Contextual Markup and Mining in Digital Games for Science Learning: Connecting Player Behaviors to Learning Goals

John S. Kinnebrew, Stephen S. Killingsworth, Douglas B. Clark, Gautam Biswas, Pratim Sengupta, James Minstrell, Mario Martinez-Garza, and Kara Krinks

Abstract—Digital games can make unique and powerful contributions to K-12 science education, but much of that potential remains unrealized. Research evaluating games for learning still relies primarily on pre- and post-test data, which limits possible insights into more complex interactions between game design features, gameplay, and formal assessment. Therefore, a critical step forward involves developing rich representations for analyzing gameplay data. This paper leverages data mining techniques to model learning and performance, using a metadata markup language that relates game actions to concepts relevant to specific game contexts. We discuss results from a classroom study and identify potential relationships between students' planning/prediction behaviors observed across game levels and improvement on formal assessments. The results have implications for scaffolding specific activities, that include physics learning during gameplay, solution planning and effect prediction. Overall, the approach underscores the value of our contextualized approach to gameplay markup to facilitate data mining and discovery.

Index Terms—Games for science learning, data mining, knowledge engineering, learning behavior

1 Introduction

Digital games provide a promising medium for science education [12], [17], [29]. The NRC report on laboratory activities and simulations [26] makes clear, however, that the design of physical and virtual learning activities, rather than the choice of a medium, has the greatest impact on learning outcomes. The meta-analysis by Clark et al. [6] underscores the central role of design in the efficacy of digital games for learning.

Unfortunately, most quantitative research on games for science learning has focused only on pre-post test comparisons to evaluate the games' effectiveness and compare different game designs (c.f., [17]). Pre-post comparisons lend themselves to straightforward comparisons of various treatments (e.g., control and experimental conditions), but they provide minimal insight into how game or level designs and player behaviors (i.e., their approach to playing the game) are linked to the learning process. Furthermore, pre-post test approaches cannot directly support formative assessment and feedback while students are playing the game. However, digital games provide ample opportunities to collect data during play that may provide insight into the evolution of students' thinking and learning (e.g., [5], [17]; [24], [25], [23]).

This paper describes an approach to making the learning context of game behavior explicit. Adding relevant metadata helps to interpret students' actions in the learning context, and then apply learning analytics and data mining techniques to model students' gameplay behavior. To assess this approach, we design learning context metadata in a physics game called SURGE Next and analyze the resulting annotated log data from a classroom study. The results provide initial evidence for the utility of contextual metadata coding and data mining for understanding links between gameplay and learning. In particular, the analysis identifies potential relationships between planning/prediction behaviors and reading informational text provided in game levels, to improvements on formal (pre-post test) assessments. In addition, this analysis points to a specific link between acceleration maneuvers in SURGE levels and learning of formal acceleration concepts. One implication of these results is that dynamic level sequencing and appropriate scaffolding for solution planning/prediction and reading activities (particularly for students who are less willing to engage in them) could aid students' learning during gameplay.

2 SURGE Next: Game Design

In this paper, we apply our metadata coding and mining approach to SURGE Next, a *conceptually-integrated game* for learning [4], where the physics concepts to be learned are integrated directly into the game mechanics. Fig. 1 illustrates gameplay in SURGE Next, which begins in an outer-

- *J. S. Kinnebrew and G. Biswas are with the EECS Department/ISIS, Vanderbilt University, Nashville, TN 37240. E-mail: john.kinnebrew@gmail.com, gautam.biswas@vanderbilt.edu.*
- *S. S. Killingsworth, D. B. Clark, M. Martinez-Garza, and K. Krinks are with the Department of Teaching and Learning, Vanderbilt University, Nashville, TN 37240. E-mail: {stephenkillingsworth, kwarizmi, kara.krinks}@gmail.com, doug.clark@vanderbilt.edu.*
- *P. Sengupta is with the Department of Learning Sciences, University of Calgary, Calgary, AB, Canada. E-mail: pratim.sengupta@ucalgary.ca.*
- *J. Minstrell is with FACET Innovations, Seattle, WA 98105. E-mail: jimminstrell@facetinnovations.com.*

Manuscript received 23 Apr. 2014; revised 19 Aug. 2015; accepted 20 Jan. 2016. Date of publication 25 Jan. 2016; date of current version 16 Mar. 2017. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TLT.2016.2521372

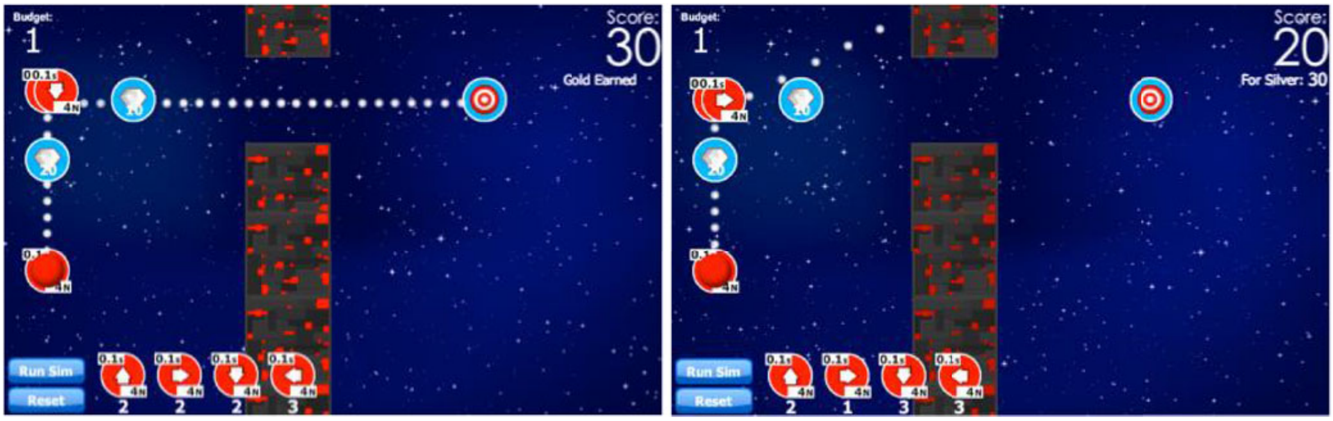

Fig. 1. Simple SURGE level focusing on fewer, but more consequential, actions.

space environment (implying there are no gravitational or frictional forces). In SURGE Next, players navigate their red spherical ship from its starting position to a target while avoiding obstacles. SURGE focuses on popular gameplay mechanics including: (a) supporting engagement and approachable entry (e.g., [15]), (b) situating the player with a principled stance and perspective (e.g., [19]), (c) providing context and identification for the player with a role and narrative (e.g., [10]), (d) monitoring and providing actionable feedback (e.g., [1]), and (e) using pacing and gatekeeping to guide the player through cycles of performance (e.g., [27]).

Rather than employing a real-time interface like the one used in the original SURGE Classic [3], where pressing an "arrow key" results in the immediate application of a force, SURGE Next requires the player to place all of the force commands (which vary in direction, magnitude, and time duration over which they are applied) on the game map in advance by dragging them from a palette (see Fig. 1). If the red ship's trajectory intersects the point where a force icon was placed, its trajectory is modified by application of that force. SURGE Next also limits the total number of force commands available in a given level to increase the salience and impact of each. Ideally, this should encourage players to consider more carefully the expected outcomes of each action as opposed to adopting trial and error methods. Thus, SURGE Next emphasizes deliberative gameplay, planning, and prediction (as opposed to reactive gameplay).

In designing SURGE Next levels, we first aligned in-game maneuvers with an application of one of Newton's three laws in a specific and explicit form. We then designed sequences of levels to highlight the relationships of the individual cases to the formal laws. As an example, levels like the one illustrated in Fig. 1 require application of Newton's first law to achieve a 90-degree turn. The right pane in Fig. 1 illustrates a common incorrect solution to this level. The player applies the correct upward force initially, but instead of applying a (downward) canceling force, the player uses a force directed to the right, assuming this will result in the ship turning to the right and moving in that direction. In this situation, the ship moves off in a diagonal trajectory to collide with a wall instead of the intended horizontal trajectory.

Our observations indicate that after an initial error with a right turn – as in the right pane of Fig. 1, some students successfully generate a correct solution in subsequent trials – as in the left pane of Fig. 1. Other students continue to increase the number of horizontal forces, assuming a large force will make the trajectory horizontal or at least get "close enough" to the desired solution. In other words, they are unable to apply the concept of canceling motion in a given direction by applying a force opposite to the direction of motion. In SURGE Next (henceforth "SURGE", for brevity), our approach to level design and logging of gameplay data facilitates automated analysis and helps us model differences in students' gameplay related to performance on assessments.

3 Contextual Metadata for SURGE Levels

Simple measures, such as success or failure in completing a level or the points earned in that level, are unlikely to support detailed inferences about students' conceptual development, students' strategies, or how well the game scaffolds students' learning. Some work on games for learning has incorporated the design of embedded assessments to measure conceptual understanding and provide formative assessments to help students learn, and for teachers or the system to adapt support within the game (e.g., [2], [24], [25]). Rather than including additional assessments, SURGE incorporates context-specific descriptions that link game actions (i.e., gameplay behavior) to relevant physics concepts. Other researchers have employed semantic annotation of objects in games, and rules linked to gameplay behavior to assess learning and conceptual understanding (e.g., [9], [16]). Our approach is similar, but its focus is not on formative assessments and feedback. Instead we codify generalizable relationships between gameplay and learning contexts, and this, enables effective mining and exploratory analysis of observed behaviors rather than focusing on assessment of known behaviors. Analyzing students' gameplay behavior with respect to the learning context can identify previously unknown errors and potentially unexpected/unknown strategies that can be connected to learning outcomes and theory beyond previously-conceived assessments.

Previous work has used researcher-guided protocols and self-report schemes for drawing connections between students' problem-solving approaches, game strategies, and conceptual change. This, requires significant time investment and are often infeasible for studies with many students. Further, students' self-reports on use of strategies and problem-solving schemes do not always correspond to the approaches they employ [28]. Merely bypassing the inaccuracies in self-reports by recording gameplay behavior is

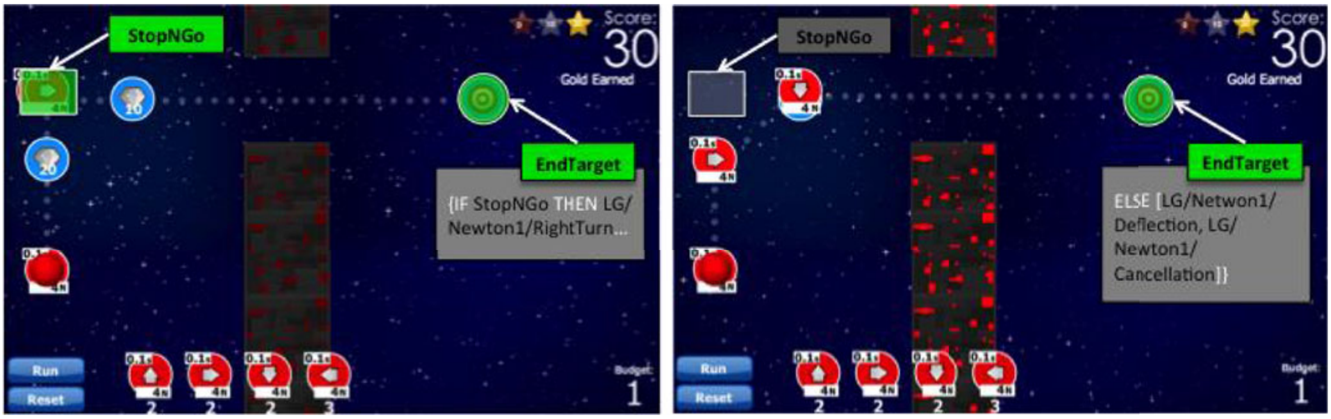

Fig. 2. Examples of metadata regions and their relationship to alternative solutions in SURGE.

insufficient. However, in SURGE, different game levels, different positions within a single level, and even different states of the player's ship (e.g., momentum and forces currently acting on it) alter the meaning of a given combination of forces. Thus, the context of a player's action(s) defines the semantics of those actions with respect to relevant physics concepts.

To contextualize actions in SURGE, we adopt a bottom-up approach that identifies relevant learning aspects of a game level during its design, in contrast toa top-down design of assessments imposed on (or external to) the game levels. Our coding scheme for game levels introduces machine-readable metadata linking potential actions to related physics concepts and learning goals at specific locations in each level. Tracking students' actions and inferring their relations to physics concepts allows us to analyze how and why students succeed (or fail) in particular levels.

SURGE's metadata tags allow the game to log the context of students' actions, and interpret them by noting the ship's collisions with visible objects and invisible "tracking" objects placed in the game environment. The invisible tracking objects detect the ship's movements into (and out of) particular regions. All encounters between the ship and obstacles and tracking objects during level play are logged along with relevant state information, such as its velocity vector, active forces, and mass.

The metadata language captures simple statements and complex conditionals. Simple statements capture contact with a tracking region and other objects associated with a particular metadata tag. More complex metadata triggers are specified by IF/THEN/ELSE constructs and Boolean logic expressions (conditions combined with the logical operators AND, OR, and NOT) associated with game objects. The conditions capture information, such as whether the player previously passed through a specified tagged region or object. Further, they may be used to express properties of the object's motion and state at the time (speed, direction of motion, or mass). For example, a very specific metadata expression could indicate that a particular feature will be associated with a trial only if the spaceship enters a tracking region at an angle between 90 and 180 degrees with a speed of 1 m/s and a mass of 1 kg. In practice, the large majority of metadata expressions are significantly less complex and require only that the spaceship passed through one or two tagged regions, sometimes with a state constraint (e.g., direction of motion).

For example, the illustrations in Fig. 2 include an invisible tracking object labeled "StopNGo," and metadata associated with the end target, labeled "EndTarget": {IF StopNGo THEN LG/Newton1/RightTurn ELSE [LG/Netwon1/Deflection, LG/Newton1/Cancellation]}. If the student's solution follows the right turn path, as shown in the left image, then the event log will indicate that the ship passed through the "StopNGo" tracking object and reached the "EndTarget." The metadata associated with the "EndTarget" will be parsed, and since the spaceship also passed through the "StopNGo" tracking object (IF StopNGo THEN . . .), the trial will be tagged with the feature indicated by the associated metadata rule (i.e., "LG/Newton1/RightTurn"). This feature indicates that the student correctly applied specific concepts related to a concrete sub-case of Newton's 1st law by stopping motion with the application of force in the opposite direction of motion while also applying an orthogonal force to produce a right turn in this level game.

Alternatively, if the student's solution follows the path shown in the image on the right side of Fig. 2, then parsing the metadata for the end target object (the "ELSE. . ." portion because the spaceship did not pass through "StopNGo") will tag the trial with the features "LG/Newton1/Deflection" and "LG/Newton1/Cancellation." These features indicate that the student correctly applied Newton's 1st law to (1) cancel the ship's current velocity by applying an opposing force, and (2) produce a deflection in the trajectory using a second orthogonal force (thus resulting in motion orthogonal to the ship's initial trajectory). Table 1 illustrates some of the other common metadata rules used in SURGE.

More generally, our coding scheme divides the solution space into a finite number of regions and states that relate learning goals to student actions by simple rules. Students are given a finite number of forces in each level, and a level is further constrained by the configuration (i.e., the placement of obstacles, diamonds, end targets, etc.). This makes the definition of contextual metadata relatively simple for most levels. Features derived from the contextual coding in the logged data, in turn, simplify subsequent analyses of interpreting students' attempts at solving the level, and the changes they make across the multiple attempts. This provides opportunities for applying systematic data mining and learning analytics approaches to gain a deeper, contextualized understanding of students' behaviors and learning performance, including their evolution over time [14]. In particular, techniques for

TABLE 1
Example Learning Goal Features

Learning Goal	Description
Newton1/Balance	Adding a balancing constant force (equal magnitude, opposite direction) to an existing constant force will result in constant speed. (e.g., overcoming friction)
Newton2/AccelForce	For a constant mass, acceleration is directly proportional to force (as force increases, acceleration increases)
Newton2/AccelMass	For a constant force, acceleration is inversely proportional to mass (as mass increases, acceleration decreases)
Newton3/ThrowForward	Throwing an object forward (in the direction of motion) will cause you to slow down or move backward (acceleration in the opposite direction)

modeling and identifying patterns in sequential activity data, coupled with the specific learning goal and error features resulting from the level metadata, can be used to identify patterns (in this case, sequences describing observed changes in game state and player actions that caused them) linking physics learning and gameplay behavior.

Analysis of these patterns can help identify important behaviors and suggest system refinements and scaffolding that benefit student learning in subsequent versions of SURGE and other games. In particular, connecting identified patterns with students' performance on pre-/post-tests can establish important areas of interaction between gameplay behaviors and domain learning. The following section describes the mining techniques employed to analyze SURGE data in more detail. The results and discussion presented in Sections 6 and 7 explore the affordances of this approach for generating new hypotheses related to learning through gameplay.

4 Differential Sequence Mining for SURGE Data

To investigate relationships between prior knowledge, formal assessment of learning gains, and gameplay behaviors in SURGE, we applied an exploratory data mining technique, Differential Sequence Mining (DSM) [13], [14], which iteratively analyzes sequences of student activities to identify the patterns that most clearly differentiate two groups of students by their frequency of use. This methodology applies four iterative steps to mine sequential activity patterns:

(1) *Activity abstraction*: Log files are processed to produce symbolic sequences of level *trials*. A sequence is an ordered series of trials that a particular student performed when attempting to solve a specific level. Each trial is defined by one or more events with additional measures (e.g., the medal received for that trial or the change in the number of forces between two consecutive trials) that are described in more detail in Table 2. For example, a sequence of three trials defined by the measures of change in forces and medal received might be: [2 Forces Added, No Medal] → [1 Force Added, Bronze Medal] → [1 Force Added, Gold Medal].

(2) *Differential Grouping*: In this study, we created two groups based on student learning gains from the pre- to post-test. To relate gameplay behaviors and knowledge gain, we selected students with low prior knowledge in the domain (those at or below the median score on the pre-test). For this group, the median gain from pre- to post-test was 0, so we split the students into two groups: *Gain group* – those who improved on the post-test (19 students), and *NoGain group* – those who did not improve on the post-test (23 students). We did not analyze data from the high prior knowledge group because very few of these students had improved post-test scores.

TABLE 2
Events with Corresponding Measures Used to Characterize Trials for Mining

Measure	Definition and possible values
End Cause	The manner in which the trial ended: (i) the player stopped the trial, (ii) the player reached the target (goal) to complete the level, (iii) the player collided with a dangerous object in the level, and (iv) the player went out of bounds.
Medal Achieved	The medal achieved for the trial (an indicator of the student's performance in completing the level): (i) no medal (if the player did not reach the target), (ii) bronze medal, (iii) silver medal, and (iv) gold medal.
Intro Screen Read	The amount of time the student spent on the level's introductory screen at the beginning of each trial, which provided information about the level and relevant physics concepts: (i) "Short" (≤ 4 seconds) and (ii) "Long" (> 4 seconds).
Change in Number of Forces	The net change in the number of forces employed versus the previous trial, a rough measure of the extent to which the student revised or extended their previous solution: (i) ≤ -3, (ii) -2, (iii) -1, (iv) 0, (v) 1, (vi) 2, and (vii) ≥ 3 change in forces.
Change in Number of Learning Goal Features	The change in the number of learning goal features (LG) achieved, an indicator of the student's successful accomplishment of a portion of the level related to a specific physics learning goal: (i) ≤ -3, (ii) -2, (iii) -1, (iv) 0, (v) 1, (vi) 2, and (vii) ≥ 3 change in learning goal features.
Change in Number of Error Features	The change in the number of error features (indicator of the student's failure to accomplish some portion of the level related to specific physics learning goals): (i) ≤ -3, (ii) -2, (iii) -1, (iv) 0, (v) 1, (vi) 2, and (vii) ≥ 3 change in error features.

(3) *Candidate pattern generation*: We employ a standard sequential pattern mining algorithm to identify trial patterns that were sufficiently frequent for that group. The threshold for selection was set at 5 percent to accommodate the wide variety of levels analyzed (14). Sequential pattern mining allows us to identify every subsequence that occurs in at least a given percentage of sequences. In this case, that means that all subsequences that occurred in at least 5 percent of either group's level sequences are identified as candidate patterns. These patterns were then analyzed to identify those that differentiate the two groups by frequency of occurrence.

(4) *Differential comparison*: A t-test on pattern occurrence across the groups was used as the selection criteria to identify the *differentially-frequent* patterns from the set generated in Step 3 [14]. This step filtered out patterns that did not have strong differences in frequency of occurrence across the two groups. The remaining set of differentially-frequent patterns were ranked and analyzed using the ratio of the frequency in the Gain group level attempts to the frequency in the NoGain group level attempts.

5 Method

The primary goal of the analysis in this paper is to use exploratory methods to illustrate the affordances of our metadata and mining approach in deriving insights that link gameplay to learning context. Therefore, the primary findings in this paper yield insights that relate gameplay behavior and learning of physics concepts. In this framework, we *generate* hypotheses about mechanisms and potential scaffolds that improve learning from gameplay, rather than to confirm pre-defined hypotheses. To achieve this, we employ data from a SURGE study described below.

Subjects. 68 eighth-grade students (37 male and 31 female) from six classes taught by the same teacher participated in this study and used the version of the game described in Section 2. The public middle school, located in Southeastern United States, primarily serves a diverse lower/middle-class population. The intervention was conducted during the students' science class, and amounted to approximately 3 hours of gameplay, and 1 hour of pre-post assessments spread over a week. Only data from the 42 students who completed the IRB assent form, the pre-test, and the post-test is included in this analysis.

Design. The study used a pre-test–intervention–post-test design. Before playing the game, each student completed a pre-test adapted from the Force Concept Inventory (FCI). After the pre-test, students played the SURGE game with levels designed to roughly align with the same focal concepts questions in the FCI-based test. Students had no prior classroom instruction on Force and Motion concepts and no prior experience playing the SURGE game before the study began. After playing SURGE, students completed the post-test.

Assessment of physics understanding. The pre- and post-tests consisted of 12 multiple-choice questions based on the Force Concept Inventory [11]. Questions covered basic concepts relevant to understanding Newton's Laws: (1) vector combination and diagonal motion (vectors); (2) the relationship between velocity, acceleration, and position (acceleration); (3) the influence of friction on motion (friction); (4) the influence of mass on motion (mass); and (5) the influence of gravity on motion (gravity).

Level metadata markup. As discussed, each level was coded with metadata and invisible tracking objects to facilitate post hoc analysis of the progression of students' solution attempts in terms of key learning goals and potential errors related to the learning goals. For analysis, these learning goals were further grouped into four physics-related conceptual categories explored in the game: *diagonals/deflections/turns*; *forces and cancelling*; *acceleration and friction*; *mass*; and *launching*. The diagonals/deflection/turns group included these three kinds of 2D motion. The forces and cancelling group included selecting and cancelling forces to achieve required speeds. The launching group focused on Newton's third Law. The relevant game actions for these learning goals involved placing force vectors at appropriate locations to speed up, slow down, or deflect the spaceship. Because the majority of students did not reach levels that included the launching learning goals, we did not analyze this learning goal.

6 Results

We present two sets of results. First, we explore relationships between patterns mined from logged gameplay data and pre-post learning gains. Then we explore the relationship between pre-post learning gains and game performance. t, Our purpose is not to demonstrate the efficacy of the beta version of SURGE used in the study in terms of learning gains (significant on vector questions, $F(1, 63) = 4.37$, $p < 0.05$, $\eta_p^2 = 0.07$, but not overall, $F(5, 59) = 1.50$, $p = 0.21$, $\eta_p^2 = 0.11$). Instead, we highlight the affordances of our metadata and mining approach in deriving insights by linking gameplay to learning context.

6.1 Relationships between Gameplay Patterns and Learning Gains

In order to investigate student gameplay behavior and its relationship with overall pre- and post-test performance, we employed the DSM exploratory data mining technique described in Section 4 to identify important differences in trial activity patterns between the groups of students who demonstrated learning gains (Gain, $n = 19$) and those who did not (NoGain, $n = 23$). The trial sequences were relatively short, and multiple measures were used to define each trial (see Table 2). As a result, many of the differential patterns describe a single trial, while others extended over short sequences of trials.

We ran the mining process iteratively as described in Section 4, employing 15 different combinations of measures to define trials. The DSM algorithm identified 65 differential patterns across these mining iterations. Some patterns were less interesting because they simply confirmed expected relationships. For example, when using measures for the change in learning goal features and change in error features to characterize trials, the resulting patterns simply indicated that the *Gain* group tended to achieve additional learning goal features more frequently without increasing the error features. Though this pattern is generally expected, it confirms that

TABLE 3
Selected Trial Patterns that Show Differential Usage between the Gain and the NoGain Groups

Measures	ID	Pattern	Freq. Ratio (Gain : NoGain)	% of Levels (Gain)	% of Levels (NoGain)
Change in Forces, End Cause	Gain1	[3+ Forces Added, Target Reached]	1.56	7.3%	4.7%
	NoGain1	[2 Forces Added, Collision]	0.77	11.6%	15.0%
	NoGain2	[1 Force Added, Stop]	0.77	21.0%	27.2%
Change in Forces, Change in Learning Goal Features, Change in Error Features	Gain2	[3+ Forces Added, 2 More LG, Same # Errors]	1.99	7.8%	3.9%
	Gain3	[3+ Forces Added, 1 More LG, Same # Errors]	1.57	6.9%	4.4%
	NoGain3	[1 Force Added, Same # LG, 2 More Errors]	0.61	3.3%	5.4%
Intro Screen Read Time, End Cause	Gain4	[Long Intro Read, Target Reached]	1.88	15.9%	8.5%
	Gain5	[Long Intro Read, Stop] –> [Short Intro Read, Stop]	1.71	8.7%	5.1%
	NoGain4	[Short Intro Read, Stop]	0.91	43.0%	47.4%
	NoGain5	[Short Intro Read, Stop] –> [Short Intro Read, Stop]	0.81	18.9%	23.4%
Intro Screen Read Time, Change in Learning Goal Features	Gain6	[Long Intro Read, 2 More LG]	2.74	6.9%	2.5%
	Gain7	[Long Intro Read, 1 More LG]	1.58	10.4%	6.6%
	Gain8	[Long Intro Read, Same # LG]	1.27	41.9%	33.1%
	NoGain6	[Short Intro Read, Same # LG]	0.92	79.2%	86.2%
Medal Achieved	NoGain7	[No Medal] –> [Gold Medal]	0.70	6.9%	9.9%

low prior knowledge students who gained more on the assessment also had greater success in the game.

In addition to confirming expected relationships, other trial features produced more interesting patterns that we discuss in greater detail below. We use five trial characterizations with 15 total patterns to illustrate important differences in gameplay and t the relationships between gameplay and learning. Table 3 lists these 15 differential patterns along with the percentage of levels in which the pattern occurred (for each group described in Section 4) and the ratio of those occurrences between the two groups. The patterns that appeared differentially in the Gain group's trials are in highlighted rows. For example, the first pattern, labeled Gain1 describes a trial in which at least 3 forces were added and the level was successfully completed by reaching the end target. Trials matching this pattern occurred 1.56 times more often per sequence (i.e., the series of trials performed by a student on a single level) in the Gain group than in the NoGain group. Overall, this pattern matched at least one trial in 7.3 percent of the level sequences for the Gain group and 4.7 percent for the NoGain group.

Patterns relating solution planning to learning gains. Patterns Gain1-Gain3 and NoGain1-NoGain3 illustrate that successful planning and prediction was more strongly linked to success (both on the level and in application of physics concepts) for the Gain group. These patterns involve the addition of multiple forces, which required better planning and combining predictions of the effects of forces. The first group of patterns investigated (Gain1, NoGain1, NoGain2) were from trials characterized by the *change in forces employed* (with respect to the previous trial) and the *end cause* (indicating success or failure) for the trial, where Gain1 was 1.5 times more common for the Gain group.

The hypothesized link between planning/prediction and learning is further supported by considering patterns NoGain1 and NoGain2 in which the NoGain group was more likely to cause a collision with a dangerous object when they applied two or more new forces, and more likely to stop the run after adding a single new force. Considering the gameplay mechanics and observing student play, stopping the run indicates that the student realized that additional forces would be required to reach the target or that their spaceship was not following the intended path. This suggests that the NoGain group did not fully understand the consequences of applying additional forces, and adding multiple forces often led to unexpected collisions. When they attempted to develop an understanding by adding one force at a time, they still generated deviations from the expected trajectory. Sometimes they realized that they needed to add more forces to reach the goal.

The other differential patterns that included the *change in forces measure* (Gain2, Gain3, NoGain3) further suggest that the Gain group was more effective in planning and prediction because their force additions led to achieving more learning goals (one or two) without increasing the number of errors in the solution. On the other hand, the NoGain group was more likely to increase errors in their solution even when they added a single force. The lack of progress indicated by the patterns also suggest that the NoGain group may have resorted to trial and error methods.

Patterns relating reading time and success measures. Patterns Gain4-Gain8 and NoGain4-NoGain6 illustrate that the *time spent reading* the introductory page was more frequently linked to success (on game levels and in application of physics concepts). As compared to the NoGain group, the Gain group was nearly twice as likely to successfully complete a level on a trial where they spent a longer time reading the introductory material (pattern Gain4). The Gain group was also more likely to combine long and short reads of the material (in consecutive trials) and manually stop the trial (pattern Gain5), indicating that they may have paid more attention to the introduction especially when they had difficulties in achieving the correct solution. In contrast, the NoGain group was more likely to combine one or more short reads, and then stop the trial (patterns NoGain4 and NoGain5), suggesting that they did not use the introductory material to help them correct their solutions.

Patterns Gain6, Gain7, and Gain8 also show that the Gain group's long reads corresponded to more success in attaining learning goal features. In contrast, the NoGain group's short reads did not help in achieving the learning goals. In combination with Gain4, these results suggest that the Gain group not only spent more time reading the intro screen, but they also profited from it more than the NoGain group.

Patterns of medals achieved on subsequent level attempts. At face value, pattern NoGain7 was particularly surprising because it indicates that the NoGain group was more likely to jump directly from a trial in which they did not complete the level (*no medal achieved*) to one in which they completed the level with a *gold medal* (i.e., the highest possible performance on the level). However, a deeper analysis of this pattern showed that this differential occurrence was largely confined to three short levels that introduced the effects of change in mass on acceleration. In these levels, a gold medal was earned for any completion of the level. Furthermore, the NoGain group often continued to attempt additional incorrect solutions on these levels before they returned to the correct solution. Therefore, this pattern does not indicate a particularly useful behavior by the NoGain group. Instead, this may suggest that the relationship between force, mass, and acceleration, i.e., Newton's Second Law, was particularly confusing for the NoGain group.

6.2 Relationship between Learning Gains and Game Behavior Covariates

The previous analyses did not link the learning goal features with the specific concepts to ensure that a sufficient number of comparable trial sequences for data mining, In this section, we present further exploratory analyses to examine the relationship between in-game behavior (as measured by triggering learning goal features within certain conceptual categories) and pre-post learning gains on specific concepts.

A repeated-measures MANCOVA analysis (along with separate univariate analyses) was conducted with percent correct on each post-test question type (vectors, acceleration, mass, friction, and gravity) included as separate measures. Test administration (pre vs. post-test) was included as a within-subjects factor. Counts of each different category of game learning goal features triggered by students' actions were included as covariates (to evaluate the relationship between post-test scores and game behavior). The learning goal feature covariates with their mean counts and standard deviations included *diagonals/deflection/turns (mean 11.27, SD 5.64)*, *forces and cancelling (mean 12.61, SC 4.85)*, *mass (mean 2.05, SD 2.54)*, and *acceleration and friction (mean 6.45, SD 4.63)*. All of these covariates were used in the analysis to examine the influence of each gameplay learning goal count while controlling for others.

Our analyses focused on the main effects of covariates or interactions between test administration and gameplay learning goal covariates to determine if learning gains were related to contextualized game behavior. None of the effects were significant in the multivariate tests (i.e., none of the learning goal covariates explained multivariate improvement across the range of test question types). Most importantly, univariate tests for acceleration assessment items showed a significant interaction between test administration and *acceleration and friction* learning goal count, $F(1, 59) = 4.74$, $p < .05$, $\eta_p^2 = .07$. Higher numbers of *acceleration and friction* learning goals triggered predicted significantly greater gains on acceleration assessment items, $r_{partial}(59) = .27$, $p < .05$. No other univariate effects were significant in the analyses.

7 Discussion, Implications, and Final Thoughts

Together, the analyses described in Section 6 illustrate the potential utility of the contextual metadata coding and data mining to link gameplay and learning context for more detailed investigations of student learning and performance. Though the overall learning gains were minimal, incorporating the metadata coding and data mining provided revealing information about learning within the game and illustrated the potential for these techniques to reveal relationships between game performance and improvement on formal assessments. In this section, we discuss the implications of these results and our next steps to expand upon this approach in future analyses.

7.1 Discussion

Relationships Between Gameplay Patterns and Learning Gain. Gameplay patterns provided insights into student learning processes in this pilot version of SURGE. The data mining results illustrated three important differences in level attempt behavior between the students who showed a learning gain on the pre-/post-test and those who did not.

One set of identified patterns indicated that low-prior knowledge students who developed the competencies of using three or more new force commands to correctly solve a level gained more on the formal assessment also. On the other hand, low-prior knowledge students who showed no gains on the assessment had difficulties in improving theory results from their previous trial even when they added one force command at a time. A reason for this lack of success may be that these students could not link underlying physics principles to the game scenarios, and they tried to overcome their lack of knowledge (or misunderstanding) of the physics concepts by making repeated incremental modifications that did not result in improved outcomes. Alternatively, these students may have adopted a brute force trial-and-error strategy for game play, and it worked for some simple levels. However, the lack of systematic planning, prediction, and interpretation of game play resulted in lack of success and missed opportunities for learning the physics concepts.

The observed differences may also reflect an underlying difference in students' working memory capacities. Planning, prediction, and interpretation are often a function of domain-specific knowledge and domain-general working memory capacities. Students who exhibited more planning behavior may have had more free working memory resources to extract meaningful general rules from their experiences in the game. This analysis may imply a causal link between the ability to plan and predict the outcome of game actions and the learning of the physics concepts, but additional studies are needed to confirm the link to physics learning.

A second set of identified patterns for students with learning gains indicated a similar link between time spent in (re-)reading the introductions to a level and improved performance on that level. This may be attributed to reading abilities: Perhaps better readers found the content useful, and extracted useful information that led to better level solutions. Differences in verbal working memory capacity might also explain this finding: Students who read the text for a longer period of time may have better abilities to understand and retain information from the intro materials while exploring the game map.

Another pattern implied that students with no learning gains more frequently showed jumped from a failure to a complete success for a level. Further analysis showed that this pattern was misleading – this pattern occurred only for some of the simpler levels in which mass concepts were introduced, and the sudden jump from failure to completion happened after these students' had made repeated unsuccessful attempts to solve the rather simple level. It may be that trial-and-error approaches eventually result in success for these simple levels, while greater understanding and planning is a prerequisite for completing some of the more complex levels.

Since some activities (more extensive planning/prediction and reading of the provided material about the levels) were linked with important game success measures and overall learning, one potential implication is that dynamic level sequencing and appropriate scaffolding for these activities (particularly for students who are less willing to engage in them) could aid physics learning during SURGE gameplay. Since the data mining analysis is exploratory, these results do not confirm any experimental hypotheses, but rather suggest that these areas may be fruitful for further investigation of the connections between learning and gameplay behaviors. Future refinements and experiments with SURGE will systematically test these scaffolding hypotheses.

Relationship between learning gains and game behavior covariates. Finally, incorporating the metadata coding into analyses of learning gains provided further insights about specific physics concepts and learning in SURGE. The analysis of the effects of learning goal feature covariates on test learning gains showed that gains on acceleration items were greater when students successfully performed behaviors that involved acceleration and friction.

The relationship between improvement on acceleration test items and game successes with acceleration scenarios suggests that the students who solved more of the acceleration-related levels successfully in the game learned more about the related physics concepts. However, successes in acceleration scenarios did not relate to larger test gains in other item types. This emphasizes the content-specific connection between game successes and test learning gains. Though this result is currently limited to acceleration, it provides an important first step by demonstrating that success in contextualized game behavior can be directly related to learning of specific physics concepts. The metadata coding and the DSM methods can help us identify the relations between contextually-relevant game behaviors and assessment-based learning gains to draw inferences about how domain knowledge can be acquired from a game.

7.2 Implications and Future Work

In the original SURGE (SURGE Classic), the lack of contextual metadata limited our analyses to pre-post test gains and high-level analyses of gameplay data (e.g., "how many actions did the player use per trial?"). We collected raw gameplay data, but without level-specific metadata, sequences of actions could not meaningfully be analyzed in terms of related physics concepts and the context in which they were applied [17]. In the current study, the contextual metadata provided insight into the connections between gameplay and learning. The analyses presented in this paper represent a valuable first step toward distilling large amounts of player data and interpreting them as domain-relevant learning behaviors by: (a) incorporating contextualized metadata during level design and (b) analyzing gameplay behaviors in an automated manner using this metadata rather than analyzing raw player actions. Our findings demonstrate that this approach provides a deeper understanding of how gameplay is connected to the development of both intuitive and formal understandings in ways that can support real-time adaptation of gameplay. In SURGE Next, these results can be leveraged to support players' evolving understanding of the underlying physics relationships, as well as their game-playing and problem-solving strategies.

Based on these findings, we are working to expand and refine the approach in terms of the grain-size of the focal metadata tagging. In the current metadata level coding, specific physics learning goals and corresponding errors were the focus. Each learning goal was coded primarily within the first few levels that introduced the relevant concepts. As each subsequent new learning goal was introduced, however, that new learning goal became the focus. This was an effect of both a primary focus on the "new" learning goals as they occurred in the progression and of the complexity of coding for many different learning goals/errors in a single level. In order to better analyze students' developing understanding of the physics concepts over the full course of the game, we are refining our metadata markup approach to focus on maneuvering regions in a level and a minimal amount of information about the expected maneuver, then automatically relating the student's maneuvers to relevant facets [20], [21] of physics understanding (instead of explicitly marking each learning goal and error of interest in the level). This approach builds on a knowledge-in-pieces theoretical account of student learning [7], [8]. A facet-based account of student learning assumes that students bring a large number of resources and ideas with them that combine everyday experiences and formal learning experiences. The specific ideas that students apply in a given context are cued by the particulars of the context. Learning from this perspective involves students refining these specific ideas (or facets) in terms of how they are combined, cued, and applied.

This facet-based approach is promising for data-mining in games because facets of student thinking can be inferred as corresponding to a given set of forces employed during a maneuver, in combination with the expected motion resulting from the maneuver. These combinations of actions and contexts for the actions remain consistent across levels. Therefore, it suffices to mark the regions in which maneuvering is necessary along with the expected motion (i.e., direction of motion and/or speed) in the context of the level design. Then the connection to physics understanding can be inferred by combining this information with knowledge of the student's actual use of forces within the maneuvering region. Thus, the quantity of metadata markup needed for a level is a more compact *per maneuver* rather than *per potential learning goal and error of interest*. This significantly reduces overhead for marking levels, while allowing us to identify facets across the full progression of levels and thus track and analyze learning progressions and processes.

In the level from Figs. 1 and 2, for example, only the general region in which the turning maneuver can be executed is marked, and the metadata only contains a name and the expected direction of motion upon exiting. Given this very simple markup, the example level solution can be automatically analyzed based on the student's placement of a canceling (down) force and a perpendicular (right) force overlapping each other within the maneuver region, combined with the logged information that the player entered the region moving north and exited in the expected direction of east. For this example, the resulting facets would be "canceling" (indicating the student understood they could stop the northward motion by using an opposing force) and "force as mover" (indicating the student understood that then applying a force to the east would cause movement in that direction). In contrast, if the student used the angled trajectory approach illustrated in Fig. 2B, the automated analysis would identify the resulting facets as "force as deflector" (indicating the student understood how to first deflect to the northeast) and then "component cancelling" (indicating the student understood how to cancel the northward velocity component with a force applied toward the south).

Equally important, common errors can also be identified in terms of facets with the same, simpler markup instead of requiring additional metadata to describe each potential error. If the student had only placed a force to the east (resulting in the incorrect trajectory illustrated in Fig. 2B) this would automatically be recognized as an error described by the application of the facet "force as a mover" without considering or understanding the other facets of "force as a deflector" or "canceling." In this approach, most of information about how to translate level context plus gameplay behavior into physics understanding is contained in the rules that connect facets to the combination of: (1) motion entering a maneuver, (2) expected motion exiting a maneuver, (3) actual motion exiting a maneuver, and (4) the forces employed in the maneuver. The information for (1), (3), and (4) is automatically logged by any invisible tracking object in a SURGE level, so the necessary coding to complete the contextual information for a level is simply adding tracking object(s) to indicate the region(s) in which a maneuver may occur and providing simple metadata indicating the expected motion upon exiting each region. Thus, with less tagging, we can analyze not only whether or not the student displays a normative understanding, but also whether or not the student displays behaviors indicative of specific common alternative conceptions, which then could provide the basis for productive targeted scaffolding.

Further, to track developing understanding more consistently over the course of the game, we intend to refine the level design process to incorporate repeated level *segments* throughout the progression. Individual level segments will assess student understanding of each physics concept, but these concepts will re-appear in multiple levels over the course of the game. This design allows for more systematic assessment of students' developing proficiency with individual learning goals and can support consistent formative assessment in a classroom setting. Finally, the design may provide more stable diagnostic information that could drive adaptive content.

Another issue with our current data mining analyses is that the characterization of trials (i.e., each level attempt) is limited to high-level features. In order to assess planning, solution refinement, and related activities more effectively, we propose to extend our approach with additional measures that describe each solution component. In SURGE Next, a solution component is a contextualized force placement, which can be linked to relevant physics principles (e.g., a deflection, a vector component cancellation, or a force not triggered on the current trajectory). The use of additional performance measures may help us quantify how changes in force placements from one trial to the next bring the student's solution closer to the goal (or not). This analysis, combined with the new markup language, can clarify how students' strategies and solution structures evolve.

Ultimately, observed behavior patterns can be used to drive real-time adjustments of the scaffolding and game experience. For example, predictive classifiers can be applied to dynamically adjust the scaffolding based on a student's behavior patterns in early levels. We are developing a scaffolding approach through dialog with agents (characters) in the game world, which is framed as the student helping the character to resolve a mission challenge or task. A key role of the agents in this approach will be engaging the students in prediction, reflection, and articulation of important causal relationships. This dialog first focuses on helping the student reflect on their experiences in a level and articulate the relevant causal relationships involved in core game maneuvers. The dialog then extends this reflection to articulating the specific relationships in terms of more formal disciplinary ways of thinking. We refer to our approach as "explanation games" [4], [5] based on research on self-explanations (e.g., Roy & Chi, 2005; [18]) as well as prediction, observation, and reflection (Kearny, 2004).

7.3 Final Thoughts

Commercial game designs provide powerful affordances for science learning and engagement. It is important to remember, however, that these affordances evolved under different pressures and goals that may not have links to science learning. Rethinking and redesigning these conventions to support logging and analysis of gameplay behavior with respect to both the learning context and the gaming context is of central importance. Early work on data mining in games often focused on mining of data at the level of specific actions without aggregating up to more salient levels in terms of students' understanding. Incorporating relatively small amounts of metadata to describe the learning and gaming contexts within games can produce rich data on student behavior. In combination with data mining techniques, this approach can yield deeper insights into science learning with games, which in turn can be leveraged to enhance adaptive scaffolding. The approach described here allows games to continually capture context and, therefore, assess actions in context in real time. As a result, one can systematically aggregate data in support of inferences about student understanding, which in turn could then drive real-time scaffolding based on these inferences about specific facets of students' normative or alternative conceptions.

References

[1] L. A. Annetta, J. Minogue, S. Y. Holmes, and M.-T. Cheng, "Investigating the impact of video games on high school students' engagement and learning about genetics," *Comput. Edu.*, vol. 53, no. 1, pp. 74–85, 2009.

[2] G. Bente and J. Breuer, "Making the implicit explicit: Embedded measurement in serious games," in *Serious Games: Mechanisms and Effects*, U. Ritterfield, M. J. Cody, and P. Vorderer, Eds. New York, NY, USA: Routledge, 2009, pp. 322–343.

[3] D. B. Clark, B. C. Nelson, H.-Y. Chang, M. Martinez-Garza, K. Slack, and C. M. D'Angelo ,"Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States," *Comput. Edu.*, vol. 57, no. 3, pp. 2178–2195, 2011.

[4] D. B. Clark, and M. Martinez-Garza, "Prediction and explanation as design mechanics in conceptually-integrated digital games to help players articulate the tacit understandings they build through gameplay," in *Games, Learning, and Society: Learning and Meaning in The Digital Age*, C. Steinkuhler, K. Squire, and S. Barab, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[5] D. B. Clark, M. Martinez-Garza, G. Biswas, R. M. Luecht, and P. Sengupta, "Driving assessment of students' explanations in game dialog using computer-adaptive testing and hidden Markov modeling," in *Game-Based Learning: Foundations, Innovations, and Perspectives*, D. Ifenthaler, D. Eseryel, and G. Xun, Eds. New York, NY, USA: Springer -Verlag, 2012, pp. 173–199.

[6] D. B. Clark, E. Tanner-Smith, and S. Killingsworth, "Digital games, design, and learning: A systematic review and meta-analysis," *Rev. Educational Res.*, vol. 86, no. 1, pp. 79–122, http://rer.sagepub.com/content/early/2015/10/20/0034654315582065.full.pdf+html

[7] A. diSessa, "Toward an epistemology of physics," *Cognition Instruction*, vol. 10, pp. 105–225, 1993.

[8] A. diSessa and J. Minstrell, "Cultivating conceptual change with benchmark lessons," in *Thinking Practices Mathematics Science Learning*, J. G. Greeno and S. Goldman, Eds. Mahwah, NJ, USA: Erlbaum Assoc., 1998.

[9] I. Dunwell, P. Petridis, M. Hendrix, S. Arnab, M. Al-Smadi, and C. Guetl, "Guiding intuitive learning in serious games: An achievement-based approach to externalized feedback and assessment," in *Proc. Sixth Int. Conf. Complex, Intell. Softw. Intensive Syst.*, Jul. 2012, pp. 911–916.

[10] J. P. Gee, *Good Video Games and Good Learning: Collected Essays on Video Games, Learning and Literacy (New Literacies and Digital Epistemologies)*. Bern, Switzerland: Peter Lang Pub Inc., 2007.

[11] D. Hestenes, M. Wells, and G. Swackhamer, "Force concept inventory," *Phys. Teacher*, vol. 30, no. 3, pp. 141–158, 1992.

[12] Nat. Res. Council, *Learning Science Through Computer Games and Simulations*, M. A. Honey and M. Hilton, Eds. Washington, DC, USA: Nat. Acad. Press, 2011.

[13] J.S. Kinnebrew and G. Biswas, "Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution," presented at the *5th Int. Conf. Educational Data Mining*, Chania, Greece, 2012.

[14] J. S. Kinnebrew, K. M. Loretz, and G. Biswas, "A contextualized, differential sequence mining method to derive students' learning behavior patterns," *J. Edu. Data Mining*, vol. 5, no. 1, pp. 190–219, 2013.

[15] R. Koster, "*A Theory of Fun for Game Design*, Scottsdale, AZ: Paraglyph Press, 2004.

[16] J. Maderer, C. Gütl, and M. Al-Smadi, "Formative assessment in immersive environments: a semantic approach to automated evaluation of user behavior in open wonderland," presented at the *8th J. Immersive Edu. Summit*, Boston, MA, USA, 2013.

[17] M. Martinez-Garza, D. B. Clark, and B. Nelson, "Digital games and the US national research council's science proficiency goals," *Stud. Sci. Edu.*, vol. 49, no. 2, pp. 170–208, 2013.

[18] R. E. Mayer, and C. I. Johnson, "Adding instructional features that promote learning in a game-like environment," *J. Edu. Comput. Res.*, vol. 42, no. 3, pp. 241–265, 2010.

[19] J. McGonigal, *Reality Is Broken: Why Games Make Us Better and How They Can Change The World*. New York, NY, USA: Penguin Press, 2011.

[20] J. Minstrell, "Student thinking and related assessment: Creating a facet assessment-based learning environment," in *Grading the Nation's Report Card: Research From the Evaluation of NAEP*, N. Raju, J. Pellegrino, L. Jones, and K. Mitchell, Eds. Washington, DC, USA: Nat. Acad. Press, 2000.

[21] J. Minstrell R. Anderson and M. Li, "Diagnostic instruction: Toward an integrated system for classroom assessment," in *Reconceptualizing STEM Education: The Central Role of Practices*, R. Duschl and A. Bismack, Eds. New York, NY, USA: Routledge Taylor & Francis Group, 2016.

[22] U. Munz, P. Schumm, A. Wiesebrock, and F. Allgower, "Motivation and learning progress through educational games," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 3141–3144, Dec. 2007.

[23] J. P. Rowe, L. R. Shores, B. W. Mott, and J. C. Lester, "Integrating learning and engagement in narrative-centered learning environments," in *Intelligent Tutoring Systems*. Berlin, Germany: Springer-Verlag,, 2010, pp. 166–177.

[24] V. J. Shute, "Stealth assessment in computer-based games to support learning," *Comput. Games Instruction*, vol. 55, no. 2, pp. 503–524, 2011.

[25] V. J. Shute,*Stealth Assessment: Measuring and Supporting Learning in Video Games*. Cambridge, MA, USA: MIT Press, 2013.

[26] S. Singer M. L. Hilton, and H. A. Schweingruber, Eds., *America's Lab Report: Investigations in High School Science*. Washington, DC, USA: Nat. Acad. Press, 2005.

[27] K. Squire "From content to context: Videogames as designed experience," *Edu. Res.*, vol. 35, no. 8, p. 19, 2006.

[28] P. Winne and D Jamieson-Noel, "Exploring students' calibration of self reports about study tactics and achievement," *Contemporary Edu. Psychol.*, vol. 27, no. 4, pp. 551–572, 2002.

[29] P. Wouters, C. van Nimwegen, H. van Oostendorp, and E. D. van der Spek, "A meta-Analysis of the cognitive and motivational effects of serious games," *J. Edu. Psychology*, vol. 105, no. 2, pp. 249–265, 2013.

John S. Kinnebrew conducts research that focuses on machine learning and data mining techniques to identify and model important student learning behaviors in computer-based learning environments.

Stephen S. Killingsworth conducts research that investigates how basic cognitive capacities and embodied cognition shape visual attention and support conceptual change.

Douglas B. Clark conducts research that explores students' conceptual change processes and the design of digital environments to support those processes.

Gautam Biswas combines ideas from artificial intelligence and cognitive science to develop intelligent open-ended learning environments for STEM learning.

Pratim Sengupta researches and develops programming environments for children to support science learning.

James Minstrell conducts research and development focusing on the identification and assessment of students' facets of thinking and developing diagnoser.com and other tools to help teachers address problematic as well as goal-like thinking.

Mario Martinez-Garza studies the design of digital games for learning as well as the nature of what is learned in digital games.

Kara Krinks conducts research that focuses on teacher and student thinking about physics through formal and informal representations of problems.