IPS2	FF - collect	67%	
345	CYCLE START	IPS2						
365	CYCLE END		20	Collect	Ecto	Success	93%	
413	CYCLE START	Tissue1						
428	CYCLE END		15	Collect		Success	30%	

Clean execution - no static with frenetic clicking behaviors or haphazard cycle failures.
Just the clear legit cycle + harvest cycle pattern.



 "Legit" cycle 1
 Harvesting cycle 1
 "Legit" cycle 2 - shows mastery of skill

Figure 49. An example clip of strategic failure (code: “TES”).


One example of this strategic failure (or TES) is in Figure 49. Here we see a clip from Objective 5, in which the player begins with a “legitimate” cycle, completed efficiently and successfully. Next, he/she engages in what we’ve called a “harvesting” behavior. This is the strategic failure element. In the second cycle, the player populates the grid with stem cells (called “iPS” cells in the coding snapshot), moves them around on the grid to replicate them, then collects the expanded batch of stem cells a few seconds later. The game registers this as *far failure*, because it flags the stem cells picked up as not correct (recognizing that they have *not* been treated with electricity as intended). Still, for better or for worse, the game allows these to be used for subsequent cycles. And so the show goes on, with the third cycle in our example being a perfectly executed legitimate cycle, beginning with the harvested iPS cells, treated with a growth factor, and then collected as newly minted Ectoderm cells (“Ecto” for short in the clip). Figure 49 highlights this with arrows on the right in the diagram, showing a top “legit” cycle, a middle “harvesting” cycle, and on the bottom another “legit” cycle. There are no messy, seemingly unnecessary failures, no apparently hasty clicking or ruts, no building towards a cycle

quit – just a clean sequence of Objective 5 efficiency. This is the “TES” code: far failure in harvesting as part of a layered strategy. (The designation of “Ten” or “Succ” did not often apply to this code, as both success and failure were an innate part of the strategic failure practice.)


Not Thoughtful Exploration: Careless Clicking (“C”)

The snapshot captured in Figure 50 is an example of an Objective 5 clip coded as “C” (standing for Careless, not thoughtful). A red flag here is immediately the repeated far failure of fibroblast collection, which is a behavior not associated with harvesting because there simply is no strategic advantage to doing it. It is literally impossible to run out of fibroblast cells this early in the game, and repeated collection of these useless cells in far failure was a behavior often exhibited by players during periods of frustration (as observed in numerous playsquads by the author), and a harbinger of mid-game “rage quits” and Objective 1 failure ruts (as can be seen in the log files). Another sign of haphazard clicking is the bouncing around between views (multiple times, beyond that of initial discovery) without cycle completion (c.f. Wixon et al., 2012). This view switching can be seen with the “TISSUE” “CELL” “TISSUE” “CELL” record (which signals moving to a different laboratory template, of either the tissue or the cell) in the buttons used column. These view changes, along with the unproductive failure of the cycles, was occurring somewhat quickly (e.g. 23, 13, and 10 seconds for a cell cycle was comparatively quick – the game average was 35.3 seconds per cycle.) All this lead to an impression of hastiness and unproductive, redundant failure – hence the “C” rating.


Time	Action	Start	Duration	Outcome	Cell	EVALUATION	% of turns	Buttons used
	HEADING							
	CYCLE* START	Fibro						
	CYCLE* PAUSE							TISSUE
1070	CYCLE START	Tissue3						
	CYCLE END	Tissue3	32	Grid Destroy		TF		
1099	CYCLE START	Tissue3						
	CYCLE END	Tissue3	23	Grid Destroy		TF		
1157	CYCLE START	Tissue3						
	CYCLE END	Tissue3	49	Grid Destroy		TF		
1204	CYCLE* RESUME							CELL
	CYCLE* END		66	Collect	Fibro	FF - collect	0%	
1247	CYCLE START	Fibro						
	CYCLE END		37	Collect	IPS1	Success	87%	
1257	CYCLE START	IPS2						
	CYCLE END		10	Collect	Fibro	FF - collect	0%	
1270	*							TISSUE
1295	CYCLE START	Tissue3						
	CYCLE END	Tissue3	24	Grid Destroy		TF		
1299	*							CELL
1315	CYCLE START	IPS2						
	CYCLE ABANDONED		13	Cycle Quit		QF		Obj: View
	*							



Increasingly rapid, 1-click cycles



Repeated failure with no particular pattern or improvement
Finally culminating in a quit fail




Bouncing around quickly between views without cycle completion

Figure 50. An example clip of careless clicking (code: “C”).

No Exploration and No Failure: Straight and Narrow (“S&N-Succ”)

Figure 51 is an example of another Objective 0 clip, during which the player has ONLY success, and only performs actions directly prompted by the system. Since there seems to be no exploration, we called this behavior straight and narrow (or “S&N”) as a code. The player has also only experienced success, so we also attach a “Succ” to it. The final code was “S&N-Succ”. Like most of the coding data, this was simplified to the base exploration code (just “S&N”) for analysis purposes.

Time	Action	Start	Duration	Outcome	Cell	EVALUATION	% of turns	Buttons used
0	HEADING							
204	CYCLE STAF	Fibro						
218	*							--> next
253	*							--> next
270	CYCLE END		66	Collect	IPS1	Success	0.866667	



one success only


"next" use only
(no *almanac* or "back")

Figure 51. An example clip of no exploration, without failure (code: “S&N-Succ”).

No Exploration, This Time *With* Failure (“S&N-Ten”)

The next illustration (Figure 52) shows an Objective 0 clip of no exploration, with failure. The person does not seem to be exploring the interface, or systematically investigating different kinds of failure, so the activity is labeled with no exploration, which in our scheme was called “S&N” (straight and narrow). Because there is repeated failure with eventual success, the clip is also labeled tenacious, or “Ten” for short. Thus, our code looked like “S&N-Ten”. For most analyses, this was simplified to just “S&N”.

Time	Action	Start	Duration	Outcome	Cell	EVALUATION	% of turns	Buttons used
0	HEADING							
216	CYCLE STA	Fibro						
233	*							--> next
239	*							--> next
242	*							--> next
272	CYCLE ENI	Fibro	56	Grid Destroy		NF		
284	CYCLE STA	Fibro						
317	CYCLE ENI	Fibro	33	Grid Destroy		NF		
322	CYCLE STA	Fibro						
324	*							--> next
357	CYCLE ENI	Fibro	35	Grid Destroy		NF		
363	CYCLE STA	Fibro						
404	CYCLE ENI	Fibro	41	Grid Destroy		NF		
410	CYCLE STA	Fibro						
455	CYCLE END		45	collect	IPS1	Success	1	



SAME failure
(no *almanac* or "back")

Figure 52. An example clip of no exploration, with failure (code: “S&N-Ten”).

Overall, these examples represent the most frequent codes in the gamespace across all 1,084 coded clips. The outcome designations (-Succ, -Ten, and -Quit) were useful for fine-grained coding purposes, and will likely be very informative for future study and modeling of the data. The analyses to follow, however, focus mainly on the construct of thoughtful exploration and its four designations (“TE”, “TES”, “C”, and “S&N”) in the coding scheme.

Results and Findings I: Building a Detector of Thoughtful Experimentation

The following sections of findings detail the results which emerged from the coding of the log file data. First reviewed will be an M5’ predictive model of thoughtful exploration (TE), then an exploration of TE code relationships to learning with descriptive analytics and nonparametric statistics, and lastly a J48 detector of the learning-salient TES code (thoughtful exploration with a strategic angle).

The first behavior that was modeled across gameplay was thoughtful exploration (TE), considered broadly as an aggregate number of occurrences across all 12 coded subobjectives. The goal of this analysis is to predict, based on gameplay activity, whether or not a student is engaged in thoughtful exploration. Consistent with the broad data collection methods detailed in Chapter Three, the total N for this analysis was 110 middle school students, the game's target audience. For this analysis, the TES code was considered a part of the TE umbrella, and the others were not (see Table 13). As this outcome variable was numerical, a regression tree model was chosen to build the predictive model of experimentation. Linear regression was performed in WEKA (Hall et al., 2009) using the M5' variable selection procedure (Y.-C. Wang & Witten, 1997). Linear regression was chosen as a relatively conservative algorithm, with a relatively low probability of over-fitting. Independent variables were not unitized, thus emphasizing practical significance in the model, and the regression output was cross-validated using Leave One Out Cross Validation (LOOCV) at the student level (the overall level of analysis). The final goodness metric was the post-validation coefficient of correlation.

Ultimately, the final M5' model achieved a cross-validation correlation of .627 to the behavior of thoughtful exploration, comparable to levels in similar game-based learning detector models (e.g. Baker & Clarke-Midura, 2013). Predictors included several of the fundamental features created in Chapter One, including cycle starts, time elapsed, and number of cells collected. One of the recurring metrics below is "Ph level", which represents the number of turns used before a cell collect. A higher Ph percent used indicates that more turns were used up before the cycle ended; a low Ph level used indicates that very few turns were taken before the cycle was ended. The model also featured several aspects of UI button use not core to cycle operations, such as review of instructional text with "forward" and "back" buttons, use of in-game almanac

links, and switching between cell and tissue lab screens. The model was split into three trees, divided at the top level by the criteria of number of total collects. Interestingly, this split falls roughly along the number of collects it takes to complete the game (implying possible insight into finished/nonfinished playstyle differences). If players had less than 14.5 collects (14 were required to complete the game), then linear model 1 applies. If players had more than 14.5 collects, they could fall into one of two groups based on the collection of stem cells (iPS cells) in the game – also interesting, because the collection of this kind of cell is an indication of the “strategic failure” behavior. With zero iPS cell collects in Objective 5, behavior falls along linear model 2; with 1 or more iPS cell collects in Objective 5, player group with linear model 3. Each linear model, with detailed features, is as follows:

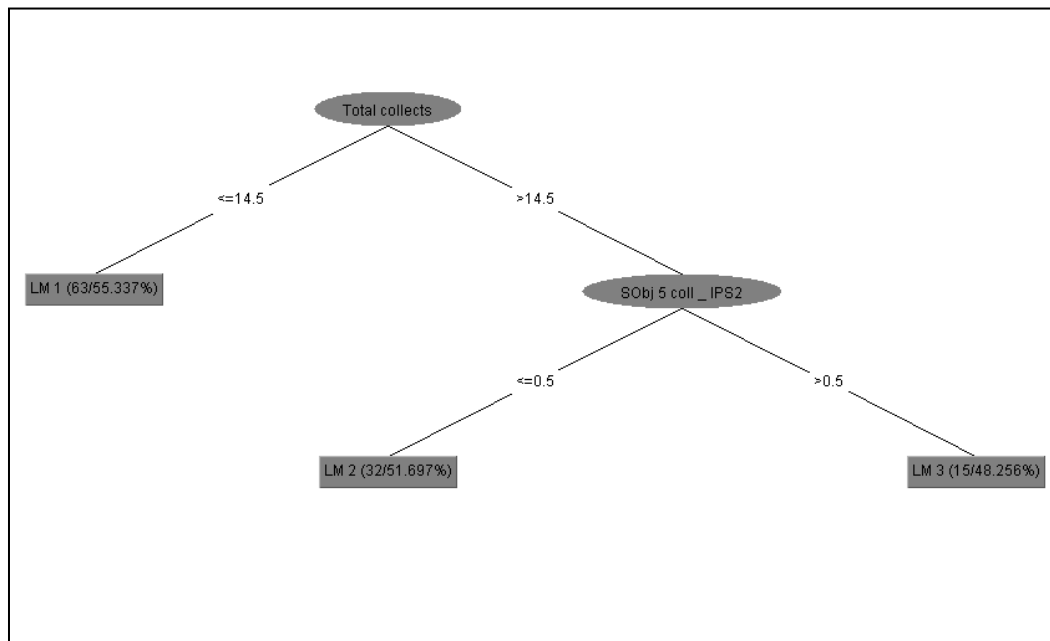


Figure 53. M5' regression tree WEKA output – predictive model of Thoughtful Exploration

Total collects <= 14.5 : Linear Model 1 (63/55.337%)

Total collects > 14.5 :

| Number of times iPS cells collected in Objective 5 <= 0.5 : LM2 (32/51.697%)

| Number of times iPS cells collected in Objective 5 > 0.5 : LM3 (15/48.256%)

Linear Model 1:

number of total Thoughtful Exploration instances =
 0.0011 * number of seconds spent on Objective 0 (training)
 + 0.0282 * total number of cells collected during Objective 0
 + 0.7095 * average Ph (%) used during Objective 0
 + 0.0256 * number of times elective UI buttons were used during Objective 0
 + 0.0026 * number of seconds spent on Objective 1
 - 0.0178 * number of fibroblast cell cycle starts during Objective 1
 - 0.2307 * average Ph (%) used during Objective 2-A
 + 0.0007 * number of seconds spent on Objective 2-B
 - 0.196 * average Ph (%) used during Objective 2-B
 + 0.104 * number of times iPS cells collected in Objective 5
 + 0.0022 * total number of cells collected during Objective 5
 + 0.0957 * number of iPS cell cycle starts during Objective 8
 - 0.0033 * number of seconds spent on Objective 8C
 + 0.1094 * number of successful cycles in Objective 8-C
 - 0.0076 * total number of times a cell or tissue collection was performed
 - 0.1585

Linear Model 2:

number of total Thoughtful Exploration instances =
 0.8926 * average Ph (%) used during Objective 0
 + 0.0322 * number of times elective UI buttons were used during Objective 0
 + 0.0008 * number of seconds spent on Objective 1
 - 0.1119 * number of fibroblast cell cycle starts during Objective 1
 + 0.0033 * number of seconds spent on Objective 2-B
 - 0.2903 * average Ph (%) used during Objective 2-A
 + 0.0009 * number of seconds spent on Objective 2-B
 - 1.4842 * average Ph (%) used during Objective 2-B
 + 0.321 * number of times an iPS cell collection was performed in Objective 5
 - 0.0143 * number of times a cell collection was performed in Objective 5
 + 0.0027 * total number of cells collected during Objective 5
 + 0.2328 * number of iPS cell cycle starts during Objective 8
 - 0.0041 * number of seconds spent on Objective 8-C
 + 0.1377 * number of successful cycles in Objective 8-C
 - 0.0095 * total number of times a cell or tissue collection was performed
 + 2.3512

Linear Model 3:

number of total Thoughtful Exploration instances =
 -0.196 * number of fibroblast cell starts during Objective 1
 + 0.8926 * average Ph (%) used during Objective 0
 + 0.0322 * number of times elective UI buttons were used during Objective 0
 + 0.0008 * number of seconds spent on Objective 1
 - 0.0687 * number of fibroblast cell cycle starts during Objective 1
 - 0.0547 * number of iPS cell cycle starts during Objective 2
 - 0.2903 * average Ph (%) used during Objective 2-A
 + 0.0009 * number of seconds spent on Objective 2-B
 - 0.7426 * average Ph (%) used during Objective 2-B
 + 0.4288 * number of times an iPS cell collection was performed in Objective 5
 - 0.0224 * number of times a cell collection was performed in Objective 5
 + 0.0027 * total number of cells collected during Objective 5
 + 0.5849 * number of iPS cell cycle starts during Objective 8
 - 0.0041 * number of seconds spent on Objective 8-C
 + 0.1377 * number of successful cycles in Objective 8-C
 - 0.0095 * total number of times a cell or tissue collection was performed
 + 2.7171

Independent variables were based on the features embedded in the text replay clips, chosen carefully as potential indicators of student exploration (see coding section for more detail). Overall, this model shows fascinating splits along the number of collects required for game completion – implying a potential finished/nonfinished play grouping – and, in the second tier, along iPS “harvesting” cycles strongly connected with strategic failure (see J48 model below). Not only does this regression tree solidify a consistent construct of thoughtful exploration in the *Progenitor* gamespace, it reveals the role of nuanced in-cycle efficiency (in Ph levels used and number of cells collected), and reinforces themes of previous chapters (e.g. negative early game repetition, and boss level success). The behavior patterns evident here, predictive of experimentation in the gamespace, also open the next section’s inquiry into the relationship between exploration and learning.

Results and Findings II: Exploration Codes and Relationship to Learning

Next, to explore the relationship between thoughtful exploration – now solidly modeled as a construct in the game – and learning outcomes in *Progenitor X*, two perspectives were taken. The first looked at aggregate code totals by full-game span. Descriptive heat mapping and base correlation were used during this first pass. Investigation quickly revealed need for greater resolution, however, and the codes were then examined in greater detail. The second perspective, then, examined the codes as sequential strings of behavior (not unlike like DNA strands) specific to objective context. For this second investigation, feature engineering, sequential probability modeling, and statistical comparison methods were employed.

For methods that involve contrasts between learning groups, this was based on the data from two sets of students: an upper quartile and a lower quartile of learners. This designation is

based on a pre-post assessment on regenerative biology (developed with content experts, and described in greater detail in the *ADAGE/Progenitor* methods Chapter Three). Relative to this performance, the quartiles are made up of two groups: *Progenitor* players with the greatest positive change in score, and players with the lowest change in score. The upper quartile consists of 33 players, and the lower quartile consists of 41 players. “UQ” is an abbreviation used throughout the dissertation for the upper quartile of learners, and “LQ” stands for lower quartile of learners. These only refer to learner groups (as determined by pre-post gains) – no other kinds of quartile groups are discussed in this dissertation. For all correlation and non-quartile analyses in this chapter, N=110. In all applicable analyses, p-values have been evaluated for significance with the R Studio QVALUE package (Dabney & Storey, 2004), controlling for multiple comparison based on False Discovery Rates (Benjamini & Hochberg, 1995; Storey, 2002). All adjusted p-values are thus called q-values, or “*q*”, in the results below.

Overview: TES – A Construct Significant to Learning

To get a visual mapping of code frequency across all objectives, a heat map was constructed (green is most frequent, red is least). The four codes put into the map (Figure 54) were no experimentation (abbreviated as “S” for “straight and narrow”), thoughtful exploration (TE), strategic failure (TES), and seemingly random or careless repeated actions (C), including dead-end failure ruts and clicking haphazardly around the UI. These codes (S, TE, TES, and C) are referred to throughout the findings section.

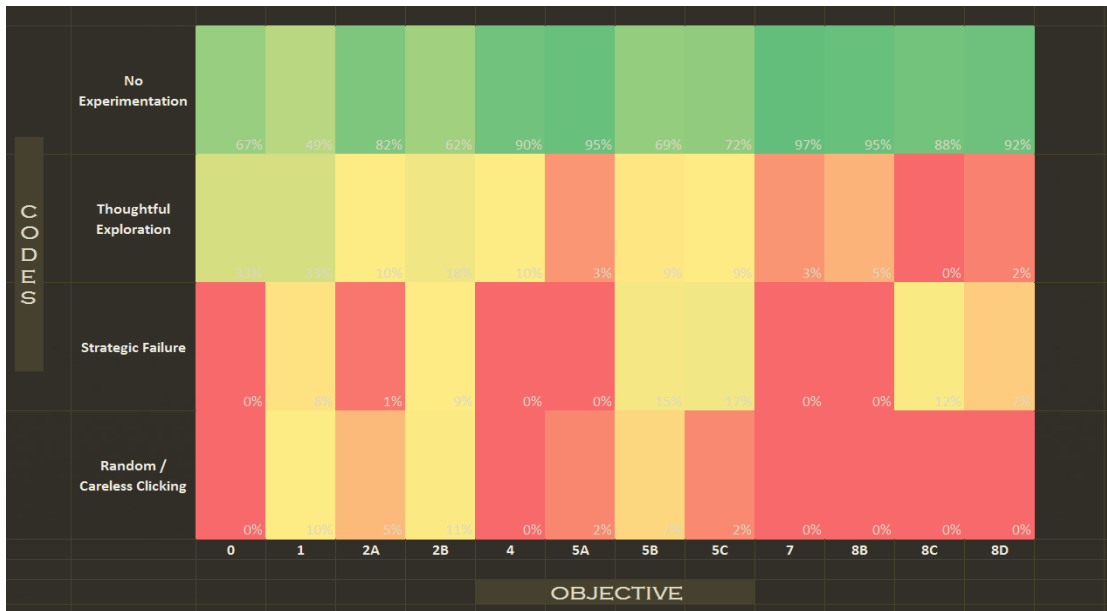


Figure 54. Aggregate heat map of exploration code frequency.

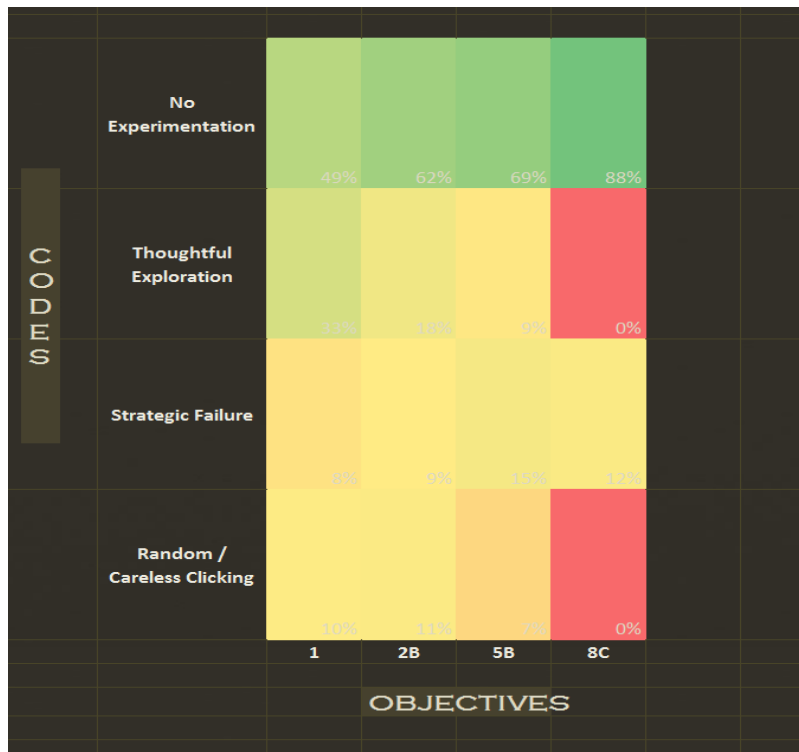


Figure 55. Map of exploration codes during cell cycles of the game (opportune for TE/TES).

Code frequencies were mapped across all coded objectives (Figure 54) as well as only phase 2 cell objectives (Figure 55), since these were inherently most affordant of TE and TES. Both show clear variation in the codes throughout the game, with a significant portion of “no experimentation” codes. The latter map specifically highlights potentially different playstyle groups, showing a clear increasing split between “no experimentation” (S) and “strategic failure” (TES) codes as the game moves forward.

Because the combinations of exploration behaviors could vary greatly, as evidenced in Figure 54, deeper investigation of codes in sequence were necessary to explore relationships with learning. Of interest especially was strategic failure (TES) in connection with other codes, since it required exploration of game mechanics deep enough to master and leverage them for metacognitive strategic ends. Because of the nature of its use in tandem with “legit” cycles (Figure 49), TES often occurred with a Straight and Narrow successful objective either just before or after it. This deliberate TES-S sequence was one that ultimately showed a positive relationship to learning, as seen in the findings below.

Sequential Codes: Nuanced Relationships with Learning

The overall findings quickly made it clear that there are great variations in patterns throughout the game, as well as great variation in the sequences of codes. Insight into the order of codes throughout play, and identification of pattern groups, were next explored to unearth deeper connections with failure and learning.

To help understand some of these variations, and unlock context of TES for greater learning insight, the codes were investigated in specific sequences for study throughout the game. To do this, several methods of feature engineering, descriptive analytics, and nonparametric statistics. These three analyses are described in detail below.

First, sequences of codes were distilled as data features for each student, and then used in the building of Markov models to show learners' likelihood of moving between codes. For example, a student who finished the game would have had 12 TE/Non-TE codes assigned to them (one for each subobjective, as shown in Figure 44). The same four base types detailed in Table 13 were used: TE, TES, C, and S&N ("S" for short). Jane's¹⁰ gameplay, for instance, may have started with no experimentation, changed to experimentation mid-game, and then focused on strategic failure for the duration. In this case, her string of 12 play codes might look like: S-S-S-S-TE-TE-TE-TE-TE-TE-TES-TES-TES. In this manner, each students' sequence of codes was defined, and separated into upper and lower quartile learning groups to help understand codes' relationship to learning gains. Two Markov models were then built (one of each quartile) in NetLogo (Wilensky, 1999) using the Markov algorithm (Berland, 2012) to better understand the connection between learning and transitions between codes (Figure 56). The models showed the probability of a student moving from experimentation (TE) in one objective to careless clicking (C) in the next, and from careless clicking to strategic failure (TES) in the next objective, and so on (bi-directionally for all four code variants). The probabilities for each learning quartile were then contrasted, with the differences highlighting moves characteristic of the greater learning group. The learning quartiles are defined consistently throughout this dissertation (see Chapter Three, or the intro of Results and Findings II in this chapter for detail).

¹⁰ Fictional player with a fictional name in a fictional string of play.

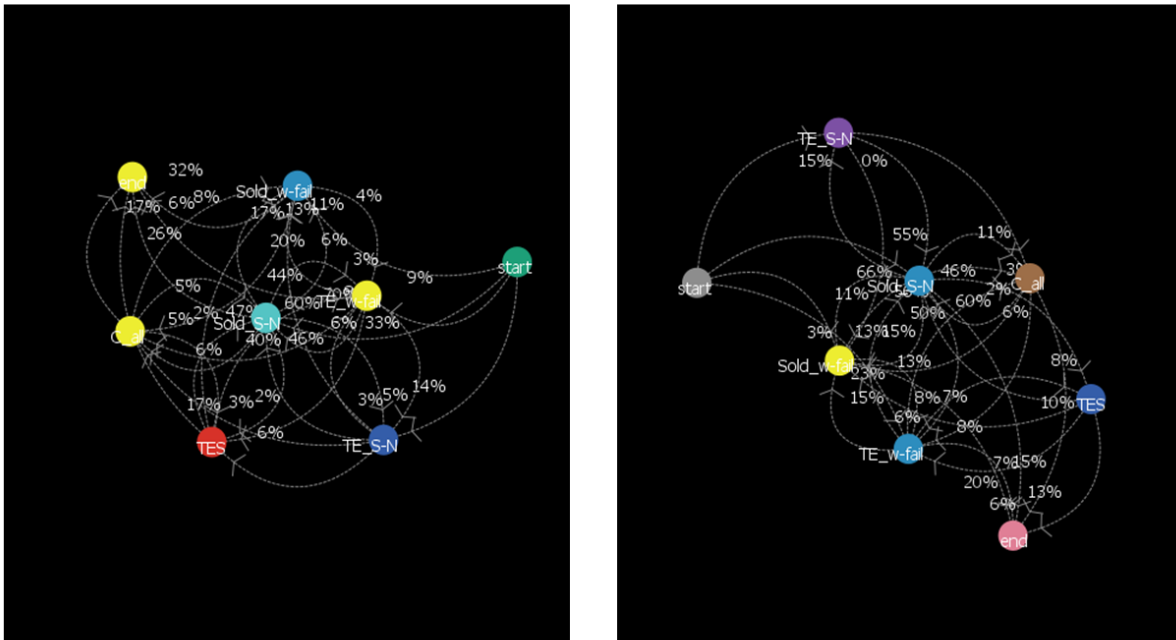


Figure 56. Markov modeling of code transitions in lower and upper quartiles (respectively).

In the second analysis of exploration and learning, features were distilled from the TE code data (Table 13), detailing first and second-order sequences of exploration codes. Taking Jane's code sequence from the previous paragraph, a first-order sequence for her (starting from objective 0) would look like: S-S. To capture sequences of codes at different points of play, these two-code combinations were taken from every possible objective point (starting with Objective 0, then Objective 1, Objective 2A....etc). Jane's hypothetical example of this first-order sequence (starting from each possible objective) is shown in Table 16. In Objective 0, her S-S sequence means that she started with "S" (no experimentation) in this level and went on to "S" again in the next level (Objective 1). Her next S-S sequence, starting in Objective 1, means that she had an "S" code in this level, and went on to Objective 2 with another "S" (and so on, throughout all possible starting points). These segments are merely a breakdown of her total experimentation code sequence of 12 (S-S-S-S-TE-TE-TE-TE-TE-TE-TES-TES-TES).

Table 16

First-Order Code Sequences, Starting from Each Objective: An Example with Jane

	Obj 0	Obj 1	Obj 2A	Obj 2B	Obj 4	Obj 5A	Obj 5B	Obj 5C	Obj 8B	Obj 8C
Jane's codes	S-S	S-S	S-S	S-TE	TE-TE	TE-TE	TE-TE	TE-TE	TES-TE	TES-TE

These code sequences were created for all students, not only for a first order chain (code-code) but for a second order chain with three codes (code-code-code). A snapshot of these second-order code chains from each possible starting point for each player is shown in Figure 57. For analysis of these new data features, the frequency of each sequence of codes (from each possible starting point) was calculated for each quartile of learners. This was to get an aggregate sense of exploration sequences, from specific starting objective points, that characterized each learning group. The frequencies (both first-order and second-order) of each learner group were then subtracted from one another, resulting in a matrix which showed the main differences in exploration sequence (specific to starting point) between the learner quartiles. This new matrix of differences was converted to a heat map (Figure 58) to more clearly show UQ tendencies (green) and LQ tendencies (red). (Heat maps were done for both the first-order and second-order features; for illustration purposes, the first-order heat map is shown here.) This way, learner exploration patterns could be seen in context of specific game objectives (information not conveyed in the context-free Markov model).

Player	Full game sequence	obj0-3letter	obj1-3letter	obj2A-3letter	obj2B-3letter	obj4-3letter	obj5A-3letter	obj5B-3letter	obj5C-3letter	obj7-3letter	obj8B-3letter
c_1	S-C-----	S-C-	C--	--	--	--	--	--	--	--	--
c_10	S-C-S-C-----	S-C-S	C-S-C	S-C-	C--	--	--	--	--	--	--
c_11	S-T1-C-S-T2----	S-T1-C	T1-C-C	C-C-S	C-S-S	S-S-S	S-S-T2	S-T2-	T2--	--	--
c_2	S-T1-S-T1-S----	S-T1-S	T1-S-T1	S-T1-S	T1-S-	S--	--	--	--	--	--
c_3	S-T1-S-C-T2----	S-T1-T1	T1-T1-T1	T1-T1-S	T1-S-S	S-S-C	S-C-T2	C-T2-	T2--	--	--
c_4	S-T1-S-----	S-T1-S	T1-S-S	S-S-	S--	--	--	--	--	--	--
c_5	T1-S-----	T1-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S
c_6	S-T1-S-C-S----	S-T1-S	T1-S-S	S-S-S	S-S-S	S-S-C	S-C-S	C-S-	S--	--	--
c_7	T1-S-T2-S-----	T1-S-S	S-S-T2	S-T2-S	T2-S-	S--	--	--	--	--	--
c_8	S-T1-S-T2-S----	S-S-S	S-S-T1	S-T1-S	T1-S-S	S-S-T2	S-T2-S	T2-S-S	S-S-	S--	--
c_9	S-T1-S-T1-----	S-T1-S	T1-S-T1	S-T1-	T1--	--	--	--	--	--	--
d_1	S-T1-S-----	S-T1-S	T1-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-	S--	--	--
d_2	S-T1-S-T1-S----	S-T1-S	T1-S-T1	S-T1-T1	T1-T1-S	T1-S-S	S-S-S	S-S-	S--	--	--
d_3	T1-S-T1-S-----	T1-S-S	S-S-S	S-S-S	S-S-T1	S-T1-S	T1-S-S	S-S-	S--	--	--
d_4	S-T2-S-----	S-T2-S	T2-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S	S-S-S

Figure 57. A snapshot of the second-order code sequences of each player (about 10% of total log shown).

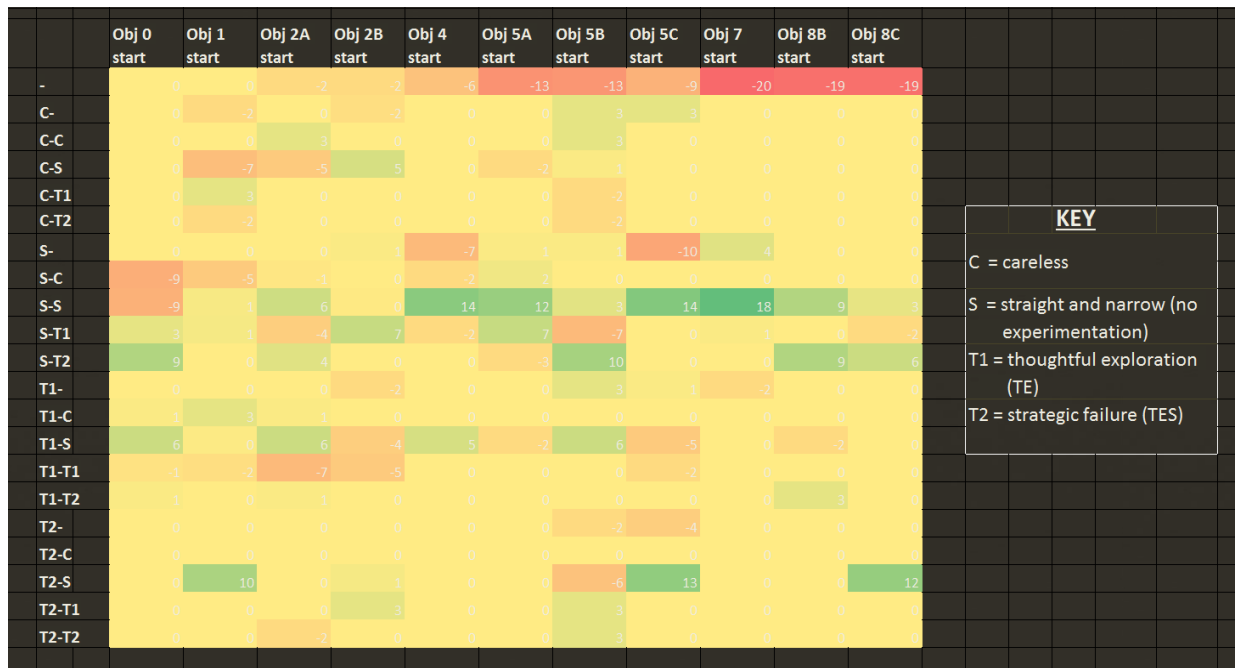


Figure 58. Contrast of sequenced-code frequency between upper and lower quartile learner groups.

Lastly, these new first- and second-order telemetry features were then compared with learning outcomes in nonparametric correlation, and the upper and lower quartile learner groups were contrasted through ranked mean comparison (via two-sample Wilcoxon). These statistics helped to support descriptive patterns found in the first two analyses.

Results from these three combined methods emerged in findings involving each exploration code. The following paragraphs will discuss different TE codes' relationship to learning throughout the progression of play.

Findings: Exploration Code Sequence and Learning

Exploration code sequences were visualized in the heat map and Markov models (Figures 57 and 55). These visualizations were used to graphically identify trends for exploration with correlation and mean comparison. One interesting result included the non-exploration codes (carelessness and straight-and-narrow) as negatively connected with learning. Specifically, when paired with a general TE behavior in a C-S-TE string, this sequence was negatively correlated with learning gains ($r=-.212$; $q=.047$). Reinforcing the trend, 71% of players who had ONLY "C" and "S" codes (no TE or TES) were lower quartile learners. Thus, it seems that carelessness without long-term strategic thinking was negatively connected with learning.

The second main finding was that TES was positively connected with learning in a number of combinations. Essentially, TES paired with non-experimentation (either before or after) was positively correlated with learning gains. This makes sense with the use of TES in a strategic arc involving a clear sequence of cycle mastery (see Figure 49). For example, an S-TES-S sequence was 12% higher in frequency on average for the upper quartile of learners than for the lower. This implies that TES was a vital strategy during these cell-cell-tissue sequences

characteristic of the greater learning group. The upper quartile had significantly higher ($q=.047$) frequency of the S-TES transition. Correlation supports this trend, showing that transitioning from “no experimentation” to “TES” (S – TES) in consecutive subobjectives was also positively correlated with learning ($r=.244$; $q=.047$). Similarly, a transition from TES to “no experimentation” was also positively associated with content gains ($r=.188$; $q=.046$). The Markov model also shows a positive connection between strategic experimentation and learning, marking the same TES to “no experimentation” as 20% higher in the upper quartile. Thus, it seems that S-TES and TES-S sequences are states of exploration more characteristic of learning, especially in cell-to-tissue levels.

Overall, understanding the connections of thoughtful exploration to learning give us insight on the roles of no experimentation, carelessness, strategic failure, and thoughtful exploring relative to sequence and game context. These findings help highlight some organic, descriptive trends in the data for informing future studies, and support overall inferences about strategic failure and learning. A model of strategic failure is next, deepening the investigation of strategic use of explored mechanics during play.

Results and Findings III: Predicting Learning-Supportive Strategic Failure

Given its recurring connection with learning, TES was chosen as the outcome variable in a detector of strategic failure. The clear use of TES by players was particularly of interest because it requires multiple layers of understanding: first, it indicates an intricate knowledge of the game’s failure mechanisms, and secondly, it implies a level of metacognition in employing those mechanisms in a deliberate success strategy. It may be this complexity of strategic thinking that supports connections with learning. In tracking gameplay actions characteristic of strategic

failure, it is possible to pinpoint a very specific combination of in-game moves which define this emergent, alternative learning trajectory. It's possible to predictively model this strategy using the EDM detector method, described in detail at the beginning of the chapter.

In order to predict this strategic thinking based on event-stream player action, a J48 classification algorithm was employed in WEKA (Hall et al., 2009) with a binary “TES / no TES” code as the dependent variable. Independent variables were based on the features embedded in the text replay clips, chosen carefully as potential indicators of student exploration (see coding section for more detail). These included several of the fundamental features distilled in Chapter Five with the intricate grain size of Chapter Six study, including cycle starts, kinds of cell collected, and start-collect combinations on the cycle-by-cycle subobjective level. Results of the analysis were evaluated using Cohen's Kappa, with cross-validation using the LOOCV (Leave One Out Cross Validation) at the student level (the overall TES level of analysis).

Overall, the model achieved a Kappa of .71 after student-level cross-validation, comparable to similar learning game detector studies (e.g. DiCerbo & Kidwai, 2013; Asbell-Clarke, Rowe & Sylvan, 2013). This value indicated the accuracy of the detector was 71% better than chance. The A' value was .87, signifying that the detector could correctly classify whether a clip contained strategic failure 87% of the time.

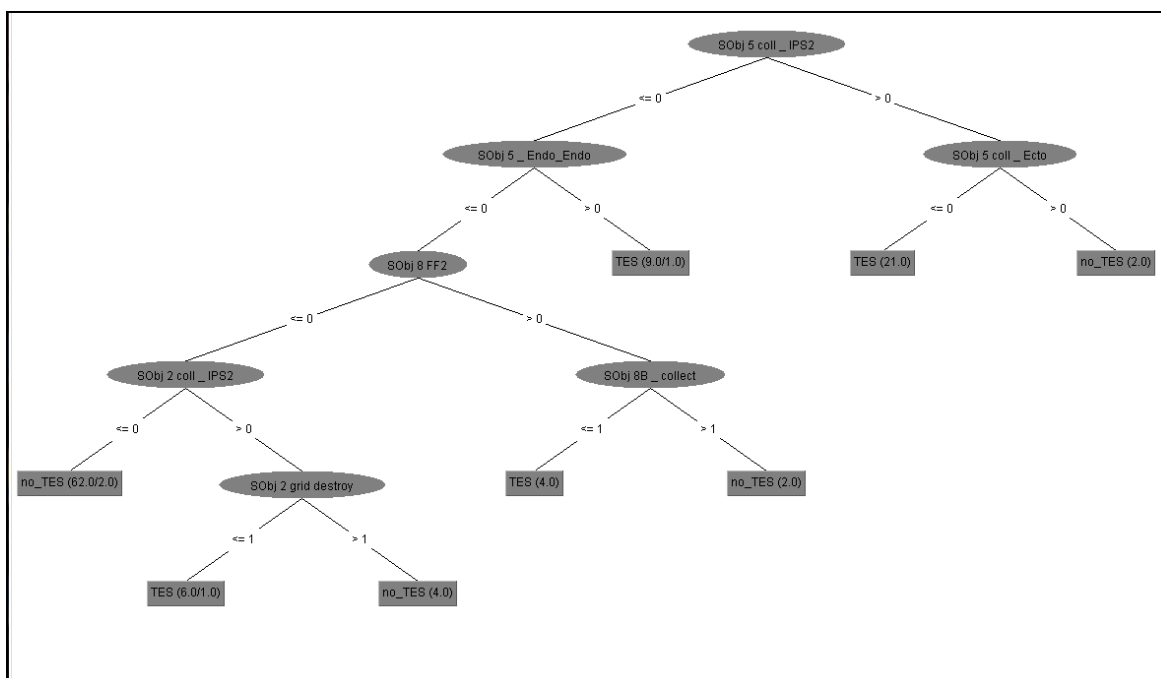


Figure 59. J48 Prediction Model – Detector of Strategic Failure

The results of this model are discussed below, starting with the top tier of the model, Objective 5 collects of iPS cells.

- The highest level predictor of strategic failure was the collection of iPS cells in Objective 5. As discussed in previous chapters, Objective 5 is the first synthesis of non-scaffolded tissue and cell cycles (presented in a cell-cell-tissue sequence). IPS cells are the core stem cells needed to build new tissue for the ailing zombies, and thus represent a core biology concept of the game. If students collected any of these cells (indicative of the “harvesting” strategy) in Objective 5, but avoided collecting the wrong kind of cell (“Ectoderm”), it was an indication of strategic exploration. However, if they did collect these wrong ectoderm cells, it decidedly put them in the “non_TES” category.

- Another branch of the tree on the second-to-highest tier has correct cells collected – “Endoderm” cells. Specifically, if the students populated the grid with endoderm cells, and simply moved them around to multiply them before collecting them again (a behavior now dubbed “harvesting”), then they are engaging in strategic exploration behavior. (This behavior is represented above in the abbreviation “Endo_Endo”.) If not, we move to the third tier of the tree. This tier is reached if players 1) have not collected Objective 5 iPS cells, and then 2) not engaged in endoderm harvesting. This tier directly addresses the failure element of the strategy, showing that an efficient fail-success schema in Objective 8 (the final level) signals purposeful failure.
- The right branch of tier three shows that if there is any far failure in Objective 8, and it is paired with only ONE (successful) cycle in the middle of the level (8B), then students are using this failure as part of TES. However, if there is more than one collect in the middle boss level (8-B), students are classified as NOT using TES; in this case, far failure more likely signals...well, just failure. The left branch of the third tier displays a path of students with NO Objective 8 far failure. If they did not collect any iPS cells in Objective 8, then they fall into the Non_TES category. If they DID collect these cells, and had very few near failures (1 or less), it was again a sign of purposeful, strategic use of failure. Conversely, if players at this point had more than 1 near failure, they were likely not practicing strategic exploration (Non_TES).

In short, players were classified as practicing TES if they met one of the following four conditions:

- 1) They collected iPS cells with NO wrong “Ectoderm” cell collects in Objective 5.

- 2) They did NOT collect Objective 5 iPS cells, but did have one or more endoderm-endoderm (start-end) cycles in Objective 5 (a “harvesting” behavior).
- 3) They had NO iPS collects and no endoderm-endoderm cycles in Objective 5, but had one or more Objective 8 far failure cycles, with only 0-1 collects in the middle of Objective 8.
- 4) They had NO Objective 5 iPS collects or endoderm-endoderm cycles, and no Objective 8 far failure. Beyond these thresholds, if they collected any iPS cells in Objective 2 (a “harvesting” behavior) with *low* near failure in that same objective (one or less), they were classified as TES.

Thus, this predictive model implies that specific use of far failure (via iPS and endoderm harvesting behavior), accompanied with low amounts of incidental failure (e.g. near failure) generally characterize the use of strategic failure in the *Progenitor* gamespace. This emphasizes earlier analyses’ findings that failure is not monolithic, and each failure type has an evolving relationship to learning throughout play. Essentially, this model of strategic failure operationalizes transgressive play in the *Progenitor* gamespace and captures it in positive characterization of learning.

Conclusion

These findings define predictive models of experimentation in the *Progenitor* gamespace, identifying one kind of strategic failure in particular significantly associated with learning gains. TES operationalizes a form of transgressive play in the gamespace, and connects this sort of play-based testing of limits positively with learning. Detectors such as these are powerful tools in informing future game design, able to provide real-time differentiation between productive failure (c.f. Kapur, 2006) and unfocused floundering in the gamespace. Distinctions like these can inform design of game cues to support learner trajectory along these organic critical

pathways to learning. Findings of this chapter also build on, and reinforce, the previous chapters' themes of early game failure ruts, the importance of the mid-game tissue levels, and the fascinating evolving role of far failure throughout play.

Tying into the previous two studies, this analysis clarifies the role of far failure, near failure, and success in patterns of experimentation, and ultimately learning. Building on the feature engineering and performance trends over time mapped in analysis one, and the play progression states visualized in analysis two, this section can supplement our understanding of performance and learner behavior relative to the immersive, exploration-friendly context of educational gamespaces.

Chapter Seven: Conclusions and Future Work

In understanding learner trajectories within game microworlds as designed experiences (c.f. Rieber, 1996; Squire, 2006), this dissertation looked through the lenses of play purpose (games as playful medium), instructional purpose (learning games as content delivery systems), and individual purpose (the play style and subjective goals each player brings to the game). The empirical arc of analysis was based in the game *Progenitor X*, and used mixed methods to examine three distinct intersections of the lenses above (Figure 1). Together, the three analyses broadly explored an overarching research question: what kinds of naturalistic player interaction with the educational gamespace (including play progression, in-game success, shades of failure, and experimentation) characterize learning?

Each of the analyses explored its own research question and intersection of lenses. Specifically, the first analysis used descriptive and nonparametric statistics (with specially engineered features) to explore the intersection between the learning game as content delivery and individual player choices. To do this, it identified procedural biology content, translated to specific verbs of play, and engineered data which showed student performance on these key tasks. Statistical analysis of these success and nuanced failure patterns then explored the research question: how does fine-grained, context-specific game performance (including shades of failure and success) connect with learning outcomes? The second analysis focused on the intersection of the game as a designed arc of play, and student choices in navigating this gamespace. Machine learning analysis was used to study player progress through each sequential cycle of the game, exploring learner pathways in play progression. Markov models were used, because they can illustrate the probability of players moving, in time order, from one level to another. For each set of progress data, two Markov chain models were made: one for the upper quartile, and one for

the lower quartile of learners. Contrasting the two quartile models of play directly addressed the second research question: how does organic play progression differ between groups of learners? The last analysis followed with a question of natural corollary: what play data features (both of performance and progression) characterize experimentation in *Progenitor X*, and how does this behavior connect with learning outcomes? Resting at the intersection of all three lenses, this third analysis focused on player agency in employing strategic performance (on academic content mechanics), while optimizing their pathways through the experimentation-encouraging medium of the game. To study this, it drew on educational data mining to build a predictive detector of player experimentation in *Progenitor X*, and then examined kinds of experimentation in relationship with learning outcomes.

These analyses gave respective insights that built three overall trends of findings in relationship to learning: harmful far failure in early levels, critical mid-game skill synthesis, and strategic failure in later-game levels (positive to learning). Broadly, these findings showed that overall play success and progress is positively connected to learning, while aggregate time on task and total failures (as a general category) were not related to learning. Some kinds of failure were, however, connected with learning patterns: specifically, tissue failure and “far” failure. Generally, tissue failure (possibly signaling extended frustration with this core mechanic) in mid-game levels had negative impact on both play completion and learning gains. Far failure in tutorial game levels were negatively related to learning, but became positively correlated with learning gains in the boss level. Identifying emergent forms of strategic failure (analysis 3), which was positively associated with learning, helped to explain the changing relationship of far failure to learning throughout the game.

The first analysis built data features that supported all three analyses, and showed that failure was not a monolithic construct in *Progenitor* gameplay and learning. In other words, kinds of failure matter, and context of that failure matter, particularly in relationship to learning. Far failure, a definition of failure as a result of acting directly contrary to game cues (e.g. wrong start or wrong collect), emerged as an important construct. Specifically, far failure's early game negative relationship to learning shifted to a positive relationship in later levels. Next, the Markov analysis created a new set of data features based on cycle-by-cycle game progression, allowing for new sequential refinement. Fundamentally, it showed differences between the learner groups consistent with the first analysis: avoiding repetition of early-level cycles (and far failure), strong tissue performance in skill synthesis levels mid-game, and consistent use of far failure in the boss level of the game were characteristic of the higher learning group. The last analysis, exploring experimentation in play and learning, built upon the performance and sequentially-detailed features of the first analyses. Aligned with the previous findings, it thematically revealed that far failure was a defining construct for learner experimentation in play. Essentially, the study was able to identify a behavior of strategic failure – deliberate use of gameplay failure for efficient objective completion – through the holistic coding of play sequence. This strategic failure, a type of experimentation in play, was positively associated with learning gains. Consistent with earlier chapters' findings, far failure in later levels of the game were positively connected with strategic failure and learning, while tissue failure and far failure in early levels were negatively related.

Implications

New Paradigms of Assessment

One major implication of ADAGE for game-based assessment is affording multiple approaches to understanding learner trajectories. ECD task models, for example, can be easily used in conjunction with ADAGE Critical Achievement structures to gather salient evidence for analysis. This approach represents a pre-formed hypothesis about an optimal learning pathway through the virtual space. Conversely, ADAGE also provides data very compatible with a purely exploratory data mining approach (as this dissertation demonstrates). Assessment data provided by ADAGE can also provide a combined confirmatory-exploratory approach, one increasingly popular in learning game analysis (e.g. Institute of Play, 2013; Baker & Clarke-Midura, 2013). Beyond basic research approach and experimental design, there are a myriad of analysis methods which can be fueled by ADAGE data. For example, ADAGE is being built out to capture textual, multi-player data, and already collects context-rich event-stream data. This myriad of interaction data can be used with classic statistics (e.g. multiple regression) and learning analytic techniques (e.g. pattern matching algorithms, predictive student modeling, association mining, or visualization). A common framework for salient assessment data aggregation across genres solves the “critical problem” of recording relevant clickstream data in the “deluge of information” that is game data (c.f. Shute, 2011). It enables more time spent on intelligent iterative game design, and facilitated connection of play patterns across games (not on reinventing the data structure “wheel” for each consecutive project).

As reviewed in Chapter Two, the ADAGE framework – as shaped by these paradigms of Evidence Centered Design and Educational Data Mining – represents new possibilities for authentic assessment in virtual learning spaces.

Today, games and other digital media allow us to track progress on multiple variables to gauge growth across time and to discover different trajectories towards mastery and innovation compared and contrasted across thousands of learners.... A single score on a standardized test taken on one day—a “drop out of the sky test”—will come to look not just thin, but unethical. (Gee, 2012, p.2)

Indeed, current “drop out of the sky”, high-stakes, annual multiple choice tests currently reinforce several arguably narrow views of learning. First, current “testing teaches there are right answers” – specifically, only *one* right answer (and presumably thinking process) per problem (Shank, 2011, p. 80). It also fixates on factual memorization (recalling Dewey’s century-old warning (1938) about the school system’s fact fetish), decontextualizes assessment from an authentic learning context, and teaches us that some subjects are compartmentalized, with some more important than others (Shank, 2011). Conversely, interaction-based assessment in digital worlds can afford learner agency in exploring multiple solutions through an interactive thinking process. This virtual, learner-centric mining of the interactive event stream can help provide formative assessment as feedback in the learning process, emphasizing cross-subject problem-solving in worlds where learning context and assessment are seamlessly integrated. Thus, assessment based on interaction mining in virtual learning spaces can help move us to a process-based assessment system rather than a knowledge-based one (c.f. Shank, 2011; Behrens et al., 2012). New paradigms like this help answer an increasing call for alternative, digitally based performance assessment from as high up as the White House (President’s Council of Advisors on Science and Technology, 2013). This movement is also supported by digital assessment data consortiums in top tier research institutions (e.g. GLS reference) and in recent efforts by assessment giants like ETS and ACT to expand research in virtual-world event-stream data analytics (e.g. Institute of Play, 2013; Encarnacao, 2014).

As this assessment movement gains momentum, it could change the face of education. The reason for this is that “assessment, especially when coupled with accountability, drives how we teach and learn” (Gee, 2012, p.1). If we can develop robust yet authentic assessment (embedded with formative feedback), with learner agency and creativity in problem solving with multiple solutions – all in virtual worlds where subjects can overlap and even share solution strategies – then this is exactly the kind of teaching and learning that will be incentivized in our schools.

Failure and Learning

Within these future paradigms of assessment , this research supports changing empirical understandings of the nature of failure and its role in learning. Technology-supported inquiry learning (Edelson, Gordin & Pea, 1999), for example, specifically highlights the role of the discovery and refinement of knowledge based on student exploration of content via hypothesis testing; necessarily, this involves failure as feedback in defining knowledge boundaries and problem constraints. In this sense, the consideration of failure in a positive role is vital to authentic assessment within broad instructional schemas like discovery-based learning. Ranging from these constructivist methods to more direct instruction, the “assistance dilemma” (Koedinger et al., 2008) questions whether failure minimization via greater assistance and heavy scaffolding is always better for learning. Kapur (2006) directly builds on this research, empirically establishing the construct of “productive failure” in minimally-scaffolded learning contexts. Jim Gee and Jesper Juul both assert, respectively, that failure in video games is designed to be pleasantly frustrating and often serves to fuel self-regulated learning (Gee, 2005a; Juul, 2013).

By design, games as a medium support productive failure. Boundary testing, and response to failure as formative feedback, is in an essential part of the game experience (Schell, 2008; Squire, 2006). Salen & Zimmerman (2004) identify limit testing as part of the profile of the “dedicated player,” who fails frequently because they intentionally do not follow the rules, but is engaged in play and interested in strategically understanding the underlying rule constraints of the game (p. 268). (This dedicated player stands in contrast to the “spoilsport” or the “cheat,” both of whom would be likely to engage in “WTF” behavior or gaming the system (e.g. Wixon et al., 2012; Baker, Corbett & Koedinger, 2004).) In the domain of transgressive play, or purposeful rule-breaking, Salen & Zimmerman assert that “rule-breaking can enhance meaningful play,” because “to strategically break rules requires an intimate knowledge of the rules themselves” (Salen & Zimmerman, 2004, p.281-282). This dissertation, for example, empirically demonstrates the existence of strategic failure, and in positive relationship to learning – which implies the sort of agency and cognitive engagement characteristic of “dedicated” transgressive play. Failure is, indeed, an incentive for pushing forward in play (Juul, 2013) and serves to fuel the core learning principle of “pleasant frustration” in video games (James Paul Gee, 2003). Failure as productive play is a vital consideration in our assessment of learning in games, particularly in considering multiple “optimal pathways” through the learning space (Ramirez et al., 2012; Owen & Ramirez, in submission).

This dissertation’s insights into failure afford vital insight for such authentic assessment of learning in virtual learning spaces. The first pivotal understanding is that failure is not monolithic; it does not always simply exist in one form throughout the entirety of a learning arc. In this study, different types of failure were worth distinguishing, both in terms of learning and play progress. A second key insight is that context of failure matters. Each kind of failure was

examined at multiple points of gameplay, and the relationship of each to learning varied based on the context. Far failure, for example, went from having a negative association with learning in the beginning of the game to a positive relationship towards the end. This evolution towards strategic failure implies the agency and engagement of “dedicated” transgressive play (Salen & Zimmerman, 2004). Thirdly, failure can be beneficial to learning. Strategic failure in *Progenitor X*, for example, was positively connected with learning outcomes. Failure, in this case, became vital for assessment purposes because it signaled significant learning gains. Nuanced failure in context-specific performance, thus, can be key in understanding (and assessing) transgressive pathways optimized for both play and learning.

Data-Driven Learning Design

The ADAGE framework and its analytic affordances have major implications for iterative game design optimized for play and learning. In the design of learning game environments, experts assert that players rarely interact with the game in exactly the way the designers envision, and thus heavily emphasize early, repeated usertesting (Schell, 2008; Salen & Zimmerman, 2004). With the added element of content-specific learning goals, or concrete growth over time in a domain-specific skill, attending and adjusting to organic play patterns becomes even more vital (c.f. Shute, 2011; Norton, 2008; Institute of Play, 2013). Specifically, through designed assessment structures, and through application of data output to various data mining methods, assessment frameworks like ADAGE can powerfully fuel a data-driven game design process that optimizes learner experience from the earliest development stages. They can use telemetry-based assessment structures and applied learning analytics to inform three stages of development: initial core design, alpha and early beta usertesting, and final design overlay of learner-adaptive gameplay.

In the early design process, using a learning game data framework can not only help pinpoint existing data features for usertesting insight and later analysis – it can inspire the creation of helpful learning measures in-game (like ADAGE “Critical Achievements”, or CAs). Moments considered important for measuring learning can be built into the design process, crafted with the goal of informative yet seamless feature of gameplay. One example of the Critical Achievement data structure supporting early design is in *Crystals of Kaydor*, a game in the *Tenacity* collaboration with GLS and the Center for Investigating Healthy Minds. *Crystals of Kaydor* is an RPG designed to cultivate the development of pro-social behavior through collaborative social interactions. The player controls a robot who has crash-landed on an alien planet. For the first kind of CA, in order to win the aliens’ trust, the player must pay close attention to non-verbal cues, tracking aliens’ facial expressions and intensity through a slider interface. Secondly, the player must then correctly select the emotion of the alien, and for the last CA, choose an emotional response to the aliens’ affect. These two CAs correspond directly with the game’s content model of teaching awareness of non-verbal cues and emotion in others. In tandem with these CAs, ADAGE play progression data has also provided a context-rich backdrop to evaluate play progression in relationship to learning (Beall et al., 2013). Building in formative assessment (like CAs) in initial phases of game design (rather than clunky late-game additions or identified post-hoc by desperate researchers) has several advantages. First, it helps beautifully integrate play progression and learning measurement mechanics for a seamless player experience. Second, these designer-specified mechanics directly inform data structures and early-phase analytics, making usertesting results even more relevant to developers. Thirdly, these key learning mechanics provide anchoring measurement points for educational researchers, who can

then provide insight into growth patterns that inform final in-game scaffolding design (see adaptive design below).

Strong data structures also enable telemetry analysis for data-driven design in the alpha and early beta phases. Visualizations and descriptive analytics can be particularly helpful in refining UI design, as well as identifying bugs and player attrition points. All of these analytics, based in click-stream data, can greatly complement qualitative usertesting methods like interviews, surveys, and think-alouds. Specifically, visualization of well-structured telemetry data can be a powerful tool in identifying bugs and player attrition points. A similar example exists for early design and testing of *Fair Play*¹¹, a game about implicit bias in graduate-level academic institutions. In this game, positional telemetry data was recorded to create a heatmap of player activity. One map level (aerial view) in the game (Figure 60) was programmed to show areas of frequent player travel in red, and areas least traveled by students in blue (Owen & Ramirez, in submission). This helped inform placement of in-game assets critical to content exposure and game advancement.

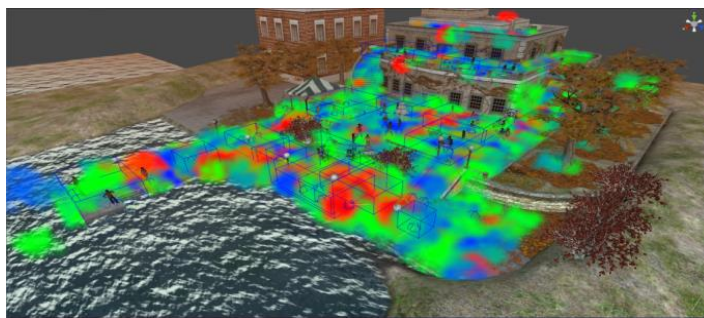


Figure 60. *Fair Play* heat map of click-stream player activity

¹¹ <http://www.gameslearningsociety.org/fairplay/>

In final stages of game development, after extensive data collection with late-beta builds, ADAGE-enabled learning analytics can be used to predict in-game actions and performance most characteristic of learning. This knowledge of ideal player behavior can then inform the final design phase: user-adaptive, fully scaffolded play for optimized in-game learning. To this end, optimal player actions, sequential pathways, and assessment growth trajectories can each be explored through learning analytics (including visualization, prediction, and pattern mining methods). These categories are based on the extensive literature review in Chapter Two (Figure 11), and examples of each (relevant to adaptive play design) are given below:

- This dissertation's early correlation of in-game success and failure with pre-post learning outcomes helps define red flags of far failure (negative to learning) in tutorial levels.
- Predictive modeling of experimentation supportive of learning is also modeled in this study's arc, using a detector methodology with classification and regression trees (CART) to predict in-game performance and learning (e.g. DiCerbo & Kidwai, 2013; Baker & Clarke-Midura, 2013).
- In similar research, Bayesian networks have also been used with this data theme, probabilistically connecting chunked performance data to creative problem solving in games (e.g. Shute, 2011). Though not yet widely used in games, Bayesian Knowledge Tracing (BKT) is an applicable algorithm that can predict learning moment-by-moment based on multiple performances on a chosen task (e.g. Baker, Corbett, & Alevan, 2008). If gameplay models show certain actions at certain points to be more predictive of learning, then player-triggered scaffolding (e.g. help resources) can be implemented in-game to help keep players on track at these crucial points.

- Sequential learner pathways, using salient event-stream data, can also be modeled using machine learning methods. Specifically, visualization and predictive modeling (including cluster analysis and pattern-mining techniques) have been used with success in learning games research to capture learner trajectories. In this study, the Markov models of play trajectory are a prime example (Chapter Five). In other relationship mining through visualization, SimCity.EDU researchers are currently building player profiles by identifying groups using hierarchical cluster analysis (Institute of Play, 2013). In another visualization example, ADAGE-based heatmaps can visualize learners' critical pathways through the game (e.g. Owen, 2014).

A concrete example of final-stage adaptive play design can be given based on this study's findings. The predictive models of experimentation and strategic failure in Chapter Six can identify real-time transgressive play patterns supportive of learning – and those NOT helpful to students. With an automated detector of strategic failure supportive of learning, an additional layer of game code could be added which encourages players along this path (and differentiates it from the kind of failure characteristic of attrition or lower learning, helping those players recover as well.) Overall, in application to games, these ADAGE-fueled assessment analytics can help designers anticipate and support in-game performance indicative of learning.

Future Work

ADAGE provides a flexible, cross-genre assessment data framework that supports multiple evaluation approaches and analysis methods. Schemas like ADAGE, defining salient information in the event-stream digital deluge of data, are vital for new paradigms of assessment made possible in this digital age of education (c.f. Behrens et al., 2012; Shute, 2011;

Steinkuehler, Barab & Squire, 2012). In future work, the growth and development of ADAGE and in-game assessment can improve with larger-scale research around multiple contexts and audiences for application.

Progenitor X began at GLS in 2011 as an idea brainstorm, led by Kurt Squire, for the National STEM Video Game Challenge. Lead game designer on the game was Mike Beall, working closely with programmers Ted Lauterbach and Greg Vaughan. *Progenitor* soon became a part of CyberSTEM, a GLS game assessment project (led at the time by Rich Halverson and Ben Shapiro). Since the game was being developed in-house concurrent with CyberSTEM, it became an ideal genesis point for the first clickstream embedded assessment study. Hence, the zombie-ridden biology game and GLS in-game assessment rose up together from mutual fertile ground. Out of playful undead tissue regeneration, thus, sprung the first version of ADAGE in 2012. Allison Salmon was the programmer who made possible the translation of the conceptual assessment frame of ADAGE to an actual implementable API. Hence, *Progenitor X* was the original click-stream assessment game of study, which worked well for pioneering the assessment framework and methods applications central to this research. However, in future work, research with expanded sample size and game genres promises to yield further insight. ADAGE has now been implemented in eight different GLS games, and has its own open-source, user-friendly website (www.adageapi.org). IRB permissions have just expanded to permit all anonymized clickstream interaction with any GLS game to be used for study (including any remote use by anyone on the internet). The foundational assessment framework and methods blueprint presented in this dissertation thus help support expansive, larger-scale future studies imminent with ADAGE.

Looking beyond n-size, the future of ADAGE is perhaps most exciting in the potential for integration across multiple interaction data sources, contexts, and audiences. Event-stream game data is a rich source of information which can be leveraged for even deeper insights with other forms of player interaction data (including observational/video, interview, survey, in-game discourse, and physiological sensor data). Powerful insights about learning in play can be made with the synchronization of these multiple sources, each one a part of a larger ecology of interaction data. To broaden in-game event stream information, ADAGE is moving from player-game records to player-player interaction structures, being built out for the multi-player GLS game *Trails Forward*. Indeed, the boxed game itself is just a small part of a larger “big G” Game ecology that involves community discourse and collective intelligence around the game (Gee, 2003; Jenkins, 2006; Steinkuehler, 2006). Similarly, a way to capture player-player interactions *outside* of the game is to study asynchronous player data – like forum posts, modding, machinima, and other game-centered community artifacts. Integration of these affinity space data (James Paul Gee, 2005b) with in-game interaction data can support even greater insight on play and learning on the “big G” Game community level.

In addition to future integration of interaction data sources and larger community context, the use of assessment data for multiple audiences is critical. In ideal future work, game-based assessment data should be collected, processed, and then exported for the benefit of several parties: researchers, students (in understanding their own play), developers, and facilitators (including parents and teachers). For researchers, future work might leverage deeper analytic methods in the machine learning categories of visualization, relationship mining, and prediction (Figure 11). For example, heat maps of critical paths and automated correlations of performance with learning outcomes can provide powerful tools for understanding student behavior (and are

planned as ADAGE-automated features). Connecting research methods and players, in-game formative feedback cycles can be fueled by intelligent learning models built into game code. One example is Bayesian Knowledge Tracing, a probability algorithm modeling learning moment-by-moment (e.g. Baker et al., 2011) that is ideal for informing adaptive “help” resource scaffolding during play. Students, especially with the development of an ADAGE student portal, should be able to see visualizations of their own progress, earn badges and achievements, and benefit from data-informed adaptive gameplay. Game development can be informed by base ADAGE assessment mechanics, as well as many machine learning analyses – including heat map visualizations, network diagrams of player navigation, context-specific features most highly correlated to learning, and moment-by-moment detectors of desired behavior – in early and iterative design stages. Facilitators, with a future ADAGE portal, should be able to see their pupils’ progress and more easily support and group students according to data-driven recommendations. Customizing assessment output for each member of the audience ecosystem can help create a sense of agency in all stakeholders – and inform new paradigms for integrated assessment design, collection, analysis, and iterative application optimizing play-based learning.

Final Summary and Conclusion

This dissertation supports ADAGE as an assessment data framework that advances new paradigms in the way we understand student learning in play. Empirically, this research demonstrates cross-method application of ADAGE assessment through the lenses of game microworlds as designed experience. ADAGE-based findings differentiate types and context of failure, reveal experimentation patterns, and demonstrate the positive relationship between strategic failure and learning. The ADAGE-based mining of these unexpected player pathways through the learning space have powerful implications for defining alternate learner pathways in

new assessment paradigms, reconsidering the role of (non-monolithic) failure in formal learning evaluation, and informing iterative educational game design for the optimization of learner-adaptive play. Ultimately, these insights can fuel new empowerment of researchers, designers, and facilitators in providing engaging, interactive, learner-adaptive play environments for those to whom the future of education belongs: our students.

References

- Anaya, A. R., & Boticario, J. G. (2011). Content-free collaborative learning modeling using data mining. *User Modeling and User-Adapted Interaction*, 21(1-2), 181–216.
doi:10.1007/s11257-010-9095-z
- Annetta, L. A., Holmes, S., Cheng, M. T., & Folta, E. (2010). Measuring student perceptions: Designing an evidenced centered activity model for a serious educational game development software. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, 2(3), 24–42.
- Arnold, A., Nallapati, R., & Cohen, W. W. (2008). Exploiting Feature Hierarchy for Transfer Learning in Named Entity Recognition. In *ACL* (pp. 245–253). Retrieved from <http://www-2.cs.cmu.edu/~wcohen/postscript/acl-2008.pdf>
- Arnold, A., Scheines, R., Beck, J. E., & Jerome, B. (2005). Time and Attention: Students, Sessions, and Tasks. (pp. 62–66). Presented at the Educational Data Mining, Pittsburgh.
- Asbell-Clarke, J., Rowe, E., & Sylvan, E. (2013). Assessment design for emergent game-based learning. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems* (pp. 679–684). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=2468476>
- Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012). The best and future uses of assessment in games. In *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 227–246).
- Baker, R. (2013). *Big Data In Education - Week 6*. Retrieved from https://www.youtube.com/watch?v=hxz9EgeO_2A
- Baker, R., & Carvalho, A. De. (2008). Labeling student behavior faster and more precisely with text replays. ... *of the 1st International Conference on*

- Baker, R., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction, 18*(3), 287–314.
- Baker, R., D’Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners’ cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*(4), 223–241.
- Baker, R. S., & Clarke-Midura, J. (2013). Predicting Successful Inquiry Learning in a Virtual Performance Assessment for Science. In *Proceedings of the 21st International Conference on User Modeling, Adaptation, and Personalization* (pp. 203–214). Retrieved from <http://www.columbia.edu/~rsb2162/UMAP-2013-BCM-v9.pdf>
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 531–540). Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-30139-4_50
- Baker, R. S., Goldstein, A. B., & Heffernan, N. T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education, 21*(1), 5–25.
- Baker, R. S., Gowda, S. M., Corbett, A. T., & Ocumpaugh, J. (2012). Towards automatically detecting whether student learning is shallow. In *Intelligent Tutoring Systems* (pp. 444–453). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-30950-2_57
- Baker, R., & Siemens, G. (2014). Educational data mining and learning analytics. *Cambridge Handbook of the Learning Sciences*: Retrieved from

<http://www.ebooksmagz.com/pdf/educational-data-mining-and-learning-analytics-columbia-university-170268.pdf>

- Baker, R. Sj., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems* (pp. 531–540). Springer Berlin Heidelberg.
- Baker, R. Sj., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. In *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems* (pp. 29–36). Retrieved from <http://www.cs.cmu.edu/afs/cs/Web/People/rsbaker/BCWFinal.pdf>
- Baker, R. Sj., & De Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 38–47). Retrieved from <http://learnlab.org/uploads/mypslc/publications/edm2008textreplayalgebrag.pdf>
- Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.
- Beall, M., Farajian, R., Owen, V. E., Slater, S., Smith, A., Solis, E., ... Davidson, R. (2013, June). *Games for Mindfulness and Pro-Social Behavior: the Tenacity Project Collaboration*. Presented at the The 9th Annual Games+Learning+Society Conference, Madison, WI.
- Behrens, J. T., & DiCerbo, K. (2013). *Technological implications for assessment ecosystems: Opportunities for digital technology to advance assessment*. Retrieved from http://researchnetwork.pearson.com/wp-content/uploads/behrens_dicerbo_technological_implications_assessments.pdf

- Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world*. Los Angeles, CA: CRESST (UCLA).
- Behrens, J. T., Mislevy, R. J., DiCerbo, K., & Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–53). Charlotte, NC: Information Age Publishing.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57(1), 289–300.
- Berland, M. (2012). *Narkov, a first-order Markov chain algorithm*.
- Berland, M., Smith, C. P., & Davis, D. (2013). Visualizing Live Collaboration in the Classroom with AMOEBA. In *CSCL 2013 Conference Proceedings* (Vol. Volume 2 — Short Papers, Panels, Posters, Demos & Community Events, pp. 2–5). Madison, WI: International Society of the Learning Sciences (ISLS).
- Boaler, J., & Greeno, J. G. (2000). Identity, agency, and knowing in mathematics worlds. In J. Boaler (Ed.), *Multiple perspectives on mathematics teaching and learning* (pp. 171–200). Westport, CT: Ablex Publishing.
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth & Brooks.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42.

- Carley, K. M. (2003). Dynamic network analysis. In *Dynamic social network modeling and analysis: Workshop summary and papers* (pp. 133–145). Retrieved from http://oobgy.googlecode.com/svn-history/r208/trunk/biyeshaji/docs/2009Institute_NA_Track_Carley_2003_dynamicnetwork.pdf
- Cetintas, S., Si, L., Xin, Y. P., & Hord, C. (2009). *Automatic Detection of Off-Task Behaviors in Intelligent Tutoring Systems with Machine Learning Techniques*. Presented at the IEEE Transactions on Learning Technologies.
- Chaffar, S., Derbali, L., & Frasson, C. (2009). Inducing positive emotional state in Intelligent Tutoring Systems. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*.
- Chen, N. S., Kinshuk, Wei, C. W., & Chen, H. J. (2008). Mining e-learning domain concept map from academic articles. *Computers & Education, 50*, 1009–1021.
- Cheng, R., & Vassileva, J. (2006). Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. *User Modeling and User-Adapted Interaction, 16*(3-4), 321–348. doi:10.1007/s11257-006-9013-6
- Clark, D. B., Martinez-Garza, M. M., Biswas, G., Luecht, R. M., & Sengupta, P. (2012). Driving Assessment of Students' Explanations in Game Dialog Using Computer-Adaptive Testing and Hidden Markov Modeling. In *Assessment in Game-Based Learning* (pp. 173–199). New York: Springer.
- Clarke-Midura, J., Code, J., Dede, C., Mayrath, M. C., & Zap, N. (2012). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based*

- assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 125–148). Charlotte, NC: Information Age Publishing.
- Clarke-Midura, J., & Yudelson, M. V. (2013). Towards Identifying Students' Causal Reasoning Using Machine Learning (pp. 704–707). Presented at the Artificial Intelligence in Education, Springer Berlin Heidelberg.
- Cocca, M., Hershkovitz, A., & Baker, R. (2009). The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (p. 507—514).
- Cohen, P. R., & Beal, C. R. (2009). Temporal Data Mining for Educational Applications. *Int. J. Software and Informatics*, 3(1), 31–46.
- Collins, A., Brown, J. S., & Newman, S. E. (1990). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser*. Hillsdale, NJ: Lawrence Erlbaum.
- Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.
- Costikyan, G. (2002). I Have No Words & I Must Design: Toward a Critical Vocabulary for Games. In *CGDC Conf.* Retrieved from <http://www.digra.org/digital-library/publications/i-have-no-words-i-must-design-toward-a-critical-vocabulary-for-games/>
- D Baker, R. S. (2010). Mining data for student models. In *Advances in intelligent tutoring systems* (pp. 323–337). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-14363-2_16

- D Baker, R. S., Corbett, A. T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems* (pp. 406–415). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-69132-7_44
- D’Mello, S. K., Craig, S. D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic Detection of Learner’s Affect from Conversational Cues. *User Modeling and User-Adapted Interaction*, 18(1-2), 45–80.
- Dabney, A., & Storey, J. D. (2004). QVALUE: The Manual Version 1.0. *University of Wasington Department of Biostatistics*. Retrieved from <http://www.bioconductor.org/packages/2.13/bioc/vignettes/qvalue/inst/doc/manual.pdf>
- Dede, C. (2012, May). *Interweaving assessments into immersive authentic simulations: Design strategies for diagnostic and instructional insights*. Presented at the Invitational Research Symposium on Technology Enhanced Assessments.
- deJong, T., Wilhelm, P., & Anjewierden, A. (2012). Inquiry and Assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.
- Dekker, G. W., Pechenizkiy, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. In *International Conference on Educational Data Mining* (pp. 41–50). Cordoba, Spain.
- Derry, S. J., & Steinkuehler, C. A. (2003). Cognitive and situative theories of learning and instruction. In (L. Nadel, Ed.) *Encyclopedia of Cognitive Science*.
- Dewey, J. (1938). *Experience and education*. New York, NY: Macmillan.

- DiCerbo, K. E., & Kidwai, K. (2013). Detecting Player Goals from Game Log Files. Presented at the 6th International Conference on Educational Data Mining. Retrieved from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_58.pdf
- Edelson, D. C., Gordin, D. N., & Pea, R. D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences*, 8(3-4), 391–450.
- Encarnacao, M. (2014). http://transformingedu.com/avada_portfolio/minguel-encarnacao-act/. Retrieved from http://transformingedu.com/avada_portfolio/minguel-encarnacao-act/
- Feng, M., Beck, J. E., & Heffernan, N. T. (2009). Using Learning Decomposition and Bootstrapping with Randomization to Compare the Impact of Different Educational Interventions on Learning. In *Proceedings* (pp. 51–60). Cordoba, Spain.
- Feng, M., Hansen, E. G., & Zapata-Rivera, D. (2009). Using Evidence Centered Design for Learning (ECDL) to examine the Assistments system. In *annual meeting of the American Educational Research Association (AERA), San Diego, California*. Retrieved from <http://web.cs.wpi.edu/~mfeng/pub/AERA-2009-Feng-Hansen-Zapata.pdf>
- Feng, M., & Heffernan, N. T. (2006). Informing teachers live about student learning: Reporting in the assistment system. *TECHNOLOGY INSTRUCTION COGNITION AND LEARNING*, 3(1/2), 63.
- Fogarty, J. A. (2006). *Constructing and evaluating sensor-based statistical models of human interruptibility*. IBM Research. Retrieved from <http://www.cs.cmu.edu/afs/.cs.cmu.edu/Web/People/jfogarty/publications/jfogarty-dissertation-final.pdf>

- Fok, A. W. P., Wong, H. S., & Chen, Y. S. (2005). Hidden Markov model based characterization of content access patterns in an e-learning environment. In *Proceedings* (pp. 201–204). Amsterdam, Netherlands.
- Frezzo, D. C., Behrens, J. T., & Mislevy, R. J. (2009). Design Patterns for Learning and Assessment: Facilitating the Introduction of a Complex Simulation-Based Learning Environment into a Community of Instructors. *Journal of Science Education and Technology*, 19(2), 105–114. doi:10.1007/s10956-009-9192-0
- Garton, L., Haythorntwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1).
- Gaydos, M., & Bauman, E. (2012). Assessing and Evaluating Learning and Teaching Effectiveness: Games. *Game-Based Teaching and Simulation in Nursing and Health Care*.
- Gee, J. P. (2003). *What Video Games Have to Teach Us About Learning And Literacy*. Palgrave Macmillan.
- Gee, J. P. (2005a). Learning by design: Good video games as learning machines. *E-Learning and Digital Media*, 2(1), 5–16.
- Gee, J. P. (2005b). Semiotic social spaces and affinity spaces. *Beyond Communities of Practice Language Power and Social Context*, 214–232.
- Gee, J. P. (2008). Learning and games. *The Ecology of Games: Connecting Youth, Games, and Learning*, 3, 21–40.
- Gee, J. P. (2012). Big “G” Games. <http://www.jamespaulgee.com/node/63>.

- Gee, J. P. (2012, September). Games Can Drive Assessment to a New Place. Retrieved from <http://gamesandimpact.org/wp-content/uploads/2012/09/Games-Can-Drive-Assessment-to-a-New-Place.pdf>
- Gifford, B. R., & Enyedy, N. D. (1999). Activity centered design: Towards a theoretical framework for CSCL. In *Proceedings of the 1999 Conference on Computer Supported Collaborative Learning* (p. 22).
- Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational Researcher*, 26(1), 5–17.
- Greeno, J. G. (2005). Learning in Activity. In R. K. Sawyer (Ed.), *Cambridge Handbook of the Learning Sciences* (pp. 79–96). Cambridge University Press.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3, 1157–1182.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.
- Halterman, J., Sanford, C., & Owen, V. E. (2014). *Interpreting behavioral patterns in game data to identify instances of transgressive play*. Presented at the Games+Learning+Society 10.0, Madison, WI.
- Halverson, R., Berland, M., & Owen, V. E. (in press). Game Based Assessment. In *Encyclopedia of Educational Technology*.
- Halverson, R., Blakesley, C., & Figueiredo-Brown, R. (2011). Video Game Design as a Model for Professional Learning. *Learning to Play: Exploring the Future of Education With Video Games, New Literacies and Digital Epistemologies*, 53, 9–28.

- Halverson, R., & Owen, V. E. (2014). Game Based Assessment: An Integrated Model for Capturing Evidence of Learning in Play. *International Journal of Learning Technology*.
- Hämäläinen, W., Suhonen, J., Sutinen, E., & Toivonen, H. (2004). Data mining in personalizing distance education courses. In *World Conference on Open Learning and Distance Education* (pp. 1–11). Hong Kong.
- Hansen, B. E. (2000). Sample Splitting and Threshold Estimation. *Econometrica*, 68(3), 575–603.
- Harlen, W., & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education*, 4(3), 365–379.
- Heathcote, E., & Prakash, S. (2007). What your learning management system is telling you about supporting your teachers: monitoring system information to improve support for teachers using educational technologies at Queensland University Of Technology. In *International Conference on Information Communication Technologies in Education* (pp. 1–6). Samos Island, Greece.
- Hershkovitz, A., & Nachmias, R. (2008). Developing a Log-based Motivation Measuring Tool (pp. 226–233). Presented at the International Conference on Educational Data Mining.
- Hickey, D., & Jameson, E. (2012). Designing for Participation in Educational Video Games. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 401–430). New York, NY: Springer.
- Retrieved from [http://books.google.com/books?hl=en&lr=&id=Y7sYxaO4W8oC&oi=fnd&pg=PR5&dq=%22second+part+presents+innovative+ways+for+assessing+learning+in%22+%22assessment+of+game-based+learning+\(Loh,+Chap.+8\).+The+Timed+Report%22+%22game-](http://books.google.com/books?hl=en&lr=&id=Y7sYxaO4W8oC&oi=fnd&pg=PR5&dq=%22second+part+presents+innovative+ways+for+assessing+learning+in%22+%22assessment+of+game-based+learning+(Loh,+Chap.+8).+The+Timed+Report%22+%22game-)

based+learning+is+provided+by+the+MAPLET+(Gosper+and%22+&ots=kPthhlyYd5&sig=HxYP3GiZ7HHKdRd4YfevxtO8PMs

- Huang, C., Tsai, P., Hsu, C., & Pan, R. (2006). Exploring cognitive difference in instructional outcomes using text mining technology. In *Proceedings* (pp. 2116–2120). Taipei, Taiwan.
- Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. In *Proceedings of the AAAI Workshop on Challenges in Game AI* (pp. 04–04). Retrieved from <http://www.aaai.org/Papers/Workshops/2004/WS-04-04/WS04-04-001.pdf>
- Ibrahim, Z., & Rusli, D. (2007). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression (pp. 1–6). Presented at the Annual SAS Malaysia Forum, Kuala Lumpur.
- Ingram, A. (1999). Using web server logs in evaluating instructional web sites. *Journal of Educational Technology Systems*, 28(2), 37–157.
- Institute of Play. (2013). *Digging Into Data with SimCityEDU* (Vol. 21). Retrieved from https://www.youtube.com/watch?v=Y5_7Y7wKE6A&feature=share
- International Educational Data Mining Society. (n.d.). Retrieved July 9, 2013, from <http://www.educationaldatamining.org>
- Iseli, M. R., Koenig, A. D., Lee, J., & Wainess, R. (2010). Automatic assessment of complex task performance in games and simulations. In *The Interservice/Industry Training, Simulation & Education Conference 2010* (Vol. 1). National Training Systems Association.
- Jenkins, H. (2006). *Fans, bloggers, and gamers: Exploring participatory culture*. NYU Press.

- Juul, J. (2013). *The Art of Failure: An Essay on the Pain of Playing Video Games*. Cambridge, MA: MIT Press.
- Kapur, M. (2006). Productive failure. In S. Barab, K. Hay, & D. Hickey (Eds.), *Proceedings of the International Conference on the Learning Sciences* (Vol. 0, pp. 307–313).
- Kelly, D., & Tangney, B. (2005). First Aid for You: Getting to Know Your Learning Style Using Machine Learning. In *IEEE international Conference on Advanced Learning Technologies* (pp. 1–3). Washington, DC.
- Kerr, D., & Chung, G. K. W. K. (2013). *The Effect of In-Game Errors on Learning Outcomes* (CRESST Report No. 835). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from <https://www.cse.ucla.edu/products/reports/R835.pdf>
- Koedinger, K. R., Pavlik, P., McLaren, B. M., & Aleven, V. (2008). Is it better to give than to receive? The assistance dilemma as a fundamental unsolved problem in the cognitive science of learning and instruction. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2155–2160). Retrieved from <http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/bmclaren/pubs/KoedingerEtAl-IsItBetterToGiveThanToReceive-CogSci2008.pdf>
- Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press. Retrieved from http://books.google.com/books?hl=en&lr=&id=CAVIOrW3vYAC&oi=fnd&pg=PA11&ots=OAsDys1EGm&sig=5vznah0Pt-e5jo_5NjQIpUv0LX4#v=onepage&q&f=false

- Lee, H. K., & Kim, J. H. (1999). An HMM-based threshold model approach for gesture recognition. In *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (Vol. 21(10), pp. 961–973).
- Lee, M. W., Chen, S. Y., Chrysostomou, K., & Liu, X. (2009). Mining student's behavior in web-based learning programs. *Expert Systems with Applications Journal*, 36, 3459–3464.
- Lee, M. W., Chen, S. Y., & Liu, X. (2007). Mining learners' behavior in accessing web-based interface (pp. 226–346). Presented at the International Conference of Edutainment, Hong Kong, China.
- Loh, C. S. (2012). Information Trails: In-process assessment of game-based learning. In *Assessment in Game-Based Learning* (pp. 123–144). New York: Springer.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge, MA: MIT Press.
- Martinez, R., Yacef, K., Kay, J., Al-Qaraghuli, A., & Kharrufa, A. (2011). Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In *Fourth International Conference on Educational Data Mining, Eindhoven*. Retrieved from http://educationaldatamining.org/EDM2011/wp-content/uploads/proc/edm2011_paper36_full_Martinez.pdf
- Mavrikis, M. (2008). Data-Driven Prediction of the Necessity of Help Requests in ILEs. In *International Conference on Adaptive Hypermedia* (pp. 316–319). Hanover, Germany.
- Mavrikis, M. (2010). Modelling student interactions in intelligent learning environments: constructing Bayesian networks from data. *International Journal on Artificial Intelligence Tools*, 19(06), 733–753.

- Mayrath, M. C., Clarke-Midura, J., & Robinson, D. H. (Eds.). (2012). *Technology-based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Charlotte, NC: Information Age Publishing.
- McQuiggan, S., Mott, B., & Lester, J. (2008). Modeling Self-Efficacy in Intelligent Tutoring Systems: An Inductive Approach. *User Modeling and User-Adapted Interaction*, 18, 81–123.
- Merceron, A., & Yacef, K. (2011). Measuring correlation of strong symmetric association rules in educational data. In C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker (Eds.), *Handbook of educational data mining* (pp. 245–256). US: Taylor & Francis.
- Millán, E., Loboda, T., & Pérez-de-la-Cruz, J. L. (2010). Bayesian networks for student model engineering. *Computers & Education*, 55(4), 1663–1683.
doi:10.1016/j.compedu.2010.07.010
- Mislevy, R. J. (2011). *Evidence-Centered Design for Simulation-Based Assessment* (No. 800). National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *RESEARCH REPORT-EDUCATIONAL TESTING SERVICE PRINCETON RR, 16*. Retrieved from <http://marces.org/EDMS623/Mislevy%20on%20ECD.pdf>
- Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three Things Game Designers Need to Know About Assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 59–81). New York, NY: Springer.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of Evidence Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.

- Mislevy, R. J., & Riconscente, M. M. (2005). Evidence-centered assessment design: Layers, structures, and terminology. *Menlo Park, CA: SRI International*. Retrieved from http://padi.sri.com/downloads/TR9_ECD.pdf
- Nardi, B. A. (Ed.). (1996). *Context and consciousness: Activity theory and human computer interaction*. Cambridge, MA: MIT Press.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- Nkambou, R., Nguifo, E. M., Couturier, O., & Fournier-Viger, P. (2007). Problem-solving knowledge mining from users' actions in an intelligent tutoring system. *Advances in Artificial Intelligence*, 393–404.
- Norton, D. (2008a). A Practical Model for Separating Games and Simulations. Presented at the 5th Annual Games+Learning+Society Conference. Retrieved from <http://www.previous.glsconference.org/2008/session.html?id=139>
- Norton, D. (2008b, July 10). *A Practical Model for Separating Games and Simulations*. Presented at the Games+Learning+Society Conference, Madison, Wisconsin. Retrieved from <http://www.previous.glsconference.org/2008/session.html?id=139>
- Owen, V. E. (2014). Learning Analytics for Educational Game Design: Mapping Methods to Development. Presented at the The 11th Annual Games+Learning+Society Conference, Madison, WI.
- Owen, V. E., & Halverson, R. (2013). ADAGE: Assessment Data Aggregator for Game Environments. In *Williams, C., Ochsner, A., Dietmeier, J., & Steinkuehler, C. (Eds.) Proceedings of the Games, Learning, and Society Conference (Vol. 3)*. Pittsburgh, PA: ETC Press.

- Owen, V. E., & Ramirez, D. (in submission). Analytics in Learning Games: Mining Educational Gameworlds using ADAGE (Assessment Data Aggregator for Game Environments). *Journal of Computer Assisted Learning*, (Special Issue: Learning Analytics in Massively Multiuser Virtual Environments and Courses).
- Owen, V. E., Ramirez, D., Salmon, A., & Halverson, R. (2014a). *Game-Based Learning Analytics with ADAGE (Assessment Data Aggregator for Game Environments)*. Presented at the Games+Learning+Society Conference.
- Owen, V. E., Ramirez, D., Salmon, A., & Halverson, R. (2014b, April). *ADAGE (Assessment Data Aggregator for Game Environments): A Click-Stream Data Framework for Assessment of Learning in Play*. Presented at the American Educational Research Association.
- Owen, V. E., Shapiro, R. B., & Halverson, R. (2013). Gameplay as Assessment: Analyzing Event-Stream Player Data and Learning Using GBA (a Game-Based Assessment Model). In *CSCLE 2013 Conference Proceedings* (Vol. Volume 1 — Full Papers & Symposia, pp. 360–367). Madison, WI: International Society of the Learning Sciences (ISLS).
- Owen, V. E., Wills, N., & Halverson, R. (2012). CyberSTEM: A Game-Based Evidence Model. In A. Ochsner (Ed.), *Proceedings of the 8th Annual Games, Learning, and Society Conference*. Pittsburgh, PA: ETC Press.
- Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. New York: Basic Books, Inc. Retrieved from http://www.medientheorie.com/doc/papert_mindstorms.pdf
- Plass, J. L., Homer, B. D., Kinzer, C. K., & Perlin, K. (2012). Games for Learning Institute (G4LI) White Paper: Ideas for Impact Games. Retrieved from

<http://gamesandimpact.org/wp-content/uploads/2012/09/PlassNYU-Ideas-for-Impact-Games-2.pdf>

- President's Council of Advisors on Science and Technology. (2013). *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*. Executive Office of the President of the United States.
- Pritchard, D., & Warnakulasooriya, R. (2005). Data from a Web-based Homework Tutor can predict Student's Final Exam Score (pp. 2523–2529). Presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Chesapeake.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE* (Vol. 77(2), pp. 257–286).
- Ramirez, D., Hatfield, D., Owen, V. E., Samson-Samuel, C., & Shapiro, R. B. (2012, October). *Meaningful failure: Instrumental feedback that guides performance*. Presented at the Meaningful Play Conference, East Lansing, MI.
- Retalis, S., Papasalouros, A., Psaromiligkos, Y., Siscos, S., & Kargidis, T. (2006). Towards networked learning analytics—A concept and a tool. In *Proceedings of the fifth international conference on networked learning*. Retrieved from <http://nlc.ell.aau.dk/past/nlc2006/abstracts/pdfs/P41%20Retalis.pdf>
- Rieber, L. P. (1992). Computer-based microworlds: A bridge between constructivism and direct instruction. *Educational Technology Research & Development*, 40(1), 93–106.
- Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development*, 44(2), 43–58.

- Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. T. (2007). Cognitive Tutor: Applied research in mathematics education. *14*(2), 249-255. *Psychonomic Bulletin & Review*, *14*(2), 249–255.
- Rodrigo, M., Mercedes, T., d Baker, R. S., McLaren, B. M., Jayme, A., & Dy, T. T. (2012). Development of a Workbench to Address the Educational Data Mining Bottleneck. *International Educational Data Mining Society*. Retrieved from <http://files.eric.ed.gov/fulltext/ED537222.pdf>
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. doi:10.1016/j.eswa.2006.04.005
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *40*(6), 601–618.
- Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368–384.
- Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. (Eds.). (2011). *Handbook of educational data mining*. US: Taylor & Francis.
- Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. MIT Press.
- San Pedro, M. O. C. Z., d Baker, R. S., & Rodrigo, M. M. T. (2011). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *Artificial Intelligence in Education* (pp. 304–311). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-21869-9_40
- Sasikala, D., Premalatha, K., & Logeswari, S. (2011). A Survey on Association Rule Mining. *International Journal for Data Mining and Knowledge Engineering*, *3*(4), 231–236.

- Scalise, K., & Wilson, M. (2012). Measurement Principles for Gaming. In *Assessment in Game-Based Learning* (pp. 287–305). New York: Springer.
- Schell, J. (2008). *The art of game design: A book of lenses*. US: Taylor & Francis.
- Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., ... Mislevy, R. (2009). Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media*, 1(2), 33–53.
doi:10.1162/ijlm.2009.0013
- Shank, R. (2011). *Teaching Minds: How Cognitive Science Can Save Our Schools*. New York, NY: Teachers College Press.
- Sheard, J., Ceddia, J., Hurst, J., & Tuovinen, J. (2003). Inferring student learning behaviour from website interactions: A usage analysis. *Education and Information Technologies*, 8(3), 245–266.
- Shelton, B. E., & Parlin, M. A. (2012). Taking Activity-Goal Alignment into Open-Ended Environments: Assessment and Automation in Game-Based Learning. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 105–121). Springer.
- Shih, B., Koedinger, K. R., & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. In C. Romero, S. Ventura, M. Pechenizkiy, & R. Baker (Eds.), *Handbook of educational data mining* (pp. 201–212). US: Taylor & Francis.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 55(2), 503–524.

- Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: foundations, innovations, and perspectives* (pp. 43–58). New York, NY: Springer.
- Shute, V. J., & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning? In *Design research on learning and thinking in educational settings* (pp. 359–387). Routledge.
- Simmons, M. P., Adamic, L. A., & Adar, E. (2011). Memes Online: Extracted, Subtracted, Injected, and Recollected. In *ICWSM*. Retrieved from <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2836/3281>
- Song, D., Lin, H., & Yang, Z. (2007). Opinion Mining in e-Learning Systems. Presented at the International Conference on Network and Parallel Computing.
- Songer, N. B. (2012). Assessing essential science of nascent inquirers. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.
- Southavilay, V., Yacef, K., & Calvo, R. A. (2010). Analysis of Collaborative Writing Processes Using Hidden Markov Models and Semantic Heuristics. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on* (pp. 543–548). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5693344
- Squire, K. (2006). From content to context: Videogames as designed experience. *Educational Researcher*, 35(8), 19–29.
- Squire, K. (2011). *Video Games and Learning: Teaching and Participatory Culture in the Digital Age*. Teachers College Press. Retrieved from <http://eric.ed.gov/?id=ED523599>

- Squire, K. (2012, October). *Participatory Assessment*. Presented at the ASU Emerging Technologies Event, Tempe, AZ. Retrieved from <http://gamesandimpact.org/video/participatory-assessment/>
- Srivastava, J. (2008). Data mining for social network analysis. In *Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on* (pp. xxxiii–xxxiv). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4565015
- Steinkuehler, C. A. (2004). Learning in massively multiplayer online games. In *Proceedings of the 6th international conference of learning sciences* (pp. 521–528).
- Steinkuehler, C. A. (2006). Massively multiplayer online video gaming as participation in a discourse. *Mind, Culture, and Activity, 13*(1), 38–52.
- Steinkuehler, C. A. (2006). Why Game (Culture) Studies Now? *Games and Culture, 1*(1), 97–102. doi:10.1177/1555412005281911
- Steinkuehler, C., Barab, S., & Squire, K. (Eds.). (2012). *Games, Learning, and Society: Learning and Meaning in the Digital Age*. New York: Cambridge University Press.
- Steinkuehler, C., & Duncan, S. (2008). Scientific Habits of Mind in Virtual Worlds. *Journal of Science Education and Technology, 17*(6), 530–543. doi:10.1007/s10956-008-9120-8
- Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(3), 479–498.
- Sweet, S. J., & Rupp, A. A. (2012). Using the ECD Framework to Support Evidentiary Reasoning in the Context of a Simulation Study for Detecting Learner Differences in Epistemic Games. *Journal of Educational Data Mining, 4*(1).
- Tufte, E. R., & Graves-Morris, P. R. (1983). *The visual display of quantitative information* (Vol. 2). Cheshire, CT: Graphics press.

- U.S. Department of Education. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief*. Washington, DC: SRI International.
- Ueno, M., & Nagaoka, K. (2002). Learning log database and data mining system for e-learning – online statistical outlier detection of irregular learning processes. In *International Conference on Advanced Learning Technologies* (pp. 436–438). Tatarstan, Russia.
- Vee, M. N., Meyer, B., & Mannock, K. L. (2006). Understanding novice errors and error paths in object-oriented programming through log analysis. In *Workshop on Educational Data Mining* (pp. 13–20). Taiwan.
- Ventura, M., Shute, V. J., & Kim, Y. J. (2013). Assessment and Learning of Qualitative Physics in Newton’s Playground (pp. 579–582). Presented at the Artificial Intelligence in Education, Springer Berlin Heidelberg.
- Vranic, M., Pintar, D., & Skocir, Z. (2007). The use of data mining in education environments (pp. 243–250). Presented at the International Conference on Telecommunications, Zagred.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Walonoski, J. A., & Heffernan, N. T. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 382–391).
- Wang, A. Y., & Newlin, M. H. (2002). Predictors of we-based performance: the role of self-efficacy and reasons for taking an on-line class. *Computers in Human Behavior Journal*, 18, 151–163.

- Wang, Y.-C., & Witten, I. H. (1997). Inducing model trees for continuous classes. In *Proceedings of the Ninth European Conference on Machine Learning* (pp. 128–137). Retrieved from <http://www.cs.waikato.ac.nz/~ml/publications/1997/Wang-Witten-Induct.pdf>
- Wiggins, G. (1990). The case for authentic assessment.
- Wilensky, U. (1999). *NetLogo*. Evanston, IL: Northwestern University. Retrieved from <http://ccl.northwestern.edu/netlogo>
- Wixon, M., d Baker, R. S., Gobert, J. D., Ocumpaugh, J., & Bachmann, M. (2012). WTF? detecting students who are conducting inquiry without thinking fastidiously. In *User Modeling, Adaptation, and Personalization* (pp. 286–296). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-31454-4_24
- Woolf, B. P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D. G., Dolan, R., & Christopherson, R. M. (2010). The effect of motivational learning companions on low achieving students and students with disabilities. In *Intelligent Tutoring Systems* (pp. 327–337). Springer Berlin Heidelberg.
- Wu, A., & Leung, C. (2002). Evaluating learning behavior of Web-Based Training (WBT) using web log (pp. 736– 737). Presented at the International Conference on Computers in Education, New Zealand.
- Xu, B., & Recker, M. (2012). Teaching Analytics: A Clustering and Triangulation Study of Digital Library User Data. *Educational Technology & Society*, 15(3), 103–115.
- Yu, C. H., Digangi, S., Jannasch-Pennell, A. K., & Kaprolet, C. (2008). Profiling students who take online courses using data mining methods. *Online Journal of Distance Learning Administration*, XI(II), 1–14.

- Yu, C. H., Jannasch-Pennell, A. K., Digangi, S., & Wasson, B. (1999). Using on-line interactive statistics for evaluating web-based instruction. *Journal of Educational Media International*, 35, 157–161.
- Zaïane, O., & Luo, J. (2001). Web usage mining for a better web-based learning environment. In *Proceedings of Conference on Advanced Technology for Education*. (pp. 60–64). Banff, Alberta.
- Zapata-Rivera, D. (2009). Assessment in Game-Based Environments. Presented at the Annual Meeting of the American Educational Research Association.
- Zapata-Rivera, D., & Bauer, M. (2012). Exploring the role of games in educational assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. (pp. 146–169). Charlotte, NC: Information Age Publishing.
- Zapata-Rivera, D., & Hansen, E. G. (2009). Analyzing the educational potential of existing games using Evidence-Centered Design and Cognitive Task Analysis. Presented at the Annual Meeting of the American Educational Research Association.
- Zhang, C., & Zhang, S. (2002). *Association rule mining: models and algorithms (lecture notes)*. Presented at the Artificial Intelligence.
- Zhang, X., Mostow, J., Duke, N., Trotochaud, C., Valeri, J., & Corbett, A. T. (2008). Mining Free-form Spoken Responses to Tutor Prompts. In *International conference on Educational Data Mining* (pp. 234–241). Montreal.
- Zhao, J. Y. (2010). *Hidden Markov Models with Multiple Observation Processes*. Unpublished master's thesis, Department of Mathematics and Statistics, University of Melbourne. Retrieved from <http://arxiv.org/abs/1010.1042>

Zheng, S., Xiong, S., Huang, Y., & Wu, S. (2008). Using methods of association rules mining optimization in web-based mobile learning system. In *Proceedings* (pp. 967–970).

Zimmerman, B. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology, 81*, 329–339.

Appendix D.

(Chapter Five) Basic Markov: transition matrix – upper quartile model

Row Labels	end	fail	obj1_cell	obj2_cell	obj3_tissue	obj4_tissue	obj5_cell	obj5_tissue	obj6_organ	obj7_tissue	obj8_cell	obj8_organ	obj8_tissue	success
fail		1												
obj1_cell				0.23										
obj2_cell		0.01			0.16									
obj3_tissue						0.96								
obj4_tissue		0.01					0.38							
obj5_cell		0.02						0.18						
obj5_tissue		0.1							0.44					
obj6_organ		0.09								0.9				
obj7_tissue		0.02									0.45			
obj8_cell		0.01											0.24	
obj8_organ											1			
obj8_tissue														0.51
start			1											
success		1												

Note: transition matrices from Chapter Five's most detailed set of models were too big to fit legibly here; I'm happy to supply them as excel docs via email upon request.

Appendix E. Progenitor Protocol: Pre-Survey

~Progenitor X Survey~

*Thank you for being a PlaySquad team member!
Your honest answers below really help us improve the game.*

1. Please enter your PlaySquad ID Number in the box below. This is the number that was given to you on the index card.

Part I: About You

2. What is your gender?

a. Male
b. Female

3. How old are you in years? _____

4. What race/ethnicity do you consider yourself?

a. Black not Hispanic
b. Hispanic or Latino/a
c. White not Hispanic
d. Asian or Pacific Islander
e. Native American or Alaskan Native
f. I prefer not to answer
g. Other (please specify) _____

5. On average, what kind of grades do you earn?

a. A's
b. A's and B's
c. B's
d. B's and C's
e. C's
f. C's and D's
g. D's and below
h. I prefer not to answer

Part II: Your Play Preferences

6. At what pace do you like to move through games? (circle a number below)

More Slowly	1	2	3	4	5	6	7	8	9	10	More Quibbly
-------------	---	---	---	---	---	---	---	---	---	----	--------------

7. How do you feel about playing videogames? (circle a number below)

Strongly Dislike	1	2	3	4	5	Strongly Like
Dislike			Neutral	Like		

8. How important are the following parts of gameplay to you?

	Not Important	A Little Important	Somewhat Important	Pretty Important	Very Important	N/A (I don't play, or my games don't have this)
a. Good storyline and characters	0	0	0	0	0	0
b. Learning how to solve the game challenges	0	0	0	0	0	0
c. Experiencing cool graphics and interacting with the game visuals	0	0	0	0	0	0
d. Good music and sound effects	0	0	0	0	0	0
e. Customizing your character's appearance	0	0	0	0	0	0
f. Customizing playstyle (combat & skill trees choice of gear/tools etc)	0	0	0	0	0	0
g. An open, explorable gameworld	0	0	0	0	0	0
h. Many different storyline choices your character can make	0	0	0	0	0	0
i. Teamwork with others	0	0	0	0	0	0

9. If there's an important part of your gameplay that isn't listed, write it below.

10. What difficulty level do you prefer when you play games? (circle a number below)

Easier	1	2	3	4	5	6	7	8	9	10	Harder
--------	---	---	---	---	---	---	---	---	---	----	--------

11. Do you consider yourself a gamer?

a. Yes
b. Somewhat
c. No

Prev5

Page 1

Prev5

Page 2

12. What is important in motivating you throughout gameplay?

	Not Important	A Little Important	Somewhat Important	Pretty Important	Very Important	N/A: (I don't play, or my games don't have this)
a. Understanding many details of the game	0	0	0	0	0	0
b. Achievements or badges	0	0	0	0	0	0
c. Exploring all the areas of the game	0	0	0	0	0	0
d. Winning the endgame	0	0	0	0	0	0
e. Finding the smartest way to use game tools	0	0	0	0	0	0
f. Competition vs. others or self (like beating your own high scores or others' scores)	0	0	0	0	0	0
g. Completing all game parts (even optional ones)	0	0	0	0	0	0
h. Collecting items in-game (gems, pets, rare objects etc.)	0	0	0	0	0	0

13. If something motivates you in gameplay that isn't listed, write it below.

14. Please choose the response that best describes your viewpoint.

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
a. Playing videogames hurts your schoolwork.	0	0	0	0	0
b. Playing videogames is a social activity.	0	0	0	0	0
c. Playing videogames is a waste of time.	0	0	0	0	0
d. Playing videogames can help you learn.	0	0	0	0	0

Part III: Science

Please circle your answers below. Don't worry if you don't know the answer - just pick the one you think is best.

15. Where can fibroblasts come from?

- a. They can come from adult humans.
- b. They come from large rocks.
- c. They can be artificially created in a lab.
- d. They can come from plants.

16. Where can IPS cells (stem cells) come from?

- a. They can be made by mixing chemical compounds.
- b. They can grow from the ground.
- c. They can come from treated fibroblasts.
- d. They can be made from plastic.

17. What are two steps used in growing IPS cells?

- a. Eliminating cells with bacteria, then collecting them with a micropipette
- b. Shocking fibroblasts with electroporation, then exposing them to the zombie virus
- c. Collecting cells with a micropipette, and then evaporating them
- d. Applying a growth factor to fibroblasts then collecting them with a micropipette

18. What does the growth factor do to IPS cells?

- a. Destroys all of them.
- b. Turns them into special cells that can be used to make human tissue.
- c. Converts them into chemicals for creating robot parts.
- d. Turns them into plant cells.

19. What kind of cells can IPS cells turn into?

- a. Roboderm, endoderm, and mesoderm
- b. Mesoderm, ectoderm, and plasmaderm
- c. Roboderm, mesoderm, and ectoderm
- d. Endoderm, mesoderm, and ectoderm

20. The largest building blocks that make up human TISSUE are _____.

- a. Plasma proteins
- b. Organs
- c. Cells
- d. DNA

21. Which of the following is formed from groups of tissues?

- a. Vessels
- b. Organs
- c. Genes
- d. Cells

Appendix F. Progenitor Protocol: Post-Survey

~Progenitor X Survey~

1. Please enter your PlaySquad ID Number in the box below.

Part I: Science

2. Where can fibroblasts come from?

- They can come from adult humans.
- They come from large rocks.
- They can be artificially created in a lab.
- They can come from plants.

3. Where can iPS cells (stem cells) come from?

- They can be made by mixing chemical compounds.
- They can grow from the ground.
- They can come from treated fibroblasts.
- They can be made from plastic.

4. What are two steps used in growing iPS cells?

- Eliminating cells with bacteria, then collecting them with a micropipette
- Shocking fibroblasts with electroporation, then exposing them to the zombie virus
- Collecting cells with a micropipette, and then evaporating them
- Applying a growth factor to fibroblasts, then collecting them with a micropipette

5. What does the growth factor do to iPS cells?

- Destroys all of them.
- Turns them into special cells that can be used to make human tissue.
- Converts them into chemicals for creating robot parts.
- Turns them into plant cells.

6. What kind of cells can iPS cells turn into?

- Roboderm, endoderm, and mesoderm
- Mesoderm, ectoderm, and ~~plasmaderm~~
- Roboderm, mesoderm, and ectoderm
- Endoderm, mesoderm, and ectoderm

7. The largest building blocks that make up human TISSUE are _____.

- Plasma proteins
- Organs
- Cells
- DNA

Post#15

8. Which of the following is formed from groups of tissues?

- Vessels
- Organs
- Genes
- Cells

Part II: Gameplay Habits

9. Circle all game types below that you play regularly.

- First Person Shooters (eg Call of Duty, Halo, Deus Ex, Dishonored, Battlefield 3)
- Fighting (eg Street Fighter, Mortal Kombat, Tekken, Marvel vs. Capcom)
- Music (eg Guitar Hero, Rock Band, Sing Star)
- Party (eg Mario Party, Fusion Frenzy, Tetris Party)
- Arcade (eg Super Mario Bros., Donkey Kong, Crash Bandicoot, Qbert, etc.)
- Classic Adventure (eg Myst, The Longest Journey, Siberia, Sherlock Holmes, Nancy Drew Adventure series, etc)
- Casual (eg Angry Birds, hidden-object games, Bejeweled)
- Word / Quiz Games (eg Electronic Scrabble, Jeopardy, Words with Friends, etc)
- Puzzle (eg Tetris, Foldit, Portal, CoGS)
- Racing (eg iRacing, Mario Kart, Grand Turismo, Need for Speed)
- Simulation (Nintendogs, SimCity, Viva Piñata, the Sims)
- Sports (eg Madden NFL, Tony Hawk's Pro Skater, Tiger Wood's PGA Tour, EA Sports Series)
- Strategy (eg Starcraft, Civilization, Company of Heroes, Age of Empires, Dawn of War)
- Survival Horror (eg Dead Space, Left for Dead, Resident Evil, Silent Hill)
- Action/Adventure (eg Grand Theft Auto, Uncharted, God of War, Batman Arkham Asylum)
- Fitness (eg Wii Fit, Kinect Sports, Zumba Fitness, NFL Training Camp)
- Indie Games (eg Minecraft, Limbo, Braid, Octodad, Cave Story, Super Meat Boy)
- Massively Multiplayer Online Games (eg World of Warcraft, Everquest, Runescape)
- Role Playing Games (eg Diablo, Legend of Zelda Elder Scrolls, Fokentop)
- Non-Electronic (card games, board games)
- Other (please list): _____

10. How many hours a week do you usually play video games?

- I don't play video games
- 0-3
- 4-6
- 7-9
- 10-13
- 14+

11. Circle all platforms below that you play on regularly.

- Portable consoles (Nintendo 3DS, PS Vita)
- Home consoles (Nintendo, Playstation 3, Xbox 360)
- Motion-tracking home consoles (Nintendo Wii, Xbox Kinect)
- Mobile phones (Android, iPhone, or iPod touch)

Post#15

Page 1

Part IV: Feelings About Games/Technology

18. How well do the following statements describe you? I play video games because...

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree	I don't play video games
a. They are challenging	0	0	0	0	0	0
b. They are exciting	0	0	0	0	0	0
c. They are relaxing	0	0	0	0	0	0
d. I can play them with my friends.	0	0	0	0	0	0
e. I can't do those things in real life.	0	0	0	0	0	0
f. I want to escape from problems in real life.	0	0	0	0	0	0

19. How easy or difficult are the following things for you?

	Very easy	Somewhat easy	Neither easy nor hard	Somewhat hard	Very hard	N/A (I don't know what this is)
a. Learning new computer programs	0	0	0	0	0	0
b. Finding information or media online	0	0	0	0	0	0
c. Using computer programming languages (Java, C++, Python, etc.)	0	0	0	0	0	0

- e. Tablets (iPad)
- f. Computer or laptop
- g. Board games, card games, or other physical games
- h. I don't like playing any of these
- i. Other (please specify): _____

12. Do you play videogames often with other people, either in the same room or online?
 a. Yes
 b. No

Part III: Today's Play Experience

13. Before today, had you ever played Progenitor X before?
 a. No
 b. Yes

14. Overall, how difficult was Progenitor X? (circle a number below)

Not difficult at all 1 2 3 4 5 6 7 8 9 10
 Extremely difficult

15. Overall, how fun was Progenitor X? (circle a number below)

Not fun at all 1 2 3 4 5 6 7 8 9 10
 Extremely fun

16. Overall, how frustrating was Progenitor X? (circle a number below)

Not frustrating at all 1 2 3 4 5 6 7 8 9 10
 Extremely frustrating

17. While playing, I read... (please be honest so we can improve the game!)

- a. most of the directions
- b. some of the directions
- c. very few directions

Part V: Reflections

20. What did you enjoy most about *Progenitor*?

21. If you could change anything about *Progenitor*, what would it be?

22. What do you think you learned about stem cells?

23. Why might stem cells be important?
