International Journal of Gaming and Computer-Mediated Simulations
Volume 9 • Issue 3 • July-September 2017

Investigating Epistemic Stances in Game Play with Data Mining

Mario M. Martinez-Garza, Department of Teaching and Learning, Vanderbilt University, Nashville, TN, USA
Douglas B. Clark, University of Calgary, Calgary, AB, Canada

ABSTRACT

In this paper, techniques of statistical computing were applied to data logs to investigate the patterns in students' play of *The Fuzzy Chronicles*, and how these patterns relate to learning outcomes with regards to Newtonian kinematics. This paper has two goals. The first goal is to investigate the basic claims of the proposed Two-System Framework for Game-Based Learning (or 2SM) (Martinez-Garza & Clark, 2016) that may serve as part of a general-use explanatory framework for educational gaming. The second goal is to explore and demonstrate the use of automatically collected log files of student play as evidence through educational data mining techniques. These techniques could also find general use, and this paper offers a demonstration of plausible methods and processes that are suited for game play data. These goals were pursued via two research questions. The first research question examines whether students playing The Fuzzy Chronicles showed evidence of dichotomous fast/slow modes of solution. The 2SM theorizes that slow modes of solution will correlate to higher learning gains. Congruent with the 2SM, students who use mainly fast iterative solution strategies achieved lower learning gains than students who preferred slow, elaborate solutions, or a more balanced mix of the two. A second research question investigates the connection between conceptual understanding and student performance in conceptually-laden challenges. The finding was that students generally improve their performance in these challenges as gameplay progresses, but that this improvement is strongly moderated by their prior knowledge of physics. Implications of these findings in terms of educational game design, analysis of gameplay logs, and further refinement of the 2SM are discussed.

KEYWORDS

Data Mining, Games for Learning, Science Education, Student Modelling

INTRODUCTION

Digital games are potentially powerful vehicles for learning (Gee, 2007; Prensky, 2006; Mayo, 2009; Shaffer, Squire, Halverson, & Gee, 2005; Rieber, 1996; Squire et al., 2003), and numerous empirical studies have linked classroom use of educational games to increased learning outcomes in science (e.g., Annetta, Minogue, Holmes, & Cheng, 2009; Dieterle, 2009; Neulight, Kafai, Kao, Foley, & Galas, 2007; Squire, Barnett, Grant, & Higginbotham, 2004). Several reviews have concluded that game-based learning offers numerous theoretical and practical affordances that can help foster students' conceptual understanding, engagement, and self-efficacy (Aldrich, 2003; Cassell & Jenkins, 1998; Kafai, Heeter, Denner, & Sun, 2008; Kirriemuir & Mcfarlane, 2004; Martinez-Garza, Clark, &

DOI: 10.4018/IJGCMS.2017070101

Nelson, 2012, Munz, Schumm, Wiesebrock, & Allgower, 2007). That said, not all games effectively support learning for all learners (Young et al., 2012). Clark, Tanner-Smith, and Killingsworth (2015) find favorable support for the use of educational games overall, but particularly in cases where games are augmented through the application of sound learning theory.

While the general question of whether games can provide productive contexts for learning is approaching consensus, *how* and *why* and *when* games work are more open questions. A large number of constructs receive attention as potentially important for game-based learning (Linehan, Kirman, Lawson, & Chan, 2011; Dondlinger, 2007), including constructs as varied as fun, feedback, engagement, flow, problem-solving, narrative, etc. Several scholars have proposed design principles to optimally leverage some or all of these constructs (e.g. Annetta, 2010; Kelle, Klempke, & Specht, 2011; Tobias & Fletcher, 2007; Plass, Homer, & Kinzer, 2014). Also, educational games claim a broad spectrum of possible learning outcomes (Martinez-Garza, Clark, & Nelson, 2013b) which, when combined with the wide range of gaming genres, gaming populations, and technology platforms educational researchers have available, creates a vast and constantly changing space of inquiry that resists generalized claims. Furthermore, digital games also present unique assessment challenges. Since games often incorporate novel student activities for which there are no well-established existing measurement methods, measures often need to be developed along with the game in an iterative fashion (Harpstead, Myers, & Aleven, 2013). Thus, some scholars have called for increased methodological rigor and emphasis on usable (i.e. generalizable) knowledge in educational games research (Dede, 2011; Foster & Mishra, 2008).

Regardless of the variations in theoretical framing, methods, or learning outcomes, the common denominator of all game-based learning research is the act of students' play. Thus, a general claim of game-based learning research can be phrased as "if a student plays this particular game, he or she will learn this particular thing." Much inquiry into game-based learning is directed towards explicating other issues that influence and structure educational gaming (e.g. design considerations, materials and curricula to support educational games, and detection of learning outcomes), although not so much play itself, i.e. what choices the student has available, what informs those choices, and what feedback the game offers in response. Generally speaking, the act of play as the central driver of learning is somewhat under-examined in the educational gaming literature. Among the possible reasons for this lack of focus are (1) the general difficulty of observing, encoding, and analyzing play systematically, and (2) the limitations of general theoretical frameworks that might help operationalize play in meaningful actionable ways.

Previous educational research efforts that analyzed digital game play at the individual level have relied primarily on observational methods (e.g., Annetta, Minogue, Holmes, & Chang, 2009; Hou, 2012; Sengupta, Krinks, & Clark, 2015). Observational studies that aim for thick description (Geertz, 1973) of gamers at play explicate this richness and often succeed in building strong cases for learning (e.g. Squire, DeVane, & Durga, 2008). However, investigations of play that use a student's *in situ* performance as an indicator of the learning are generally limited in scope and scale by the costs and demands of observation and coding. A possible way to address this limitation involves the use of log file data. Students' actions within the game environment, when recorded and compiled, can potentially produce a rich and detailed account that can be productively analyzed using methods of statistical computing (Martinez-Garza, Clark, & Nelson, 2012). These statistical computing methods, variously known as *learning analytics* (LA), or *educational data mining* (EDM), could be used not only for assessment of learning (as we proposed in Clark, Martinez-Garza, Biswas, Luecht & Sengupta, 2012) but also to find underlying structure and regularity in students' play that may inform meaningful generalizations about what constitutes learning through play in a game environment. Using a combination of log file data and learning analytics, educational games scholarship could potentially transcend the limitation of cost, time, and human effort without abandoning deep qualitative analysis (Berland, Baker, & Blikstein, 2014).

GOAL AND STRUCTURE OF THIS PAPER

This paper has two goals. The first goal is to investigate the basic claims of the proposed Two-System Framework of Game-Based Learning (Martinez-Garza & Clark, 2016), a cognitive perspective that may serve as part of a general-use explanatory framework for educational gaming. The second goal is to explore and demonstrate the use of automatically collected log files of student play as evidenced by educational data mining techniques. These techniques have drawn interest from researchers seeking a more nuanced understanding of student action within digital environments. The data mining techniques featured in this paper could potentially find general use, and this paper aims at offering a demonstration of plausible methods and processes that are suited for the specific challenges of game play data.

The context for this research is an educational game intended to help middle school students develop a better understanding of Newtonian kinematics. Among its other functionalities, this particular game stores all student actions and collects them in a central database. The Conceptual Framework section describes this game, titled *The Fuzzy Chronicles,* in some detail. Then, a summary of the Two-System Framework (or 2SM) is presented, followed by specific discussion of the implications of the 2SM in the context of The Fuzzy Chronicles. A brief overview of current research that makes use of log files from digital educational environments as evidence rounds out the Conceptual Framework section.

Plan of Work

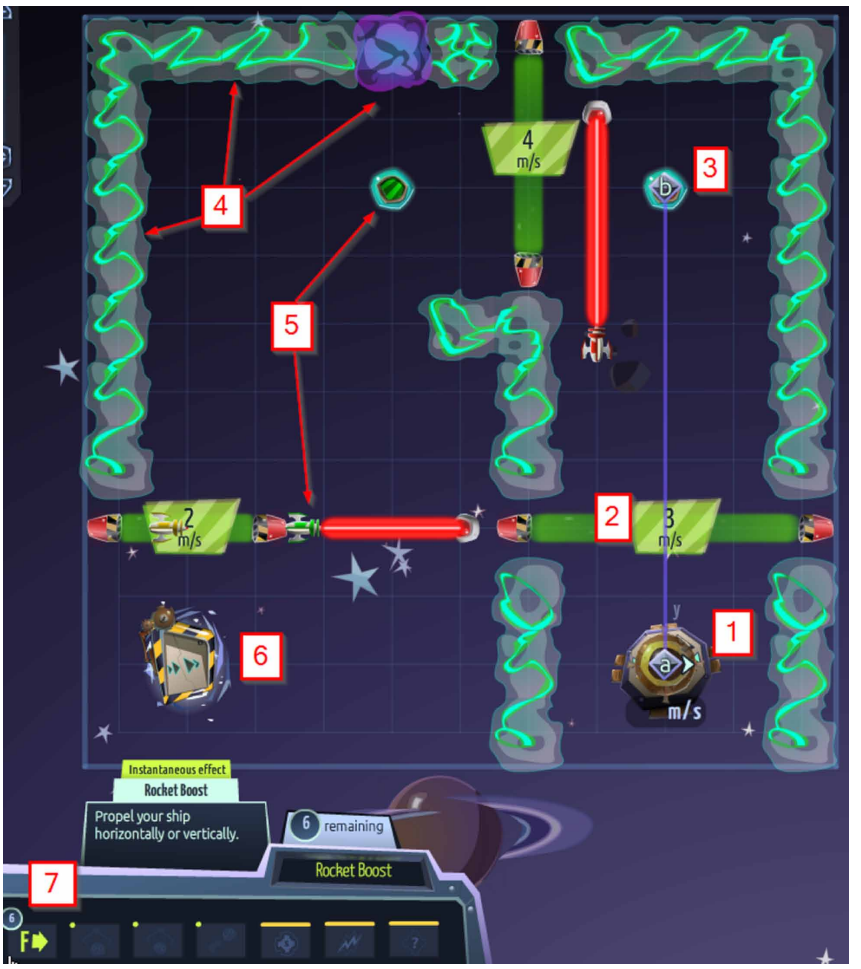
After laying out the necessary groundwork, we articulate goal of investigating the central claims of the 2SM more specifically as two research questions. Research Question 1 (RQ1) asks, "can the two epistemic stances theorized in the 2SM be observed through the study of log files gathered from student play?" The epistemic stances described in the 2SM are best suited as general descriptions of styles or strategies of play, and thus, a more targeted approach is warranted to investigate the effects of these styles on specific conceptual understandings that gameplay intends to promote. Research Question 2 (RQ2) provides this specificity by asking, "Do differences in gameplay in the specific game situations correlate with differences in performance on a conceptual knowledge test?" Each question is investigated in its own section, with separate Results and Discussion subsections. In the Conclusions, we outline some of the opportunities and difficulties of using educational data mining on digital game play logs, future directions for this kind of research, and also propose improved design factors for educational games that might better promote students' behaviors during play to more closely align with those behaviors found linked to positive learning outcomes.

CONCEPTUAL FRAMEWORK

Overview of the Game Environment: The Fuzzy Chronicles

For this study, we used the educational game titled *The Fuzzy Chronicles*, codenamed EPIGAME (Clark, 2012; Clark, Sengupta, Brady, Martinez, & Killingsworth, 2015). The Fuzzy Chronicles is the third iteration of the SURGE line of digital games intended to help students advance their understanding of Newtonian kinematics. *The Fuzzy Chronicles* (hereafter, EPIGAME) takes the form of a series of puzzles presented as a science fiction adventure. Students play as the space navigator Surge, who must find and rescue space capsules piloted by *Fuzzies*, adorable but somewhat hapless creatures who are stranded in space. In order to accomplish these rescues, the student must navigate Surge's spaceship through a two-dimensional spatial grid (see Figure 1 and Figure 2) by tracing a *Trajectory* to the stationary Fuzzy, then placing *Actions* at *Waypoints* along that Trajectory. Most Actions take the form of *Boosts* that propel Surge's ship in one the four cardinal directions with an amount of force that the student chooses. Gameplay is divided into *Levels*, each comprising a separate navigational and/or rescue challenge. All Levels have a *Start Point* and an *End Gate,* and may also optionally contain obstacles, such as impenetrable *Nebulas* and *Radiation*, as well as *Velocity Gates*

Figure 1. Anatomy of an EPIGAME level. (1) Start Point (2) Velocity Gate (3) Laser Deactivator (or "Button") (4) Nebula (5) Matching Button and Laser (note green color of both) (6) End Gate (7) available Actions.

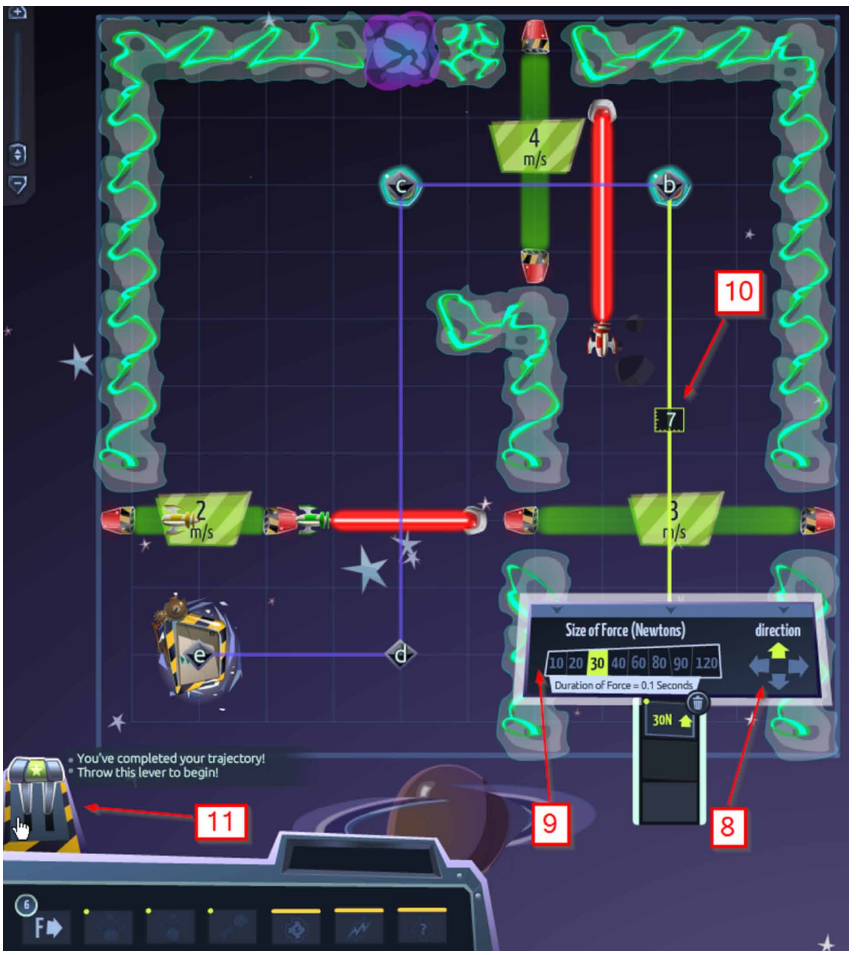


and *Mass Gates* that impede Surge's progress. These Gates signal an attribute of Surge's capsule (i.e., a specific velocity or mass) that the student is required to match before the Gate will open. Colliding with a Nebula, a Radiation field, or a Gate causes the destruction of the Surge capsule and any rescued Fuzzies, and failure of the Level.

The interactive structure of EPIGAME has two phases - a *planning phase* and an *action phase*. In the planning phase, students decide their trajectory and place their actions appropriately. The student signals the end of the planning phase by hitting the *run lever*, thus starting the action phase. In this phase, Surge's capsule follows the student's plan, which may result either in a successful navigation to the end gate and the rescue of any stranded Fuzzies or the destruction of Surge's capsule. If successful, the student moves on to the next level. If the student is not successful and Surge's capsule is destroyed, he or she is returned to the planning phase in order to change the planned trajectory and/or add or remove actions before triggering a new action phase. Together, a planning phase and its resultant action phase are called an *attempt* (which may be successful or unsuccessful).

In theory, a student may complete the game having needed only one attempt, (i.e. one planning phase and one action phase) per level. In reality, students often require multiple attempts before they

Figure 2. An attempt in process. The student is setting direction (8) and force (9) parameters on an action. The student has set a trajectory (10) through several waypoints (a-e). To begin the attempt, the student presses the launch lever (11).



successfully advance. In a given level, the student is free to construct a plan for the entire trajectory for the entire level and place all necessary actions before first activating of the run lever. Alternatively, students may choose to segment the trajectory and place only a few actions at a time, thereby solving the level incrementally (i.e., draw part of a trajectory, place a few actions, activate the Run Lever, see what happens, and adjust and extend the trajectory and actions iteratively through multiple cycles of attempts). The game neither suggests nor encourages either approach, so a student may select whichever method he or she finds more suitable.

A full game of EPIGAME as designed for this study consists of 32 levels of generally increasing complexity. Each subsequent level more often than not requires more actions than the previous ones, contains more challenges and obstacles, and demands more effort by the player to plan and strategize for success. Because of this, it is likely that any students of EPIGAME will find at least one level that requires multiple attempts in order to succeed. Some levels, particularly near the end of the game, allow only a very limited margin of error. Therefore, progress in the game requires the student to be persistent at times, take several different approaches when faced with apparently insurmountable levels of difficulty, and explore and experiment with different combinations of actions to find a correct solution for each level.

The Two-System Framework of Game-based Learning

A goal of this paper is to investigate a theory of game-based learning called the Two-Stance Model framework, or 2SM (Martinez-Garza & Clark, 2016). The 2SM framework seeks to support a more sophisticated understanding of how and what people learn from digital games. It was motivated by the contrast between recent scholarship that finds uneven evidence that people learn much from digital games (Young et al., 2012) and the observation that students inhabit rich ecologies of knowledge about the games they play (Gee, 2007) that include often-impressive feats of cognition.

Many digital games can be accurately described as software models of scientific phenomena encased within game-like structures that are intended to increase student engagement. In the case of educational games, the intention is that students develop an understanding of the principles that underlie these phenomena through the thoughtful and purposeful exploration of their scientific models. The premise of the 2SM framework is that students of educational games do not necessarily form accurate mental analogues of the software models that drive the phenomena they experience in-game (i.e. the encased "simulation"); rather, they create a second-order model (as in, a model of a model of a phenomenon) that is oriented towards explaining the functioning of the encased simulation, predicting its future states, and allowing the student to feel that he or she understands the simulation or game, and has some measure of control over it.

These two stances can be conceptualized further using features from the two-system model of cognition (Evans, 2008). Two-system models of cognition distinguish between effortless thought, or "intuition", and deliberate purposeful "reasoning". These modes of cognition are neutrally labeled as System 1 and System 2, respectively. The former is described as fast, automatic, associative, emotional, and opaque; the latter as slower, controlled, serial and self-aware. In the 2SM framework, System 1 is associated with the "player" stance and System 2 with the "learner" stance.

Students might have two distinct goals when interacting with a game's encased simulation. The first involves develop their second-order model to better understand the simulation and use it as a laboratory the objects and relationships within the simulation can be investigated. The second goal involves executing various game actions to manipulate the simulation to create the desired state (i.e., success). These two sets of goals imply different forms of thinking about the information being presented by the digital game. Our hypotheses are that (a) the first goal prioritizes or incentivizes an inquiry stance oriented towards the purposeful and systematic investigation of the operating principles of the encased simulation and that (b) the second goal prioritizes or incentivizes a heuristic-driven problem-solving stance oriented towards efficiently achieving the player's goals. A student in the inquiry (or "learner") stance might probe the simulation for information that confirms their understanding. A student in the problem-solving (or "player") stance might only engage in exploratory actions and observe whether these actions lead to positive results.

Starting from the two-system model of cognition, we proposed the following mechanistic explanation for how people play and learn from digital games. A person begins play, and a goal will be suggested to the player's thinking, immediately triggering a self-query, "how do I achieve this goal?" The self-query shifts the person towards the learning stance, and in response to the query a second-order model is constructed. This model's functional requirement is that it suggest actions that would bring the state of the game closer to what the person has identified as a goal state. These actions are rendered as execution steps ("Do that") and enacted in the simulation through the game's interface. Actions that prove effective are reinforced and actions that have a negative effect are rephrased as avoidance steps ("Don't do that"). With repeated reinforcement, effective rules are matched to the context cues from the environment and stored as conditionals, i.e. "If this, do that." These conditionals are easy to remember, quick to access, and require nearly no cognitive effort to execute: they fit the functional definition of heuristics.

Whenever the student finds herself in a situation that is covered by a stored rule, she will in most cases default to doing what that rule stipulates. In other cases, the student must shift to a learner stance, reinstate the second-order model, and use it to find new possible actions. If the student always

knows the rule to apply, the model is most likely deactivated and the student will default to System 1-style processing, or fast, effortless, intuitive heuristics. Thus, through play, a person gathers three forms of knowledge about the game: (a) the conditions that the game presents, (b) a set of heuristics, or rules of action with activation criteria that match these conditions, and (c) a second-order mental model i.e., an idiosyncratic explanation of how the game produces the observed conditions. In the case of educational gaming, these three forms of knowledge combine to form part of the learning benefit that students may develop from playing the game.

The 2SM is a novel application of the two-system theory of reasoning to educational games. There are suggestive findings from adjacent programs of research have examined forms of reasoning within and around digital learning environments that hint at its validity (e.g. Parnafes & Disessa, 2004; Gijlers & de Jong, 2013). One of the goals of this paper is to explore the fundamental claims of the 2SM, namely that traces of students' System 1 and System 2 reasoning can be observed during play, and that preference for one stance over another has a significant effect on learning. These possible effects are explored in more detail in the following section.

Implications of the 2SM for Learning

In the 2SM, stances are defined as collections of resources (Hammer & Elby, 2003). The framework stipulates that the two stances can be associated with cognitive processes described in the two-system theory of cognition (Sloman, 1996; Stanovich, 1999; Kahneman, 2003; Evans, 2008). Thus, a stance or collection of resources organized around System 1 would be optimized for processing speed and effortless thought, while a stance organized around System 2 would be primed for information use and deliberative reasoning. Stances, like resources, are cued around task demands; certain tasks, e.g. driving a car, are structured in a way that they discourage analytic reasoning, while others, like academic writing, are less amenable to quick, associative thinking. That said, human beings are biased in general towards System 1 reasoning as an effort-saving and time-saving strategy (Reyna & Ellis, 1994).

The question then becomes, which of the two stances is most conducive to learning? Intuitively, it would seem that the effortful, analytic processes described as System 2 that drive the learner stance would be preferred over faster, less deliberate thinking. This would be particularly true in the case of games that are *conceptually integrated* (Clark & Martinez-Garza, 2012) because such games are designed in such a way that thinking about game rules and challenges closely parallels thinking about science concepts and relationships. However, it is unlikely that an educational game can sustain System 2-type processing over long periods. First, students will tend to find ways to save time and effort when negotiating cognitively-demanding challenges, i.e. the "cognitive miser" of Fiske and Taylor (1991). Secondly, players facing a game they consider *too* challenging may simply disengage, thus negating any educational benefit the game might offer. Thus, a "happy medium" may be more desirable in which players both (a) reflect deeply about concepts and ideas represented in the game and (b) put their understanding into practice in motivating and interesting ways.

As many educational games, EPIGAME is intended to invite learners to think and reason about the concepts and relationships the game portrays and not to merely passively experience them. Players of EPIGAME encounter obstacles and situations of increasing difficulty that are designed not only to provide opportunities for learning but also to adapt to players' increasing knowledge and proficiency over the course of the game. Ideally, students encounter game levels whose difficulty matches but does not significantly exceed their own skill - this alignment keeps interest and engagement high even in the face of ostensibly higher cognitive demands (cf. "flow" in Csikszentmihalyi, 1991). This adaptation is not perfect: students may encounter game levels that are too difficult or too easy. The goal is ultimately not to shield students from difficulty but to provide enough scaffolding and feedback so that the *perceived* difficulty remains manageable.

We propose that a student's response to perceived difficulty cues the stances. Which stance is cued may depend largely on each student's developing understanding of the concepts and relationships

underlying the game. Early in the game, the perceived difficulty may be influenced by the student's prior experience with similar games or familiarity with the game's targeted concepts and relationships. Thus, the student's prior knowledge of the game or the principles behind the game's encased simulation may also be a significant factor that cues and organizes the stances. For instance, students with low prior knowledge might prefer a slower, more methodical approach, while students who feel confident in their understanding might play faster, and with less tentativeness, because they may have a more detailed and functional internal model. Later in the game, once all students have had similar opportunities to engage with the game's challenges, these differences might not be so stark, or they may disappear altogether. Therefore, it becomes important to examine the students' gameplay to ascertain how the game's varying set of structures and experiences influence students' learning.

Learning Analytics in Educational Gaming

Digital environments that promote learning should prompt a change in student behavior within that environment. If an educational game is designed in such a way that students are able to apply what they learn in the context of the game, then these changes in behavior should be reflected not only in external measures of learning but in play itself. If so, then these changes are potentially recoverable and traceable from log data *post hoc*. However, even comparatively simple games allow for a broad range of student interactions, all of which leave their varied and distinct traces. Changes in student behaviors that signal learning can, therefore, be easily lost in the vastness and complexity of the available data. Methods based on learning analytics (LA) can provide researchers with tools to classify, predict, and discover latent structural regularities even in data sets as voluminous and idiosyncratic as game play logs (Berland et al., 2014). LA techniques not only can help us characterize and describe learning behavior, but they can also deploy Markov-type approaches, such as Bayesian knowledge tracing and performance factors analysis, to provide some insight into latent student knowledge. Interestingly, these Markov-type models could be used for prediction, and not just description; for example, they could be used to guide adaptive scaffolding and feedback. That said, while more research is required for these applications to achieve their full promise, significant ongoing work is already exploring and refining the use of learning analytics on data logs from educational environments.

The use of in-game performance data as evidence of learning outcomes has been proposed by Shute (Shute & Ventura, 2013) and others. Shute and colleagues propose that a learner's actions within the game environment can be used as a form of assessment when evaluated against an evidence model, as per the evidence-centered design (ECD) assessment framework (Mislevy, Almond, & Lukas, 2003). Under this framework, evidence models are preceded by activity models, which are contextualized and tailored to the particular affordances and constraints of the learning environment. One implementation of EDC which seems particularly suited to educational games, "stealth assessment", aims to collect model data directly from the learning environment, bypassing the need for overt knowledge testing that may detract from the play experience. Using this methodology, Shute and Ventura have measured both learning of specific knowledge, e.g. as qualitative physics (Ventura, Shute, & Small, 2014), and also broad cognitive skills and traits, such as persistence (Ventura, Shute, & Zhao, 2013) and 21st-century skills (Shute, 2011).

Activity models can become highly complex, especially in the case of games in which many different interactions are possible. This complexity often leads to a large number of observable variables, which in turn complicates the task of formalizing them into an evidence model. For this reason, researchers have found value in machine-learning (ML) techniques of computational statistics that can make finding patterns and relationships between large numbers of variables more tractable. Examples of educational games where researchers have used ML techniques to analyze student performance data along an EDC paradigm are the investigation of systems thinking in *SimCityEDU* (Mislevy et al., 2014) and inquiry skills in *Mission Biotech* (Lamb, Annetta, Vallett, & Sadler, 2014). ECD models that are focused on content-specific outcomes that apply ML techniques are also feasible, such as the investigation of student learning of biological processes of stem cells

in *Progenitor X* (Halverson & Owen, 2014); of fraction arithmetic in *Save Patch* (Kerr & Chung, 2012) and of Newtonian mechanics in *Impulse* (Rowe, Asbell-Clarke, & Baker, 2015). There are several more exemplars of ML techniques that are used to characterize students' performance in digital environments, although these focus either on learning environments that are simulation-based (rather than game-like) or do not align exactly with an EDG paradigm. Researchers have successfully applied ML techniques, for example, to describe (a) students' science inquiry activity in *Science Assessments* (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012) and in *Virtual Performance Assessments* (Baker & Clarke-Midura, 2013; Clarke-Midura & Dede, 2010); (b) students' developing engineering thinking in *Nephrotex* (Chesler et al., 2015); and (c) students' understanding of genetics in *BioLogica* (Buckley et al., 2004).

RESEARCH QUESTIONS

The groundwork laid thus far has discussed the 2SM as a theoretical perspective for examining gameplay and discussed learning analytics as an approach for analyzing game play through data logs. The next step is to articulate the specific hypotheses and the kinds of evidence that might support them. As mentioned in the Goals section, this paper has two research questions, which we expand upon in greater detail in the following paragraphs.

Question 1: Can the Two Stances of the 2SM, as Specified by the Framework, be detected in Game Play Data?

The first question is intended to test a cornerstone claim of the 2SM, while also evaluating whether the 2SM is a useful lens for interpreting game play data as recorded in The Fuzzy Chronicles. The hypothesis is that game play logs exhibit an underlying interpretable structure when features relevant to the 2SM are selected and analyzed. Alternatively, in the case of the null hypothesis, there will be no such structure, or it will not be easily interpretable, or the structures revealed do will not correlate significantly with learning outcomes. Such a result would indicate that gameplay is more like a stochastic process, or idiosyncratic, or that players are using purely reactive or irrational processes rather than those grounded in cognitive models of performance.

Question 2: How Do Changes in Students' Functional Understanding of the Game Relate to Performance on a Test of Conceptual Understanding?

The second question refers to the feasibility of directly assessing students' emergent understanding of the concepts of Newtonian kinematics represented in The Fuzzy Chronicles based on their solutions to small, localized challenges. Each maneuver the students are asked to make in EPIGAME (starting and stopping, changing directions, keeping to a set velocity, picking up or throwing an object, etc.) is designed to reify a relevant concept or cognitive resource. By identifying and analyzing students' actions with regard to challenges *of the same type,* both within a student and over time, or between students, we can better understand how these challenges focus thought and learning for individual students. Since EPIGAME is intended to be a conceptually-integrated game (Clark & Martinez-Garza, 2012), the hypothesis is that improved performance in these conceptually-laden challenges indicates a greater understanding of the underlying principles of Newtonian kinematics. If the null hypothesis is true, variations in student performance will not correlate significantly with learning outcomes.

METHODS

Studies and Participants

To investigate the research questions, we performed two experimental runs using EPIGAME in the months of March and April 2015. The first run was used to address possible confounds as well as

pilot the gameplay data "pipeline," or the entire process of collecting, collating, testing, and analyzing EPIGAME logs. We report on study 1, the pilot study, only briefly as foundation and comparison for study 2. The second study, which is the focus of the current manuscript, deployed the full data analytic process to investigate both research questions. The two studies used the same EPIGAME version, the same assessments, and had roughly the same duration.

Study 1 (Pilot Study)

The participants were 86 9th grade students from a public high school in Middle Tennessee. In this study, the students were divided into four groups, each randomly assigned into a Solomon four-group design (Solomon, 1949) (Figure 3). The two non-treatment groups participated in their normal classroom curriculum on the topic of force and motion, while the treatment groups only played the game for three 90-minute sessions. Approximately 20 minutes were reserved at the beginning and end of the entire study for a 21-item multiple-choice test intended to assess the students' conceptual and qualitative understanding of Newton's First and Second Law. Two of the groups, one treatment and one non-treatment, completed pre-tests; all four groups completed post-tests 5 days after the experiment began.

The 4-group Solomon experimental design was used in order to obtain a test of the internal validity of the posthoc effect sizes and test for interactions between the pre-test and the intervention. Our initial conjecture, in line with the 2SM, was that high pre-test score indicating high prior conceptual understanding of physics would enable students to form more advanced play strategies. The use of these strategies would then be reflected in post-test gains. However, students might also be primed by the relationships and situations that appear in the pre-test, and post-test gains might correspond not to differences in gameplay or in prior knowledge, but in a testing effect. Thus, the goal of Study 1 was (1) to determine whether the version of EPIGAME was effective as a learning experience, (2) to investigate any possible testing effects, and (3) to prototype the data collection protocol and some of the analytical techniques. The statistical treatment of the four-group design that allows this disentanglement can be found in Braver and Braver (1988):

Two-way within-subjects ANOVA (Table 1) performed on the assessment data showed that students in Study 1 made significant pre-post gains ($F = 10.61$, df = 104, $p < 0.01$), with no strong evidence in favor of testing effects ($F = 1.11$, df = 104, $p = 0.29$) or interactions between pre-test scores and treatment ($F = 0.36$, df = 104, $p = 0.55$). This represents strong evidence that whatever

Figure 3. The Solomon 4-group design. Graphic from Braver & Braver (1988).

	Time			
	Period One (Pre)			Period Two (Post)
Experimental Group One	*R*	O_1	*X*	O_3
Control Group One	*R*	O_2		O_4
Experimental Group Two	*R*		*X*	O_5
Control Group Two	*R*			O_6

O = Observation
R = Random Assignment
X = Treatment

Table 1. Two-way within-subjects analysis of variance for Study 1

Effect	DF_n	DF_d	*F*	*p*	ges
pretest	1	104	1.1161	0.29	0.01061
treatment	1	104	10.614	0.002 **	0.09260
pretest:treatment	1	104	0.3552	0.55	0.00340

knowledge students are bringing into gameplay was not gleaned from the pre-test, nor did the pre-test prime students as to which relationships or interactions were important and thus biasing performance in the post-test.

Study 2 (Research Study)

Study 1 helped to discard two competing hypotheses: that EPIGAME is not effective as a learning tool, so any patterns or changes in gameplay cannot affect learning, and that pre-testing rather than gameplay is the source of any observed pre- to post-test gains. The remaining hypothesis, that differences in gameplay are the source of pre- to post-test gains, is the focus of Study 2. In this second study, 123 7th grade students from a public middle school in Middle Tennessee used the EPIGAME software as part of their normal classroom instruction for five consecutive class periods lasting 45 minutes each.

As in the prior study, each student had his or her own computer and was specifically instructed to avoid sharing information. The blanket policy was to provide encouragement or hints in lieu of direct assistance, but help was provided to students who appeared intractably stuck, were having technical issues, or had urgent questions about the game interface. As in study 1, approximately 20 minutes were reserved at the beginning and end of the intervention for a 21-item test of conceptual understanding of force in motion. In this study, all students who were present at the first and last day of the intervention were asked to complete the assessment.

Thus, students who participated in each of the two studies generated two forms of data: pre-post assessment data and game play data. The pre-post assessment data was anonymized and students with missing pre- or post-test scores were dropped from the study. In the case of students with complete pre- and post-test scores, a unique ID was generated for each; that unique ID was used to link the assessment data with the game play data.

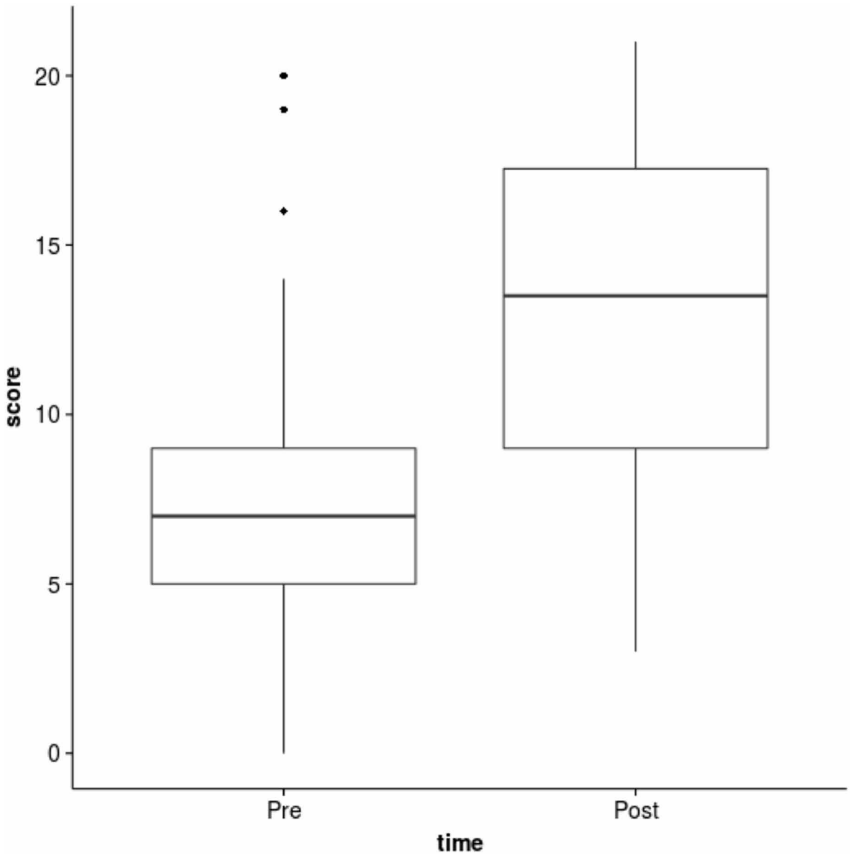
Of the 123 students who participated in the study, 104 provided both pre- and post-tests. A matched-pairs *t*-test showed a statistically significant increase in test performance ($t = 11.702$, df = 103, $p < 0.0001$) (Figure 4). The value of Cohen's *d* suggests a large effect size ($d = 1.62$).

Emulating the Evidence-Centered Design Approach

In the Learning Analytics in Educational Gaming section (above), a significant portion of the research reviewed that used learning analytics to make sense of students' process or log data used an evidence-centered design (ECD) framework for assessment as well. ECD offers several notable advantages for this form of research, *viz.*:

1. The Student Model serves to constrain the number of latent variables that the ML algorithm must infer, aiding in model fit.
2. The Evidence Model provides identification rules and ready-made coding schemes, boosting the interpretability of the final model.
3. The Task Model pre-selects observed variables that are likely to be significant, obviating the need for dimensionality-reducing steps, such as a Principal Components Analysis to help reduce the number of observed variables to a tractable number.

Figure 4. Boxplot of pre- and post-test results for Study 2



Considering these advantages, it is clear that learning analytics and ECD processes are well-suited for each other. Unfortunately, it is likely unworkable to apply the ECD framework retrospectively, as the products of ECD are intended to address the specific purposes of that particular assessment (Mislevy, Almond, & Lukas, 2003). Thus, the goal would be to emulate some useful features of ECD, such as the student model and the evidence model. The student model can be operationalized in terms of the hypothesized dynamics of the 2SM. The evidence model would then map these dynamics into the observable variables. The end result would not be nearly as robust as the full ECD evidentiary argument, but would at least qualify as a cognitive model of task performance, or an illustration of the thinking processes underlying the knowledge and skills students apply *in vivo* when solving educational tasks in a specific domain (Leighton & Gierl, 2007, p. 10).

An important feature of learning analytics and machine learning methods is that they generally do not aim to produce results that have inherent meaning. Unlike statistical treatments of parametric data, such as pre- and post-test results, in which a statistically-significant result indicates a change in the participants' behavior along a measured construct, machine learning and data-mining algorithms generate, at most, descriptions of likely patterns and structures present in the data. It is up to the analyst to interpret what those patterns and structures mean, and evaluate whether or not they support the proposition being researched (Vellido, Martin-Guerroro, & Lisboa, 2012).

Ideally, the interpretation of patterns and structures revealed by learning analytics are supported by robust theory. That is, features discovered in the data align with existing constructs and relevant explanations for the learning phenomena being studied. In this case, the proposed interpretive lens is

provided by the 2SM. Under the 2SM framework, students use collections of resources, or stances, that organize around the cognitive processes that are optimized for fast ("player") or slow ("learner") processing. Thus, the first task is to theorize how these stances would manifest as students play EPIGAME; in other words, we needed to determine how the "fast" and "slow" resources would affect gameplay. Evoking the evidence centered design paradigm, we will call this operationalization the "student model". The second task is then to create an "evidence model," that is, to deduce how the actions and strategies defined in the student model will appear in the gameplay data logs. The goal of the evidence model is to select, from all the information contained in the logs, which pieces of data are most likely to characterize the operations defined in the student model.

The Student Model

The trial-based dichotomous pass/fail task structure of EPIGAME suggests two general strategies for arriving at a solution, one mainly using "fast" processing, and the other using "slow" processing. These strategies, or modes, are:

1. Additive-Iterative Mode, in which a student solves a level through a step-by-step iterative accumulation of actions, each checked for efficacy in a separate attempt.
2. Solve-and-Debug Mode, in which an entire solution is drafted whole-cloth, then corrected only if and as necessary.

While both of these approaches imply that the learner is *thinking,* they differ in what students are thinking *with,* and what they are thinking *toward.* A student using the Additive-Iterative Mode does not necessarily have to have a working knowledge of the game's concepts and relationships in mind; all he or she requires is that EPIGAME provide an unambiguous signal that each added action is a step towards a solution (which EPIGAME provides, in the way of visually-clear animations, e.g., of Surge's capsule exploding or of the Exit Gate being activated). The Additive-Iterative Mode can be thought of as related to Parnafes and diSessa's (2004) "constraint-based thinking." On the other hand, a Solve-and-Debug approach necessitates that the student has a vision of a solution. Armed with a good working knowledge of the rules of operation, a student might feel more capable of taking more actions within each trial because he or she has a reasonable expectation that those actions will be effective. The Solve-and-Debug Mode can be thought of as related to Parnafes and diSessa's (2004) "model-based thinking."

Evidence Model

The two strategies described above represent the best estimate of the forms of play that students are most likely to use. While these forms of play sound very different mechanistically, it is useful to think of them as opposite ends along a continuum. On one end of this continuum, the Solve-and-Debug Mode is slow to plan, is more likely to be correct, and if it is not, it may require only small, effective fixes. On the other end, the Additive-Iterative Mode is fast, less likely to be correct since a student using this mode may not always define a full solution, and the iterative fixes are more error-prone. Thus, the differences between these two approaches may be captured with only a few contrasting parameters (Table 2).

The first and third parameters, Response time and Actions per attempt, are straightforward and directly observable in the data. A longer Response Time indicates slower, more deliberate processing; shorter Response Time corresponds to quick decision-making. Similarly, the number of Actions per attempt is likely specific to each Mode: more Actions taken in the same attempt implies a more elaborate, thought-out plan, while fewer Actions might indicate iterations or corrections.

The second parameter, *Error rate*, will have to be computed from other variables. Broadly speaking, the difference in Error rate between the two Modes represents the willingness of students to accept failed Attempts. Failure during an Attempt is more or less required in the Additive-Iterative

Table 2. Forms of solution and their likely parameters

Parameters	Additive-Iterative	Solve-and-Debug
Response time	Low	High
Error rate	High	Low
Actions per attempt	Low	High

Mode, since a student may consider failure as a "partial success" if it creates a baseline upon which he or she can iterate. A student using this Mode may also create a partial solution with some set of parameters he or she knows, and guess at the remaining parameters, counting on the fact that the game will provide actionable feedback. On the other hand, failed Attempts when using the Solve-and-Debug mode are more likely to be unintentional or unforeseen mistakes, rather than intentional probes or guesses. Students using the Solve-and-Debug mode seek to avoid error rather than accept it as inevitable. Thus, the Error rate parameter should incorporate information on how often students fail a level repeatedly, as this continued error would indicate unsuccessful guessing and/or low-information processes such as exhaustive testing of all the available actions.

Treatment of the EPIGAME logs

The data analysis of EPIGAME logs from Study 2 proceeded in four phases:

1. data normalization and integrity checks
2. variable selection and dimensionality reduction
3. clustering of student gameplay data and sequence mining, for Question 1
4. contextual feature mapping, for Question 2

Phase 1

The initial corpus of gameplay, recovered directly from the classroom WISE server, was composed of 16,239 records. Each record was comprised of one particular student's attempt to solve one particular level. The particular build of The Fuzzy Chronicles used in this study had 32 levels; thus, each student produced an average of 132 attempts, approximately 4 attempts per level. Each record comprised a JSON object detailing the specific parameters of the attempt the student performed, such as where on the map an action was placed, how much time the student took to plan their actions, and which values the student chose for each parameter of each action. The dataset contained approximately 1.1 million of these gameplay parameters.

We then extracted a set of variables to help characterize each attempt. Broadly speaking, we extracted two kinds of variables: observed and derived variables. Observed variables are characteristics of gameplay directly recorded by the EPIGAME software, such as planning time. *Derived* variables are those discovered through logical tests or comparisons performed on observed variables, akin to a coding scheme. A total of 23 observed and derived variables were defined, each capturing an element or aspect of gameplay (see Table 7 in Appendix A for a complete description of these variables). These 23 variables were selected on the basis of their ability to describe differentially the parameters for the forms of solution described in Table 2.

Phase 2

Generally, when using LA techniques, it is most desirable to have a data set with the smallest, most meaningful set of variables possible. Datasets with large numbers of variables are computationally very expensive to process, and such data is vulnerable to a variety of phenomena that distort results

and complicate these types of analyses. In order to select only the most meaningful variables, we performed a Principal Components Analysis (PCA) on the dataset (16,239 attempts x 23 variables) using the *FactoMineR* software for *R* (Husson, Josse, Le, and Mazet, 2007). The PCA returned 3 components with eigenvalues greater than 1, with a total of 72.1% variance explained by those three components. The full results of the PCA are included in Table 3 (below). The variables associated with the components were:

1. Component 1:
   a. *tl.Modifys*, a count of how many modifications a student made to the parameters of placed Actions, e.g. changing a Boost from 10N to 20N increases *tl.Modify* by 1.
   b. *tj.Adds*, a count of how many Waypoints were added to the Trajectory.
   c. *planningTime.log*, the observed time students spend planning and placing elements, in seconds, logarithmically transformed to amplify the difference between planning times of lower times, such as 5 and 8 seconds, but de-emphasize the difference between higher times, such as 47 and 50 seconds.
   d. *eff.actions.added*, a derived variable counting how many new Actions were executed effectively on a given attempt compared to the previous attempt.
2. Component 2:
   a. *par*. A model-based effectiveness score derived from a Markov-chain model of the combined series of outcomes of all the students' plays of each level. Each student generated a chain of Attempts for each level, and each attempt had a particular outcome, e.g. one Attempt ends in a navigation error, then two Attempts ended at Velocity Gates, then the next Attempt ended at the Success Gate. This chain of Attempts captures each student's transversal of a level. When all students' chains of Attempts for a given level are taken together, we can use a Markov-chain model to calculate the probability that a student will transition from one outcome to another on a per-attempt basis. The model is then used to calculate *par,* which is the posterior probability of a Success state occurring randomly at the end of an Attempt given the state at the end of the previous Attempt. These probabilities can range from [0,1], with 0, or no chance of success on the next attempt, being indicative of random play, and 1, or certainty of success in 1 more attempt, indicating expert play. In other words, the *par* metric asks, "if this student were playing totally randomly – that is, following only the transitions observed for all students as a whole - given that his or her last attempt ended in a certain outcome, what is the probability that he or she will find the Success Gate through sheer

Table 3. Results of the principal components analysis

	Component				
	1	2	3	4	5
Modifications to Timeline	0.4746	0.0015	0.0000	0.0001	0.0014
Additions to Trajectory	0.4173	0.0078	0.0000	0.0001	0.0000
Effective Actions added	0.1787	0.0000	0.0014	0.2471	0.0000
Par metric (square root transformed)	0.0345	0.4694	0.0000	0.0000	0.0005
Change in Par metric from previous Attempt	0.0001	0.7498	0.0000	0.0002	0.0005
Planning time (log transformed)	0.3297	0.0028	0.0000	0.0274	0.0000
Test of consecutive similar failure	0.0854	0.0048	0.1082	0.0000	0.0890
Test whether Level was aborted	0.0000	0.0014	0.6434	0.0000	0.0220

Note: values are given as squared cosines

chance in one more attempt?"[1] An important property of this metric is that it penalizes very long chains of Attempts and rewards navigating to the Success Gate on the first Attempt. The *par* score was later transformed into *par.sqrt* via a square-root transformation to make the probabilities more legible.

b. *par.delta.sqrt*. The change in the value of the *par.sqrt* metric from attempt *n*-1 to *n* for the current level and student.

3. Component 3:
   a. *is.abort*, an observed variable that tests whether or not the student manually aborted the attempt using the Abort button.
   b. *fail.same,* a derived variable that tests, if an Attempt was failed, whether or not a student failed that Attempt at the same place on the map as the immediately-previous Attempt *and* whether both Attempts failed for the same reason. A TRUE value indicates a consecutive unsuccessful attempt by a student to navigate past a specific obstacle on the map.

Further analysis revealed that since *par.sqrt* and *par.delta.sqrt* were linear combinations of each other, *par.sqrt* could be discarded in favor of *par.delta.sqrt*, which has the higher squared cosine for Component 2. At this point, further treatment of the data followed the line of inquiry specific to each research question. Relevant details can be found in their respective sections below.

RESULTS RQ1: CAN THE TWO STANCES OF THE 2SM BE DETECTED IN GAMEPLAY DATA?

The main claim of the 2SM is that the stances organize around fast- and slow-processing mechanisms. Therefore, it is reasonable to look for play strategies that embody fast and slow play. After the dimensionality reduction process above, we are left with a manageable number of variables which are nonetheless theoretically significant and useful in describing these strategies. To explore Research Question 1, we apply LA techniques exploring the variables in terms of clustering and then in terms of sequence mining. We then discuss the implications of the findings in terms Research Question 1 and the proposed 2SM framework.

Clustering

The next step in the analysis is to examine the dataset to determine whether students' play has some latent order or structure that can be brought into focus using our theoretically-relevant variables. To find this possible structure, we will use *clustering*, an unsupervised classification method. The goal of a clustering algorithm is to find the groups of observations whose features are more similar within-group than with regard to the data at large. Since this technique is unsupervised, we do not provide a pre-determined classification scheme for the software to "learn"; the rationale for this choice is that if a clustering algorithm returns a reasonably-interpretable set of clusters and these clusters were created by interactions between theoretically-significant variables, then that is a solid indication that the theory describes latent structures of the data.

With the final list of seven variables already selected, we proceeded to create a similarity matrix using Gower's coefficient to account for the mixed data types. Then, we performed affinity propagation clustering with the resulting similarity matrix. Affinity propagation (AP) is a clustering method that takes as input measures of similarity between pairs of data points and simultaneously considers all data points as potential exemplars. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges (Frey & Dueck, 2007). This method was selected as preferable to the more conventional k-means/k-medoids method because of its ability to produce a set of meaningful exemplars for each cluster – a vital consideration given the need to later interpret the characteristics of each cluster.

The AP clustering algorithm converged on a set of 145 "proto-clusters" after 260 iterations. These "proto-clusters" were then collapsed using an agglomerative method akin to hierarchical clustering. The resulting cluster dendrogram is given in Figure 5. The lower segment of each line indicates a separate proto-cluster, and the height of the joint between two proto-clusters indicates how similar they are, with greater height indicating more similarity between the clusters being joined.

Visual inspection of the cluster dendrogram suggested that a "cut" at 0.905 altitude would reduce the number of clusters to a manageable six. This clustering solution was codenamed *part.6.* The "goodness of fit" of an AP clustering solution is difficult to ascertain via standard methods such as Rand coefficients because AP clustering does not necessarily aim to produce compact clusters. Rather, it seeks to maximize the "representativeness" of the chosen exemplars. In order to determine the adequacy of the *part.6* solution, we created a heat map from the similarity matrix (Figure 6).

The heat map revealed 3 well-delimited and cohesive clusters along the diagonal, as well as one large cluster with some internal structure, and two additional smaller clusters. We iterated on the *part.6* solution several times in an attempt to resolve Cluster 2 (corresponding to the yellow region) into 3+ smaller clusters as suggested by the heat map, but no satisfactory solution was found that preserved the other clusters, and thus the *part.6* solution prevailed. The distribution of Attempts across the six clusters of the *part.6* solution are given in Figure 7.

Before proceeding to the sequence mining, we studied the properties of the *part.6* clustering. As noted above, the preliminary variable reduction through PCA left us with only 7 theoretically-significant variables out of the original 26. The *part.6* solution represents a mathematical arrangement of students' attempts that have some similar structure in terms of these 7 variables. Figure 8, below, shows a generalized pairs plot (Emerson et al., 2012) that helps visualize how the structure of each cluster responds to each of the featured variables.

From each of the clusters, we visually examined the exemplar chosen by the AP clustering algorithm, the two nearest neighbors to the exemplar, and two random members of that cluster. The 5 members of each cluster were interpreted, both by themselves and in the context of the sequence of level attempts in which they occurred. Based on this analysis, we labeled the clusters qualitatively according to a general description of the students' actions therein:

- **Cluster 1 (in Red): *ABORTS***. Students recognize that the level is going to fail and press the "Abort the Mission" button to preserve the momentum of play rather than allow the simulation to end on its own.

Figure 5. Cluster dendrogram of the AP clustering result

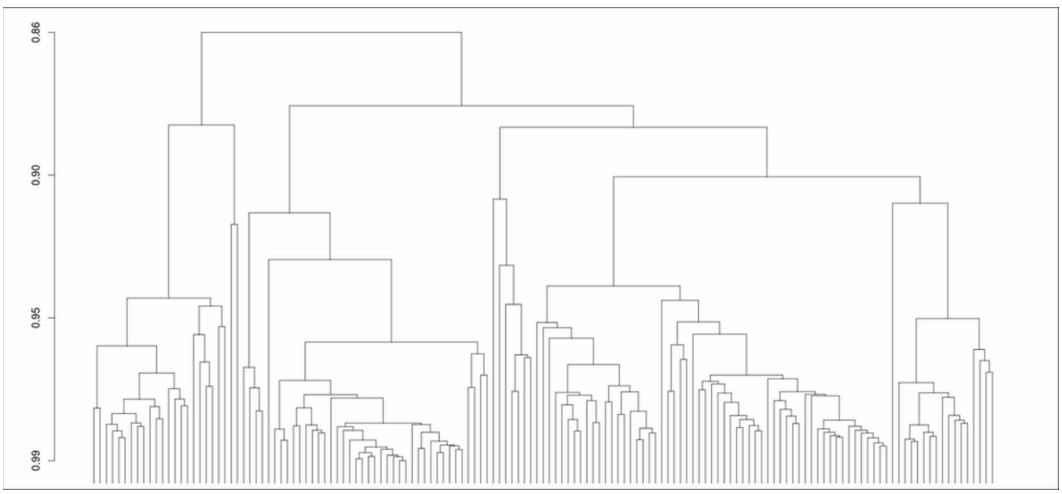


Figure 6. Heat map of the similarity matrix, along with the dendrogram of the part.6 clustering solution. According to the dendrogram, Clusters 2 (yellow) and 5 (blue) are most similar, while Cluster 4 (cyan) is the most distinct.

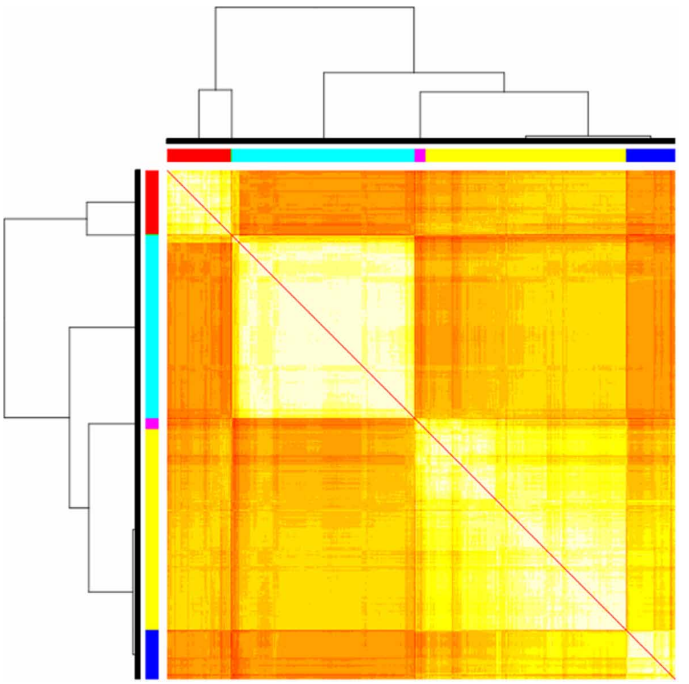


Figure 7. Histogram of the distribution of Attempts across part.6 clusters

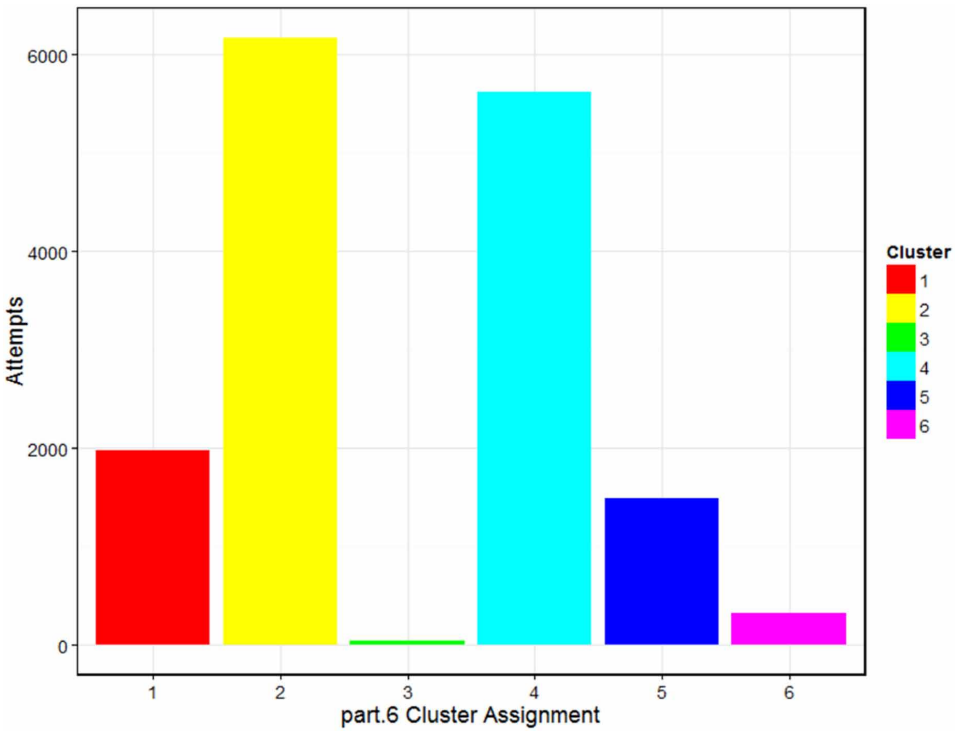
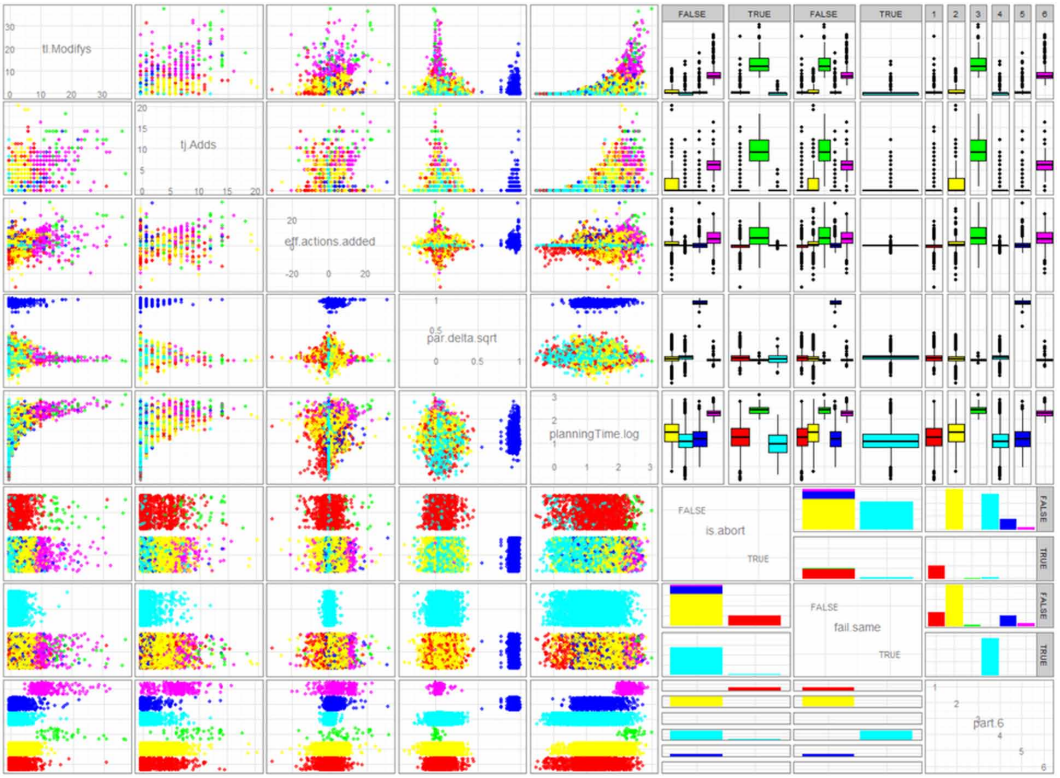


Figure 8. Generalized pairs plot (Emerson, et al., 2012) of the 7 theoretically-significant variables with the highest eigenvalues, plotted against each other, and classified according to the part. 6 solution (rightmost column).

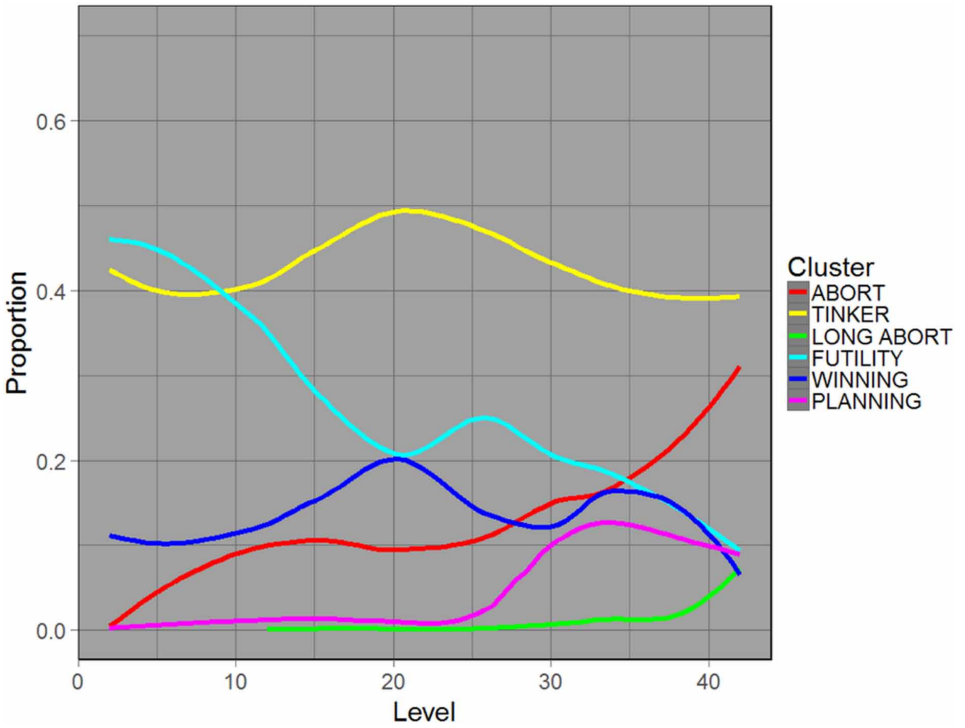


- **Cluster 2 (in Yellow):** ***TINKER***. Students add a few actions, advance a little further along in the level, and fail, but not in the same place on the map as the previous Attempt
- **Cluster 3 (in Green):** ***LONG ABORTS***. Very long planning episodes (> 100 seconds) that end in Abort. A very sparse cluster, barely distinguishable from Cluster 1. Possibly indicates a deletion and restart of the solved level in progress.
- **Cluster 4 (in Cyan):** ***FUTILITY***. Students make a few changes but fail exactly in the same place in the map against the same obstacle as their previous attempt.
- **Cluster 5 (in dark Blue):** ***WINNING***. Students make one or more changes or additions that result in a successful attempt, thus completing the level.
- **Cluster 6 (in Pink):** ***PLANNING***. Students spend a long time and add actions as well as trajectory elements (i.e., added both categories of elements). These attempts are occasionally successful, but not always.

Further investigation revealed that cluster assignments have some structure both in terms of *when* they occur in the order of play (i.e. early levels vs. later levels in Figure 9) and in terms of learning outcomes of the student that produced them (i.e., in terms of pre-post learning gains in Figure 10).

As we can see, the relative distribution of the cluster assignments may be sensitive to the learning outcome of the student (Figure 10). In other words, the levels played by students at a given level of pre-post test performance may have a different ratio of cluster assignments than those of students at a different level of performance. The different frequency profiles in Figures 5 and 6 suggest, furthermore, that the differences are not entirely due to how far students progress into the game. It is clear, then,

Figure 9. Frequency of part.6 cluster assignment by level. Graph shown is smoothed via local least-squares regression fitting (α = 0.65).



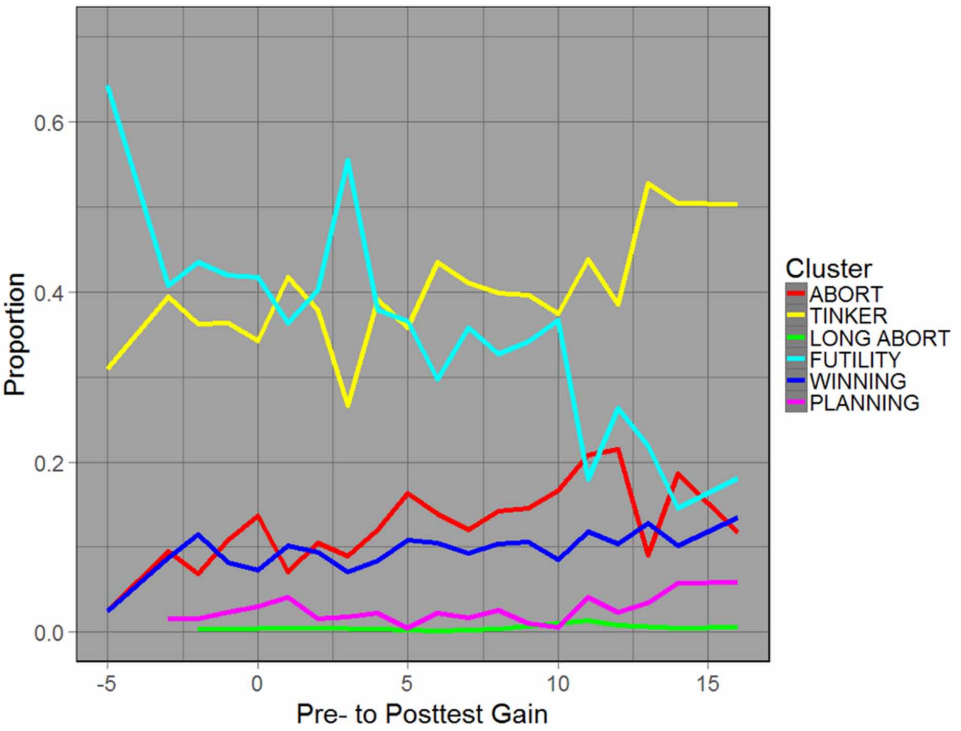
that the *part.6* cluster solution provides not only a set of meaningful code assignments that describe students' play, but also that these assignments are related somehow to learning outcomes. Figure 10 further suggests that cluster patterns evolve as students progress through the game.

Sequence Mining

Sequence mining is a methodology intended to find patterns in sequence data, such as words in a sample of natural language or genes in a protein. The main requirement is that the order of the components is as significant, or more significant, than their frequencies. The question sequence mining asks is, "given a set of items that form sequences, what are the most common smaller sequences to be found within and across those sequences?" In the case of EPIGAME data, the components to be sequenced are cluster membership codes; in other words, our goal is to investigate how students' actions, described individually in general terms by the clustering procedure, appear in succession as a part of a chain of actions intended to solve a level.

The dataset contained 2730 such sequences, meaning that the students' combined attempts to solve any level totaled 2730, or an average of 22.2 levels attempted per student. Each sequence was comprised of the series of each student's attempts to solve a single level; thus, the length of these sequences ranged from 1 to 140, the minimum and maximum number of consecutive attempts recorded in a single level. To perform the sequence mining, we used the *TraMineR* package for *R* (Gabadinho, Ritschard, Muller, & Studer, 2011). This package has the capability to calculate the relative importance of subsequences of elements within the element chains of sequence data. The relative importance of subsequences is measured not in terms of their frequency but in terms of "support", or the proportion of sequences in the overall sample that can claim a given subsequence as a subsequence of itself. The mining algorithm was configured to seek only first-order subsequences, meaning that only events that

Figure 10. Frequency of part.6 cluster assignment by pre-post test score gain



happen exactly consecutively are considered to be in sequence, and the minimum support level was set at 0.01. Thus to qualify for analysis, a subsequence had to be supported by at least 27 sequences. An additional parameter was set so that the support of subsequences of *n* identical codes would be consolidated across all sequences found of one or more identical codes. The algorithm returned 47 candidate subsequences, which were then ordered by support. The results of the sequence mining are given in Figure 11, below:

The height of the bars in the graph indicate the support for that subsequence, and they are ordered by decreasing support. Support for the unitary subsequences, e.g. (2), the most common one, are quite high since, for example, a sequence of (2) - (2) - (2) can claim the subsequence (2) a total of 6 times. Recalling that Cluster 2 stands for TINKER, thus, there is a high proportion of sequences containing long chains of TINKER, and similarly high proportions of chains of FUTILITY. The high support value of WINNING is to be expected since 97% of all sequences end with WINNING, which is how students advance in the game after all.

To investigate the relationship between play sequences and learning, we then classified students according to their pre-post test performance. Since the group of students as a whole gained significantly in their pre- to post- test scores, we chose a classification strategy that would qualify their gains relative to the group. The resulting classification scheme is summarized in Table 4 (below). The "High Prior" group consisted of students who scored in the upper quartile in both pre- and post-tests. The "Low Prior" group is likewise formed of students who scored in the bottom quartile of the pre- and post- test. A third group, "Learned" contains students whose pre-test scores were in the lower three quartiles but who improved their score by at least one quartile. A fourth group, "Null", collected students whose pre-test was in the higher three quartiles but did not show a significant increase in their scores. The number of students in each classification was 14, 13, 23, and 54, respectively.

Figure 11. The 25 highest-supported subsequences. Numbers in parenthesis indicate cluster assignment of the sequenced items, following the part.6 solution (above).

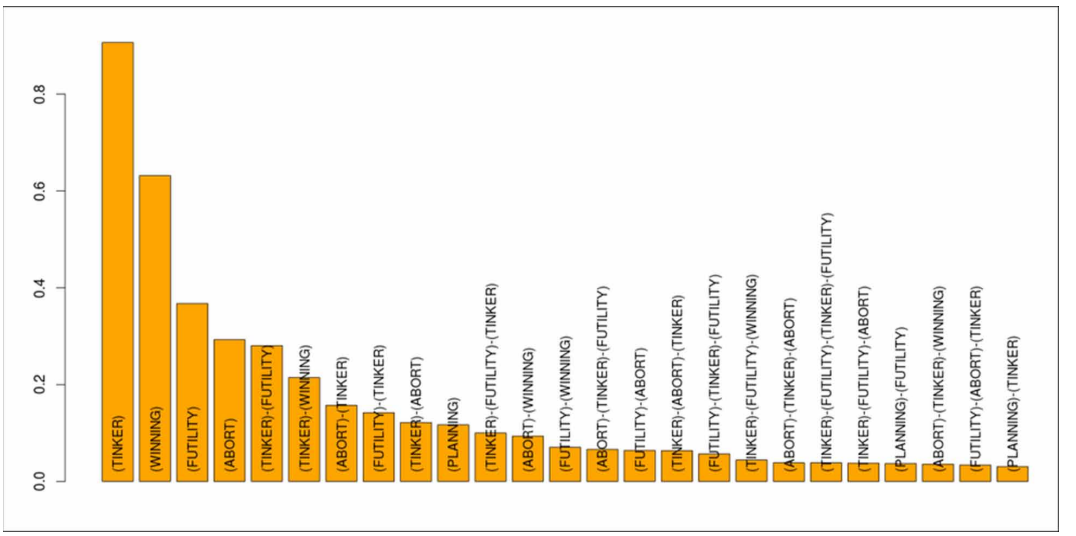


Table 4. Classification of Students by relative pre-post gains

	Post-test score (quartile)		
Pre-test score (quartile)	**1st**	**2nd – 3rd**	**4th**
1st	Low Prior	Learned	Learned
2nd – 3rd	Null	Null	Learned
4th	Null	Null	High Prior

These assignments were used as discriminant groups so that each detected subsequence's support could be tested for correlation with learning outcomes via a Chi-square test. Table 5 contains the 18 subsequences with the highest Chi-square statistic. Support for these subsequences thus varies by discriminant group in a statistically significant way. The graph of the resulting support values for each subsequence according to the student classification group is provided in Figure 12 (below).

In Figure 12, red bars indicate subsequences with significantly less support than under the assumption of independence. Conversely, blue-colored bars indicate significantly more support. Sequences in white show no statistical significance across all four groups. These significances are computed at the 0.01 level; light-blue and light-red bars indicate significance at the $p = 0.05$ level. For significance testing, the p-values were Bonferroni-corrected for the multiple comparison. This correction increases the probability of false negatives is compared to the probability of false positives but protects against incorrectly rejecting the null hypothesis, i.e., that the support values for the subsequences do not vary across discriminant groups.

This group-discriminant sequencing analysis suggests that students in the High Prior knowledge group have sharply fewer FUTILITY subsequences, fewer TINKER-FUTILITY and FUTILITY-TINKER, ABORT-TINKER, TINKER-FUTILITY-TINKER, and FUTILITY-TINKER-FUTILITY cycles, and substantially more PLANNING chains. Conversely, students with Low Prior knowledge are more likely to present longer FUTILITY chains, and more TINKER-FUTILITY cycles. These students are also more likely to follow FUTILITY with ABORT, ostensibly because they recognize

Table 5. Sequence analysis by discriminant group

			Support by Group			
Subsequence	***p***	**Chi-Sq**	**High Prior**	**Learned**	**Low Prior**	**Null**
(WINNING)	0.0000	115.71	0.42	0.60	0.81	0.69
(FUTILITY)	0.0000	113.28	0.17	0.32	0.54	0.43
(TINKER) – (FUTILITY)	0.0000	98.56	0.11	0.24	0.45	0.34
(FUTILITY) – (TINKER)	0.0000	33.92	0.06	0.13	0.22	0.17
(TINKER) – (FUTILITY) – (TINKER)	0.0000	32.16	0.03	0.09	0.15	0.12
(PLANNING) – (WINNING)	0.0001	29.05	0.04	0.01	0.00	0.01
(PLANNING)	0.0003	26.88	0.19	0.12	0.08	0.10
(FUTILITY) – (TINKER) – (FUTILITY)	0.0006	25.28	0.01	0.05	0.10	0.07
(FUTILITY) – (ABORT) – (TINKER)	0.0006	25.28	0.02	0.04	0.10	0.04
(ABORT) – (TINKER)	0.0013	23.77	0.09	0.15	0.23	0.17
(ABORT) – (TINKER) – (FUTILITY)	0.0021	22.78	0.02	0.06	0.12	0.08
(FUTILITY) – (ABORT)	0.0051	20.89	0.04	0.06	0.14	0.07
(ABORT)	0.0178	18.25	0.22	0.28	0.36	0.32
(TINKER) – (ABORT) – (TINKER) – (FUTILITY)	0.0560	15.79	0.01	0.02	0.06	0.03
(TINKER) – (WINNING)	0.0983	14.55	0.15	0.23	0.28	0.22
(LONG ABORT)	0.1368	13.80	0.03	0.02	0.01	0.01
(FUTILITY) – (WINNING)	0.1510	13.57	0.03	0.07	0.08	0.09
(TINKER)	0.1528	13.54	0.86	0.90	0.94	0.92

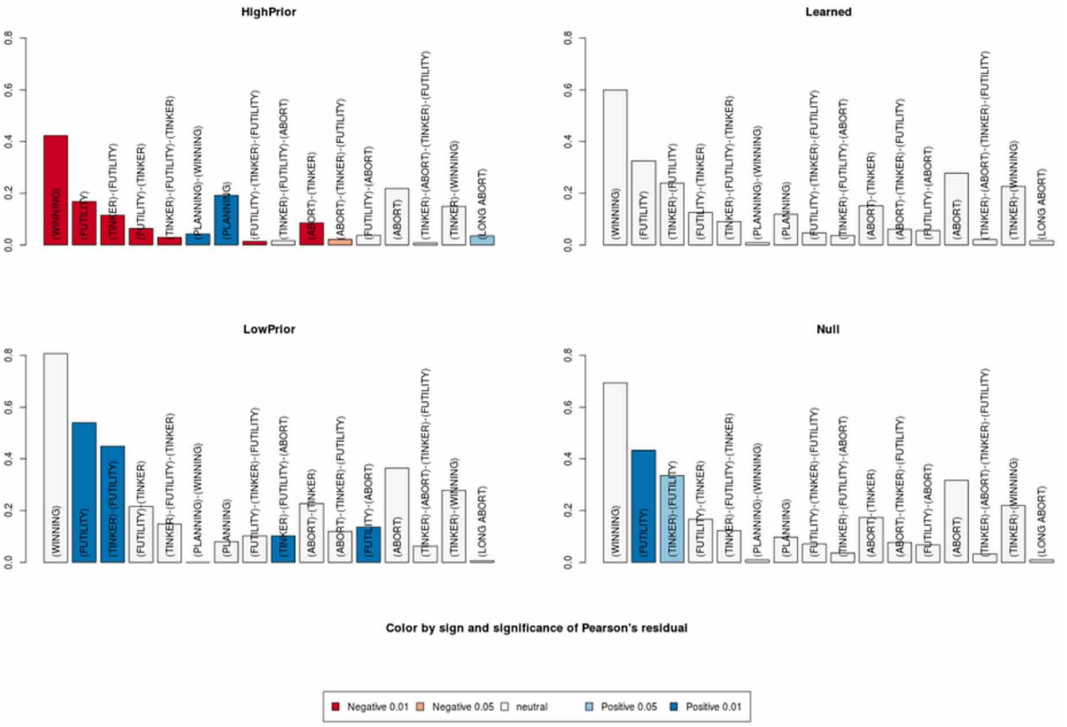
the probable outcome of that attempt would also have been FUTILITY. Students in the middle two quartiles who do demonstrate a relative increase in their conceptual understanding also show more FUTILITY chains and slightly more TINKER-FUTILITY chains.

RQ1 Discussion

The sequence analysis reveals that students with High or Low Prior knowledge play very differently than their peers. Students who have High Prior knowledge plan more and exhibit very few sequences of attempts in which they are stuck. They are not as likely to attempt small iterative fixes, preferring more complex and thought-out solutions. On the other hand, if students consistently demonstrate repeated failure on the same obstacle over large numbers of attempts, such as in a FUTILITY sequence, it is less likely that they improved their learning, regardless of their level of prior knowledge. No particular way of playing, or subsequence exhibited by students in the *Learned* group, seems to correlate with relative learning gains independently of prior knowledge.

This finding suggests that students' gameplay choices are strongly influenced by their prior knowledge. It may be fairly argued that *High Prior* knowledge students played a very *different game* than their *Low Prior* peers. The former group approaches the game as a "planning game," preferring the creation of complete solutions that require only small adjustments, making full use of the Solve-and-Debug strategy hypothesized in the Student Model. The latter group likely sees the game as a "guess and check game" or "tweaking game", where a solution emerges gradually out of extended iterating cycles of more-or-less purposeful trial-and-error, described earlier as an Additive-Iterative strategy.

Figure 12. Sequencing analysis by group



Why are students in the Low Prior group more likely to use the Additive-Iterative strategy? From the 2SM perspective, these students could be said to prefer low-effort, low-information, control-oriented processing strategies. The 2SM conceptualizes these as being closer to the Player Stance, which privileges feedback from the game environment to evaluate success. Students who play in the Additive-Iterative mode are more reliant on feedback from the game, since such feedback, rather than evaluation of internalized models, represents their main source of information about how the game operates. On the other hand, students who play the "planning game" can rely more on their own ability to visualize and predict how the game will respond to their input, and thus probably require less feedback from "tweaking" or "guessing and checking." This distinction correlates well with the general descriptions in the Two-System Framework of the Player Stance and Learner Stance, respectively.

RESULTS RQ2: GAME/TEST PERFORMANCE RELATIONSHIPS

The main learning goal of EPIGAME is to help students build a deeper understanding of Newtonian kinematics. Thus, the game's rules and systems deal with inertia and the relationship between force and velocity. Ideally, as students improve their ability to solve inertial challenges, their conceptual understanding, per an external measure, should likewise improve. From the previous analysis (see Question 1 section, above), we know that students with different degrees of prior knowledge approach the game differently and play in sharply different ways. In terms of Question 2, we investigated whether these differences in performance on the tests correlate with differences in gameplay in the specific game situations intended to help students develop concepts of inertia.

The first step in this analysis was coding the conceptual challenges. Each challenge is a situation on the game map where a student has to apply one or two maneuvers to advance past that situation.

The selected challenges all deal with *inertia* and/or *Newton's second law of motion.* These concepts can be portrayed in EPIGAME in one of four ways:

1. The student must navigate Surge from rest up to a certain velocity by applying an unbalanced force (Figure 13). There are 46 such challenges, and they were coded as *fromStop.*
2. The student must bring Surge from a constant velocity to a stop by applying one or more forces opposed to the direction of motion (35 challenges, coded as *toStop).* (Figure 14)
3. The student must increase the velocity of Surge to a certain level while Surge is in motion by applying an unbalanced force in the direction of motion (4 challenges, coded as *speedUp).* (Figure 15)
4. The student must decrease the velocity of Surge to a certain level while Surge is in motion by applying an unbalanced force opposite the direction of motion (5 challenges, coded as *slowDown).* (Figure 15)

Each challenge was identified through visual inspection of the levels, its location and type recorded (*fromStop, toStop, speedUp,* or *slowDown*), and a consecutive serial number assigned. Only the first 90 challenges students encounter while playing EPIGAME were coded. The rationale for this limit is that the conceptual nature of these challenges changes in the latter levels, first when changes in mass are introduced, and then when students have to deal with forces applied in action-reaction pairs. Thus, the first 90 challenges students encounter before these increases in complexity are the most conceptually similar and can be safely compared. Furthermore, these first 90 challenges are where we might be most likely to see trajectories of improvement because it tracks students from the beginning of the game where the learning curve may prove the clearest.

Figure 13. A fromStop challenge. Students begin motion from rest at point B and navigate toward C.

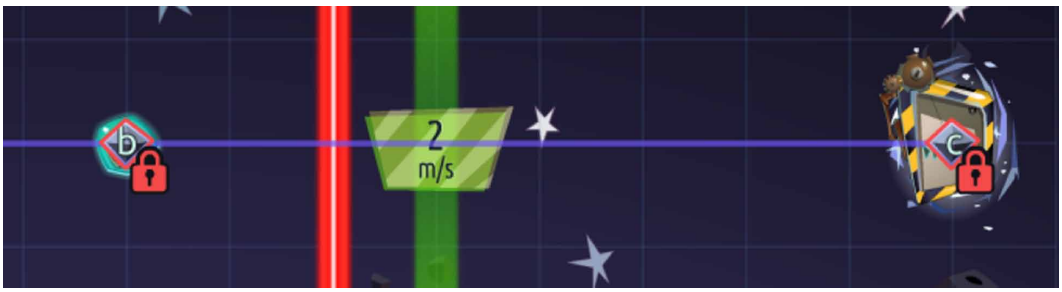


Figure 14. A toStop challenge. Students must completely stop at B before proceeding to C.

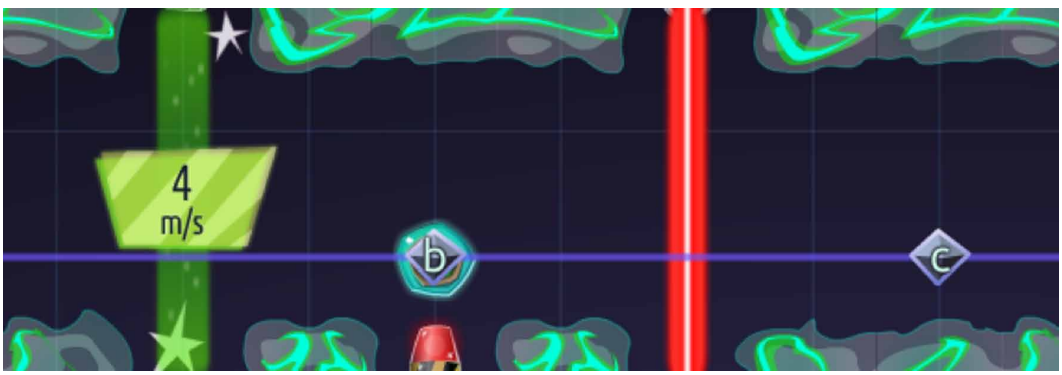
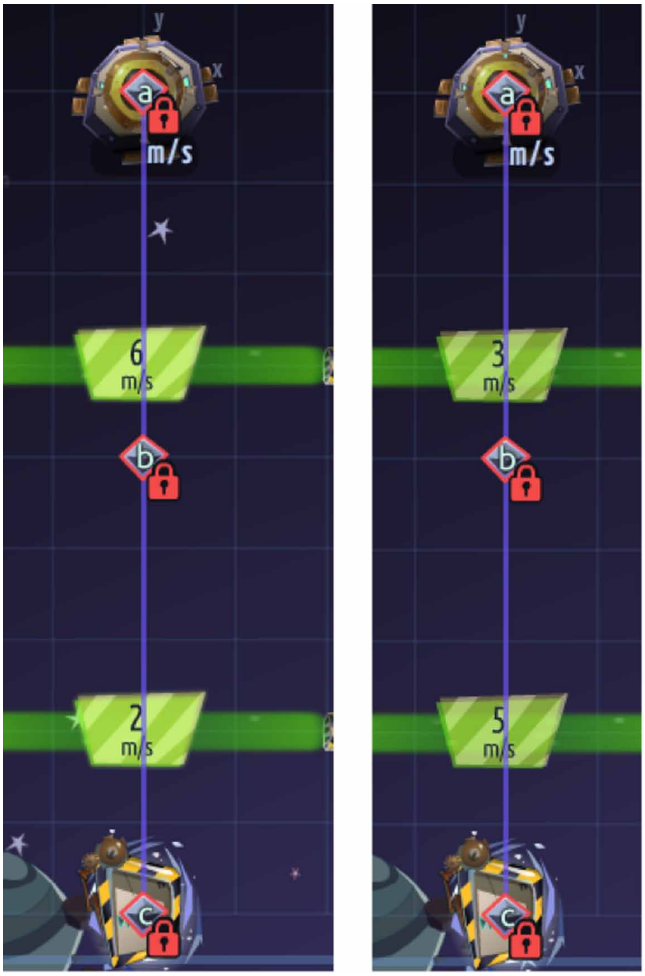


Figure 15. A slowDown (left) and a speedUp challenge (right). In both cases, the student must apply an unbalanced force at B.



To analyze these first 90 challenges, the overall gameplay dataset was filtered through a conditional join in order to identify which attempts ended at one of the coded challenges. A total of 2175 attempts were identified. Later, we decided to reduce the sample to 1282 attempts corresponding to the first 15 challenges of each type, under the rationale that the unbalanced number of challenges per type (e.g. 46 *fromStop* vs. 4 *slowDown*) would likely lead to problems with the model fit if we used the challenge type as a covariate.

Next, we proceeded to fit a generalized linear model to the data. Since the dependent variable is a count comprised of positive whole numbers only, a Poisson regression would be most appropriate. However, the data showed considerable overdispersion, and thus a negative binomial regression was chosen.

Generalized Linear Model

The statistics of the generalized linear model are provided in Table 6. In this model, the High Prior classification and *fromStop* challenge type are the model references. The statistically significant predictors of student errors per Conceptual challenge are Challenge instance, and as noted, the type of challenge is not a statistically significant predictor. Furthermore, a previous iteration of the model

Table 6. Coefficients of the negative binomial regression model

	Dependent variable: number of errors per Challenge		
	Estimate	**Std. Error**	***p*-value**
Challenge instance	-0.064***	0.008	>0.001***
Learned	0.484***	0.148	0.001***
Low Prior	0.744***	0.153	>0.001***
Null	0.654***	0.140	>0.001***
slowDown Challenge	0.112	0.090	0,21
speedUp Challenge	0.008	0.093	0.93
toStop Challenge	0.040	0.067	0.55
Constant	1.117***	0.150	>0.001***
Observations	1,282		
Log Likelihood	-3,221.669		
theta	1.371***	0.066	
Akaike Inf. Crit.	6,459.338		

Note: ***p<0.01

showed that the interaction terms of the predictors were also not statistically significant. Thus, the variables best suited to predict the number of errors students commit are the number of similar challenges already faced and the students' prior knowledge.

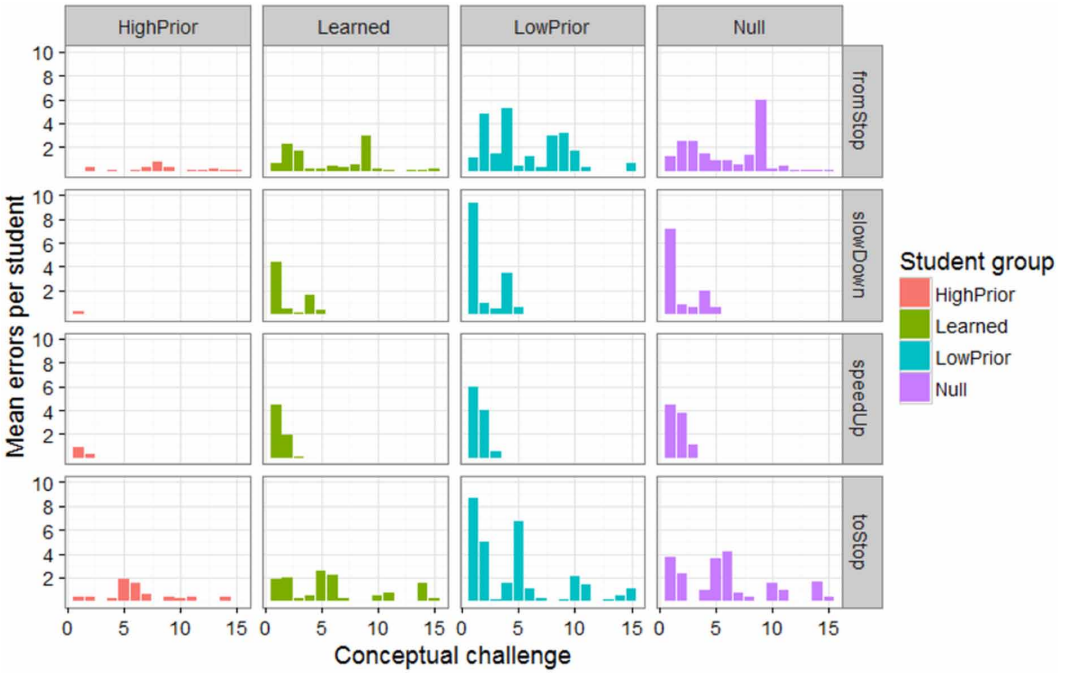
RQ2 Discussion

The generalized linear model fit to the challenge data confirms the hypothesis that students tend to make fewer errors on a challenge each successive time they encounter a challenge of the same type. More surprising is that the *mean* number of errors can also be predicted on the basis of a student's prior knowledge grouping. In other words, the first 15 times students face challenges of a given type, students who score highly on the pre-test are likely to commit as few as half as many errors as students who did not score highly.

A possible explanation is suggested by the bar chart matrix on Figure 16. We can see there that students in the High Prior column make fewer errors overall, but more importantly, commit nearly no errors the first time they face a challenge of a given type. Students in other groups commit at least 3 errors on average and often more. Unless High Prior students have played EPIGAME before, which they have not, one could assume that High Prior students would make at least a few errors when they initially encounter a challenge, while they internally navigate how their understanding of physics does or does not apply to the situations and rules of the game. However, the near-total absence of errors on initial contact with challenge types suggests that High Prior students *already know* something directly relevant to these challenges.

There are at least two other sources of knowledge besides any prior EPIGAME experience that students might be drawing on when they face new challenges. First, they may be drawing on inferences made from the pre-test. However, we demonstrated in Study 1 that EPIGAME and the EPIGAME assessment are free of testing effects (see Methods section), so a "priming" effect is unlikely. The other source might be the tutorial animations embedded in EPIGAME. There are two types of tutorials. At levels 1, 4, 8, 10 and 11, the tutorial animations are essentially *worked examples*. Students watch as the Mentor character demonstrates skills such as how to apply forces, how to draw Waypoints, how to start the trial. These animations are intended to guide students as they learn the game's

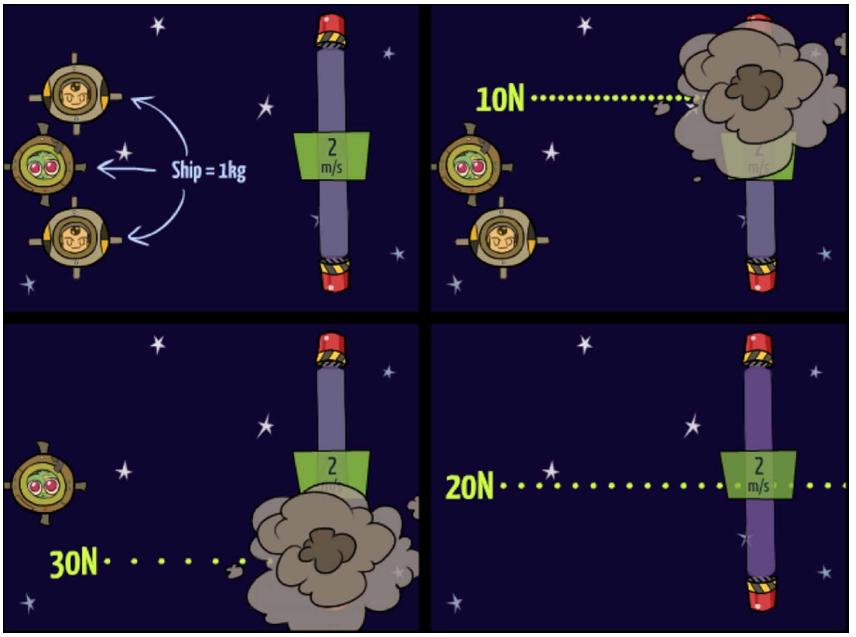
Figure 16. Mean errors per student per Conceptual challenge. Student achievement groups are in columns. Challenge types are in rows.



interfaces, design conventions, etc. On the other hand, the tutorials at levels 2 and 7 are *contrasting cases* (Figure 17). These animations take the form of experiments; a challenge is approached with several combinations of parameters, of which only one is correct. The student must deduce from this demonstration *why* that particular maneuver was effective. While the "worked example" tutorials show the *hows* of EPIGAME, the "contrasting case" tutorials show the *whys.*

Our explanation for the low rate of error of High Prior knowledge students during initial trials relative to their lower-prior-knowledge peers is grounded in the 2SM. The 2SM defines two broad classes of knowledge regarding "how to play": heuristics and internal models. The "worked example" tutorials, with their emphasis on how to execute specific maneuvers, have more "heuristicness" than "modelness". Conversely, the "contrasting cases" tutorial focus strongly on the variables and relationships at play, suggesting more model quality. It may be that the main difference between High Prior students and their peers is *which form* of the tutorial they chose to focus on. Since each form of tutorial primes a different form of knowledge about "how to play", students with a strong preference for one form of tutorial over the other may approach the game with different kinds of knowledge and thus play in different ways. And in fact, these differing styles of play do emerge (see Question 1), with High Prior students showing a marked preference for slow, deliberate play and a small tolerance for error. In contrast, students in the Low Prior and Null learning groups prefer iterative, "tweaking" gameplay that is inherently more fast-paced, yet they tend to accrue errors at each challenge, often as many as 10, 20 or more (see Figure 16). In summary, it may be that the tutorials, necessary parts of the game experience, can "prime" the 2SM stances according to (a) the forms the tutorials take (prescriptive vs. descriptive) and (b) how salient and useful the student finds the information presented in the tutorials themselves.

Figure 17. A "contrasting case" tutorial. The use of 10N and 30N are both incorrect for a 2m/s Velocity Gate.



OVERALL CONCLUSIONS

This study was designed to explore to primary questions. First, can the two stances of the 2sm, as specified by the framework, be detected in game play data? Second, how do changes in students' functional understanding of the game relate to performance on a test of conceptual understanding? Related to these two questions, a third implicit question explored the viability of our approaches to data-mining and learning analytics of game-play data to explore the two explicit research questions. In the following sections, we analyze the two research questions in terms of the overarching implications for the 2SM. Next, we consider the overarching implication for the design of digital games to support conceptual learning. Finally, we close with an examination of the third implicit questions about the viability and generalizability of our approaches to data-mining and learning analytics of game-play data to explore theoretical questions about learning.

What the Findings Say about the 2SM

The 2SM is intended as a general-purpose framework for student cognition during gameplay; it is comprehensive and not intended to be specific to any kind of game or any target domain. Because of this generality, it requires many constructs and mechanisms to explain phenomena of play. Furthermore, most of these constructs and mechanisms are entirely latent, existing only in the student's mind and perhaps only for brief moments of time. For these reasons, it is unlikely that a single study, however ambitious, could prove the 2SM as a theory.

The findings in this paper suggest that the basic underpinnings of the 2SM pass muster in terms of the existence of the two stances, but the findings do not support the hypothesized swapping back and forth between stances. We see the indicia of both fast, low-information play and slow, deliberative play. More importantly, these styles of play co-vary strongly with learning outcomes, indicating that fast styles of play may not support students in developing knowledge of a form transferable beyond to the game, in this case, focusing on Newtonian kinematics. Some students perseverate, however, in guess-and-check iteration, relying entirely on the game to provide the necessary feedback, instead

of using all available information to infer some generalizable rule they can use to increase their effectiveness. We can see from the Contextual Mapping analysis that some students never seem to stop making errors in parts of the game relating to a specific concept, even when they've already cleared a similar challenge 10 times or more. Essentially, these students never switch to a more model-oriented thoughtful stance. Conversely, some students begin in a more thoughtful stance and remain in that stance throughout play.

Related to this absence of stance switching, the finding that prior knowledge strongly influences play, even in the early stages, is problematic in terms of the 2SM. First, because it inverses the proposed way that Stances get cued. In the original framing of the 2SM, the Learner Stance is cued by a task that is too demanding, where the student has no fast effortless rule to apply. However, the results in this paper strongly support the claim that the opposite may be true, or that perceived high task demands cue the Player stance as an effort-saving strategy that is ultimately maladaptive in terms of learning. The second challenge to the 2SM comes from the necessity of having students "learn to play" the game before they actually "play" it. This instructional phase and its consequences were not addressed originally in the 2SM. Yet as we have discussed previously, the Tutorial materials and other instructional affordances might bias students towards one form of reasoning or another, independently of how the student would otherwise organize his or her epistemic Stance.

These findings suggest that revisions of the 2SM are warranted in at least two lines. On one hand, (1) the role of prior knowledge as an epistemic resource, largely ignored in the original framing. The 2SM envisions a student with well-defined goals for play but a "blank slate" in terms of pre-existing knowledge about the game. Further research that specifically targets the effect of prior knowledge, and of knowledge gleaned early in play from tutorial materials is warranted, and those findings integrated into the 2SM. Another possible revision involves (2) the issue of task demands and their possible role in cueing other resources, such as mastery or performance orientations (Pintrich, 2000). The 2SM does not explicitly consider whether a student finds a given game situation "easy" or "difficult"; rather it only considers what epistemic resources the student has at hand, such as heuristics and second-order models. Yet the findings in this study highlight that Player Stance related patterns of play may also be a coping strategy to deal with game situations students find too difficult. For the 2SM to properly account for these coping strategies, a study might be designed where versions of the game of various difficulty levels are assigned to students at different levels of achievement, either by pre-test score or by an automated adaptive functionality.

Implications for Game Design

The results of the gameplay data analysis from RQ1 generally support the notion that patterns of play related to the Player Stance are not optimal for learning. Students who persist in fast strategies are not likely to improve their learning relative to their peers, and students who make the highest relative gains do not prefer fast strategies overall. The analysis shows that a tolerance or preference for Attempt sequences with a high reliance on FUTILITY are associated with lower learning outcomes.

In the 2SM framework, multiple FUTILITY attempts with low average time per attempt can be understood as a strategy for obtaining feedback from the game's model as a way to avoid having to use slower, more intensive reasoning processes such as the second-order model. The goal of this strategy is to serve the student's agency and sense of control and preserve the momentum of play. It may be argued that use of the Player Stance helps students remain motivated and engaged even in the face of failure, and long after the novelty of the game has worn off. Yet, as we have seen, in the case of EPIGAME, the Player Stance and its associated play strategies are associated with lower learning gains. Then, the immediate question becomes, can the Player Stance be disrupted in order to promote learning? Or in the context of EPIGAME, can a student playing the "tweaking game" be nudged towards playing the game more as a "planning game?" Can a game be designed in such a way that this "nudge" occurs automatically?

In the case of EPIGAME, the tutorials might provide a clue as to how this "nudge" can occur early in play (see Question 2). Whatever the eventual form that this encouragement takes, the effectiveness of this feature depends on having a method to detect whether or not a student has settled in a Player Stance. This "detector" could be built upon the analysis here described: the game could use a similar process of unsupervised clustering we used to arrive at the *part.6* solution as a guide to classify students' actions in real time, and then detect the sequences of play which, as we have seen, are not strongly associated with learning. This added functionality would allow specific feedback to be provided to students early in their play before they commit to playing a "tweaking game" (c.f. Clark, Martinez-Garza, Biswas, Leucht, & Sengupta, 2012). Lastly, if the game can be made so that, once it has gathered enough student data to predict a student's play characteristics, the game can modulate its difficulty to make sure that the student faces challenges appropriate to the student's level of skill and knowledge, while compensating for the tendency of students to choose low-effort strategies if doing so preserves the momentum of play. These three additional functionalities could all be potentially very powerful ways to promote student learning with games, and they are all made possible by an expanded understanding of how students actually play.

These findings should also motivate discussion about how much and what kinds of support students should receive during game-based learning opportunities. Lower-performing students' over-reliance on fast strategies might be more of an adaptive response to being forced to play a game that is too difficult as opposed to an intentional strategy choice in response to their perceptions of what the game is about. In this case, automated feedback and adaptation as discussed above would also be useful. Students who are facing intractable difficulty could be detected and helped automatically. It is also possible that students perseverating in fast strategies are doing so transgressively (see Aarseth, 2007) as a rejection of the game's challenge and a personal disinvestment from the game's outcomes. This low-effort position is radically different from the low-ability position described above, but in terms of data logs, it would look rather similar. The analytics used in this study are not well-suited to detect the difference between low effort and low ability, although some scholars have had success with specific detection algorithms for disengagement in the context of science simulations (Gobert, Baker, & Wixon, 2015).

In this study, as in much of classroom-based educational game research, we relied on pre-existing classroom norms for expectations on student behavior and effort. Also, the presence and expert eye of the teacher to help identify and gently correct students who were off-task and offer guidance to those few students who may have found the game too demanding were indispensable. We observed and respected these practices while fully knowing that their effects would disturb the central assumption that the data logs record students' actions *and only students' actions*. This tension points to an inherent limitation of the data logging approach. Data logging can only account for what happens within the student-computer interaction, and classroom technology use often involves, or even privileges, person-to-person interactions. It is during these kinds of interactions that teachers and often peers help students make sense of the game when the game itself doesn't offer the necessary scaffolds, whether motivational or content-related. These interactions may have effects on participating students' play that would be captured by data logging but would be difficult for LA techniques to correctly explain or attribute. It may be that future work that harnesses data log analytics for adaptive feedback might approach, or perhaps even duplicate, the classroom teacher's ability to identify apathy and helplessness in the classroom context, or the knowledgeable peer willingness to dispense timely hints. Until that time, however, we accept some imperfection and "mangling" of the record and look for opportunities to more deeply integrate log-based analytics with observational and grounded methods.

Regardless, in order to access the potential and intended benefits of an educational game, students must first *learn to play* the game itself. This step, while commonsensical, can easily be glossed over during design; when it comes to introducing unfamiliar digital games into the classroom, we might hold the notion that young students can simply "pick it up" and "figure it out", since they may already be "gamers". Thus, materials intended to help students orient themselves in the game environment

and learn how to reach gameplay goals may not receive as much design attention as they otherwise would. Furthermore, when games are used in an educational setting, these materials compete for classroom time with the main game, where the target curricular material is most likely to reside. Ideally, we would prefer if students spend only a little time "learning to play" and as much time as possible simply "learning."

Our findings problematize these design assumptions. First, as shown, prior knowledge can structure gameplay to a great extent (see RQ1 analyses). Students who enter the game experience with a good working knowledge of the concepts and relationships are less reliant on more feedback-rich and iterative, yet ultimately more laborious, "tweaking" styles of play. Second, the analysis also suggests that the way the game teaches students to play, by following a procedure or operationalizing a relationship, may also be an important influence on students, even when this learning is focused squarely on game-specific knowledge and not on curricular concepts and relationships.

If prior knowledge and differential use of tutorial materials can structure and influence play and, thus, learning, then a greater emphasis must be placed on game functionality that supports students who do not initially enjoy or leverage these advantages. For example, lack of prior knowledge can be addressed with *scaffolding*, and gameplay difficulty can be adapted to reduce repeated error. These and other measures should be considered as means to ensure that all students can access substantially similar game experiences and thus, hopefully, more equitable positive learning outcomes.

Data-Mining and Learning Analytics as a Tool

The work described in this paper has followed a methodology that is not limited to investigating EPIGAME logs. The general methodology is versatile and feasible for use in other contexts. Starting from a robust and detailed record of students' interactions with a digital environment and a theoretical framework that supports conjectures as to why certain patterns of action create opportunities for the desired change, researchers can define the important features of those patterns and then use those features to investigate the data record using whatever LA techniques are most appropriate for that particular type of data.

A more novel focus of this analysis, which is highlighted in Research Question 2, is that it aims to track the development of students' conceptual understanding at the level of particular concepts of inertia using finer-grained observations centered on particular gameplay regions. These regions are intended to highlight specific content, and thus student performance in these regions is more closely tied to conceptual understanding than gross-level summative measures. These summative measures have been successfully used in the past and may be appropriate and sufficient for some research questions. However, the use of finer-grained contextual data offers the advantage of supporting claims of students' conceptual understanding of individual concepts such as inertia or First Law rather than broad performance constructs, such as knowing how to play EPIGAME).

This is not to say that the EPIGAME data structure and focus and the associated analyses are universal. The trial-retrial structure of gameplay and the grain size of the data capture are not necessarily common to all educational games. The specific combination of play structure and grain size warranted the sequence mining and contextual feature mapping. Other digital environments will have different interactive structures, and thus algorithms and techniques possibly better suited to the questions being asked. Fortunately, the state of the art of learning analytics is increasing both the accessibility and variety of statistical computing software, making it suitable for a wider variety of data structures, game mechanics, and learning foci.

One thing that will likely remain invariant, however, is the expertise of the analyst and his or her familiarity with the context and the data. In this paper, our own long association with EPIGAME data and our observations accumulated over multiple opportunities to facilitate students' play of EPIGAME facilitated the creation of the derived variables, the process of interpreting the *part.6*

clustering, and the use of sequencing as a way to add meaning to the cluster assignments. It is unlikely that this kind of intimate understanding of the affordances and constraints of particular games and data can be substituted by generic software although it can, perhaps, be supplemented. Until that time, however, the skill of the analyst, as in all interpretative observational methods of research, will be crucial to success.

ACKNOWLEDGMENT

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, and the National Science Foundation through grants R305A110782 and 1119290 to Vanderbilt University. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the National Science Foundation.

REFERENCES

Aarseth, E. (2007). I fought the law: Transgressive play and the implied player. *Situated Play. Proc. DiGRA*, 24-28.

Annetta, L. A. (2010). The "I's" have it: A framework for serious educational game design. *Review of General Psychology*, *14*(2), 105–112. doi:10.1037/a0018985

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M.-T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers & Education*, *53*(1), 74–85. doi:10.1016/j.compedu.2008.12.020

Baker, R. S., & Clarke-Midura, J. (2013). Predicting successful inquiry learning in a virtual performance assessment for science. In S. Carberry, S. Weibelzahl, A. Micarelli, & G. Semeraro (Eds.), *User Modeling, Adaptation, and Personalization* (pp. 203–214). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-38844-6_17 doi:10.1007/978-3-642-38844-6_17

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology*. *Knowledge and Learning*, *19*(1-2), 205–220. doi:10.1007/s10758-014-9223-7

Braver, M. W., & Braver, S. L. (1988). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, *104*(1), 150–154. doi:10.1037/0033-2909.104.1.150

Buckley, B. C., Gobert, J. D., Kindfield, A. C. H., Horwitz, P., Tinker, R. F., Gerlits, B., & Willett, J. et al. (2004). Model-based teaching and learning with Biologica: What do they learn? How do they learn? How do we know? *Journal of Science Education and Technology*, *13*(1), 23–41. doi:10.1023/B:JOST.0000019636.06814.e3

Chesler, N. C., Ruis, A. R., Collier, W., Swiecki, Z., Arastoopour, G., & Shaffer, D. W. (2015). A novel paradigm for engineering education: Virtual internships with individualized mentoring and assessment of engineering thinking. *Journal of Biomechanical Engineering*, *137*(2), 024701. doi:10.1115/1.4029235 PMID:25425046

Clark, D. B. (2012). Designing Games to Help Players Articulate Productive Mental Models. *Keynote commissioned for the Cyberlearning Research Summit 2012*, Washington, DC. Retrieved from http://www.youtu.be/xlMfk5rP9yI

Clark, D. B., & Martinez-Garza, M. (2012). Prediction and explanation as design mechanics in conceptually-integrated digital games to help players articulate the tacit understandings they build through gameplay. In C. Steinkuehler, K. Squire, & S. Barab (Eds.), *Games, learning, and society: Learning and meaning in the digital age*. Cambridge, Mass: Cambridge University Press. doi:10.1017/CBO9781139031127.023

Clark, D. B., Nelson, B. C., Chang, H.-Y., Martinez-Garza, M. M., Slack, K., & D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers & Education*, *57*(3), 2178–2195. doi:10.1016/j.compedu.2011.05.007

Clark, D.B., Sengupta, P., Brady, C., Martinez-Garza, M., & Killingsworth, S. (2015). Disciplinary Integration in Digital Games for Science Learning. *International STEM Education Journal, 2*(2), 1-21. doi: Retrieved from http://www.stemeducationjournal.com/content/pdf/s40594-014-0014-4.pdf10.1186/s40594-014-0014-4

Clarke-Midura, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research on Technology in Education*, *42*(3), 309–328. doi:10.1080/15391523.2010.10782553

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. doi:10.1007/BF01099821

Csikszentmihalyi, M. (1991). *Flow: The Psychology of Optimal Experience* (1st ed.). Harper Perennial.

Dede, C. (2011). Developing a research agenda for educational games and simulations. In S. Tobias & J. D. Fletcher (Eds.), *Computer Games and Instruction* (pp. 233–247). Charlotte, NC: Information Age Publishing.

Dondlinger, M. J. (2007). Educational video game design: A review of the literature. *Journal of applied educational technology, 4*(1), 21-31.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., & Wickham, H. (2012). The Generalized Pairs Plot. *Journal of Computational and Graphical Statistics*, *22*(1), 79–91. doi:10.1080/10618600.2012.694762

Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. doi:10.1146/annurev.psych.59.103006.093629 PMID:18154502

Foster, A., & Mishra, P. (2008). Games, Claims, Genres and Learning. In R. E. Ferdig (Ed.), *Handbook of research on effective electronic gaming in education* (pp. 33–50). Hershey, PA: IGI Global. doi:10.4018/978-1-59904-808-6.ch002

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976. doi:10.1126/science.1136800 PMID:17218491

Gabadinho, A., Ritschard, G., Müller, N. S., & Studer, M. (2011). Analyzing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, *40*(4), 1–37. Retrieved from http://www.jstatsoft.org/v40/i04/ doi:10.18637/jss.v040.i04

Gee, J. P. (2007). What Video Games Have to Teach Us About Learning and Literacy (2nd ed.: rev. and Updated). New York, NY: Palgrave Macmillan.

Geertz, C. (1973). Thick Description: Towards an Interpretive Theory of Culture. In The Interpretation of Cultures (pp. 3–30). Basic Books.

Gijlers, H., & de Jong, T. (2013). Using Concept Maps to Facilitate Collaborative Simulation-Based Inquiry Learning. *Journal of the Learning Sciences*, *22*(3), 340–374. doi:10.1080/10508406.2012.748664

Gobert, J. D., Sao Pedro, M. A., & Baker, R.S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, *4*(1), 111–143.

Halverson, R., & Owen, V. E. (2014). Game-based assessment: an integrated model for capturing evidence of learning in play. *International Journal of Learning Technology*. Retrieved from http://www.inderscienceonline.com/doi/abs/10.1504/IJLT.2014.064489

Hammer, D., & Elby, A. (2003). Tapping Epistemological Resources for Learning Physics. *Journal of the Learning Sciences*, *12*(1), 53–90. doi:10.1207/S15327809JLS1201_3

Harpstead, E., Myers, B. A., & Aleven, V. (2013). In search of learning: facilitating data analysis in educational games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 79–88). New York, NY: ACM. doi:10.1145/2470654.2470667

Hou, H. T. (2012). Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG). *Computers & Education*, *58*(4), 1225–1233. doi:10.1016/j.compedu.2011.11.015

Husson, F., Josse, J., Le, S., & Mazet, J. (2015). *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining*. Retrieved from http://CRAN.R-project.org/package=FactoMineR

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, *93*(5), 1449–1475. doi:10.1257/000282803322655392

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining*, *4*(1), 144–182.

Lamb, R. L., Annetta, L., Vallett, D. B., & Sadler, T. D. (2014). Cognitive diagnostic like approaches using neural-network analysis of serious educational videogames. *Computers & Education*, *70*, 92–104. doi:10.1016/j.compedu.2013.08.008

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3–16. doi:10.1111/j.1745-3992.2007.00090.x

Linehan, C., Kirman, B., Lawson, S., & Chan, G. (2011). Practical, Appropriate, Empirically-validated Guidelines for Designing Educational Games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1979–1988). New York, NY: ACM. doi:10.1145/1978942.1979229

Martinez-Garza, M., Clark, D. B., & Nelson, B. C. (2013). Digital games and the US National Research Council's science proficiency goals. *Studies in Science Education*, *49*(2), 170–208. doi:10.1080/03057267.2013.839372

Martinez-Garza, M. M., Clark, D., & Nelson, B. (2013). Advances in Assessment of Students' Intuitive Understanding of Physics through Gameplay Data. *International Journal of Gaming and Computer-Mediated Simulations*, *5*(4), 1–16. doi:10.4018/ijgcms.2013100101

Martinez-Garza, M. M., & Clark, D. B. (2016). Two systems, two stances: a novel theoretical framework for model-based learning in digital games. In P. Wouters & H. van Oostendorp (Eds.), *Instructional Techniques to Facilitate Learning and Motivation of Serious Games* (pp. 37–58). Switzerland: Springer; doi:10.1007/978-3-319-39298-1_3

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29. doi:10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., & Corrigan, S. … John, M. (2014). *Psychometric considerations in game-based assessment.* (white paper). Retrieved from http://www.instituteofplay.org/work/projects/glasslab-research/

Parnafes, O., & Disessa, A. (2004). Relations between Types of Reasoning and Computational Representations. *International Journal of Computers for Mathematical Learning*, *9*(3), 251–280. doi:10.1007/s10758-004-3794-7

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*(3), 544–555. doi:10.1037/0022-0663.92.3.544

Plass, J., Homer, B., & Kinzer, C. (2014) *Playful Learning: An Integrated Design Framework.* (White paper #02/2014). Games for Learning Institute. doi:10.13140/2.1.4175.6969

Prensky, M. (2006). Don't bother me Mom - I'm learning!: how computer and video games are preparing your kids for twenty-first century success - and how you can help! Paragon House. Retrieved from http://www.worldcat.org/isbn/1557788588

Reyna, V. F., & Ellis, S. C. (1994). Fuzzy-trace theory and framing effects in children's risky decision making. *Psychological Science*, *5*(5), 275–279. doi:10.1111/j.1467-9280.1994.tb00625.x

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. doi:10.1007/BF02300540

Rowe, E., Asbell-Clarke, J., & Baker, R. S. (2015). Serious games analytics to measure implicit science learning. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics* (pp. 343–360). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-05834-4_15 doi:10.1007/978-3-319-05834-4_15

Sengupta, P., Krinks, K. D., & Clark, D. B. (2015). Learning to Deflect: Conceptual Change in Physics during Digital Game Play. *Journal of the Learning Sciences*, *24*(4), 638–674. doi:10.1080/10508406.2015.1082912

Shaffer, D. W., Squire, K. D., Halverson, R., & Gee, J. P. (2005). Video games and the future of learning. *Phi Delta Kappan*, *87*(2), 104. doi:10.1177/003172170508700205

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. Cambridge, MA: MIT Press.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer Games and Instruction* (Vol. 55, pp. 503–524). Charlotte, NC: Information Age Publishing. Retrieved from http://pdf.thepdfportal.net/PDFFiles/6536.pdf

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22. doi:10.1037/0033-2909.119.1.3 PMID:8711015

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*(2), 137–150. doi:10.1037/h0062958 PMID:18116724

Squire, K. D., DeVane, B., & Durga, S. (2008). Designing centers of expertise for academic learning through video games. *Theory into Practice*, *47*(3), 240–251. doi:10.1080/00405840802153973

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Tobias, S., & Fletcher, J. D. (2007). What Research Has to Say about Designing Computer Games for Learning. *Educational Technology*, *47*(5), 20–29.

Vellido, A., Martin-Guerroro, J. D., & Lisboa, P. (2012). Making machine learning models interpretable. In *Proceedings of the 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN),* Bruges, Belgium (pp. 163–172).

Ventura, M., Shute, V., & Small, M. (2014). Assessing persistence in educational games. *Design Recommendations for Intelligent Tutoring Systems*, *93*.

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, *60*(1), 52–58. doi:10.1016/j.compedu.2012.07.003

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10). Retrieved from http://courses.had.co.nz.s3-website-us-east-1.amazonaws.com/12-rice-bdsi/slides/07-tidy-data.pdf doi:10.18637/jss.v059.i10

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., & Yukhymenko, M. et al. (2012). Our Princess Is in Another Castle: A Review of Trends in Serious Gaming for Education. *Review of Educational Research*, *82*(1), 61–89. doi:10.3102/0034654312436980

ENDNOTES

[1] The key insight that led to the development of the *par* metric was the observation that the transition probabilities described in the Markov-chain model (i.e., the probability of a certain outcome for an Attempt given the outcome of the previous Attempt) were closely related to both the spatial layout of each level and the number of times that students had encountered that level's challenges before. For example, in the EPIGAME level pictured in Figure 1, we would expect to see the transition probability of Velocity Gate to Navigation Error to be high, because the layout of the level places a 90-degree turn after a Velocity Gate, and both challenges are likely to cause player errors. Conversely, we would expect the transition from Navigation Error to Success Gate to be very low, since there is at least one obstacle between the last 90-degree turn and the Gate. Few students would be able to follow up a Navigation error with a clean transversal of the Laser and Velocity Gates. Furthermore, a given outcome can transition into *itself*, indicating an outcome that is very often repeated across students and chains of Attempts. Higher "self-transitions" probabilities are more common when new game elements are introduced, and students make repeated attempts to advance past the new challenges. Self-transitions generally become rarer, however, as students gain more proficiency in dealing with each challenge, i.e. from one level to the next.

APPENDIX A

Observed and Derived Variables in the Epigame Log Data

Table 7. Catalog of variables in the Study 2 dataset

Variable name	Meaning	Type	Notes
Student ID		Identification	Anonymized to a serial number
Experiment ID		Identification	
Date and Time		Identification	
Step Visit	Number of times student has visited that Level (step)	Observed	
Attempt		Observed	Only Attempt = 1 was used
attemptTrial	Order of this Attempt within a series of Attempts (i.e. a Trial)	Observed	
totalTrials	Combined number of Attempts in all Trials of this Level by this student	Observed	
endState	Did the player succeed (=1), fail (=0), or abort (=2)?	Observed	
endScore	Score obtained by that student at the end of that Trial	Observed	
scoreImproved	Did the student increase their Score this Attempt?	Derived	
trialTime	Length of time a between this Attempt and the end of the previous Attempt	Observed	Incorrectly named in software, should be “attemptTime”
actionsUsed	How many Actions were placed on Waypoints during the Planning Phase	Observed	
isExit	Did the student leave the Level after this Attempt?	Observed	
timeLine	Position of the time cursor on the Timeline at level end	Observed	More relevant in the Timeline version of EPIGAME. Students in the present studies did not have access to the time cursor.
attemptTrial.max	Maximum value of the variable attemptTrial for that student for that level	Derived	
p.attemptTrial	Measure of progress of Attempts within the Trial	Derived	Calculated as (attemptTrial / attemptTrial.max)
p.totalTrials	Measure of progress of Attempts within the combined chain of Attempts over all Trials	Derived	Calculated as [attemptTrial + (sum of all attemptTrial.max of all previous trials) / totalTrials]
ending.event	The state of the game that caused the Level to end.	Observed	Allowed states: Success Gate, Navigation Error, Mass Gate collision, Velocity Gate collision, Laser collision, Abort.

continued on following page

Table 7. Continued

Variable name	Meaning	Type	Notes
ended.at.action	Number of actions that fired successfully	Observed	
Par	Model-derived metric of effectiveness.	Derived	See "Treatment of EPIGAME logs" for a complete description.
planningTime	Duration of the Planning Phase	Observed	
tl.Adds	Addition of Actions to the Timeline	Observed	
tl.Deletes	Deletion of Actions from the Timeline	Observed	Very rare (mean = 0.05 deletions per Attempt)
tl.Modifys	Modification of parameters of Actions already in the Timeline	Observed	
tl.Moves	Actions moved within the Timeline	Observed	Very rare (mean = 0.17 moves per Attempt)
added.Tl.Total	Sum of Timeline Adds, Deletes, Modifys and Moves	Derived	
tj.Adds	Waypoints added to the Trajectory	Observed	
tj.Modifys, tj.Moves, tj.Deletes	Analogous to the Timeline (prefix: tl.) count variables	Observed	These variables exist in the record but no instance of these types of events was recorded.
locX, locY	Coordinates of Surge's spaceship when an event or Action occurred	Observed	
fail.same	Did this Attempt fail at the same location, for the same reason?	Derived	
eff.actions.added	How many Actions fired in this Attempt compared to the number that fired in the preceding Attempt?	Derived	Could be negative.
par.delta	Difference in the par metric between this Attempt and the previous one	Derived	
par.sqrt, par.delta.sqrt	Square-root transformations of the par and par.delta variables	Derived	
is.abort	Did the student press the Abort button before the level otherwise ended?	Derived	The software also registers the Abort button press in the endState variable.
ActionLog	Combined variable that registers the Actions applied to Surge during the Attempt	Observed	Includes position, type, and location of each Action applied
EventLog	Combined variable that registers important moments of gameplay not caused by Actions	Observed	Not fully functional in this version of EPIGAME
Serial	Serial number of the Conceptual challenge	Derived (from ActionLog)	

continued on following page

Table 7. Continued

Variable name	Meaning	Type	Notes
failed.to	In case of failure of a Conceptual challenge, the specific action the student did not do	Derived (from ActionLog)	Possible values: fromStop, toStop, speedUp, slowdown
is.colinear	Does this Conceptual challenge also require students to execute a turn?	Derived (from ActionLog)	
constant.mass	Do students have to account for changes to Surge's mass during the Conceptual challenge?	Derived (from ActionLog)	Only Conceptual challenges that pass this test were analyzed here
Pre, Post	The student's pre- and post-test scores, respectively	Observed	
bin.1	Classification of students according to beginning and ending quartile in assessment score	Derived	See "Question 1: Sequence Mining" for detailed description of this classification

Mario M. Martinez-Garza is a graduate of the Learning Sciences and Learning Environment Design program at Peabody College of Education of Vanderbilt University. His main areas of interest are investigating the potential of play as a vehicle for learning through cognitive perspectives, data science and student modeling, and applications of theory-based design principles to support learning through game environments of all kinds. He holds a Master's degree in Education and has served as a middle-school math and science teacher, and as a game designer for several commercial and educational games companies.

Douglas B. Clark is a Professor of Design-Based Learning and the Learning Sciences at University of Calgary.. He researches students' conceptual change processes and approaches for scaffolding those processes in games and other technology-rich environments. Clark is currently PI of the NSF DRK12 Enhancing Games with Assessment and Metacognitive Emphases (EGAME) grant that focus on scaffolding students' conceptual change processes in digital game based environments as well as developing approaches for analyzing game-play data in real-time for formative evaluation and adaption of student's experiences in those environments. Clark is also Co-PI on Extending CTSiM: An Adaptive Computational Thinking Environment for Learning Science through Modeling and Simulation in Middle School Classrooms. Clark was PI of the Department of Education IES Explanation and Prediction Increasing Gains and Metacognition (EPIGAME) grant and the exploratory NSF DRK12 grant Scaffolding Understanding by Redesigning Games for Education (SURGE), CoPI on the NSF Cyberlearning grant Fostering Computational Thinking in Middle Schools through Scientific Modeling and Simulation (CTSiM), and was also on the leadership team for the NSF CLT grant Technology Enhanced Learning in Science (TELS).