

Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

# Computers & Education

journal homepage: [www.elsevier.com/locate/compedu](http://www.elsevier.com/locate/compedu)

## Self-explanation and digital games: Adaptively increasing abstraction

Douglas B. Clark<sup>a,\*</sup>, Satyugjit S. Virk<sup>a</sup>, Jackie Barnes<sup>b</sup>, Deanne M. Adams<sup>a</sup><sup>a</sup> Department of Teaching and Learning, Vanderbilt University, USA<sup>b</sup> Game Design Program, Northeastern University, USA

### ARTICLE INFO

#### Article history:

Received 7 May 2016

Received in revised form 14 September 2016

Accepted 17 September 2016

Available online 21 September 2016

#### Keywords:

Self-explanation

Digital games

Science education

### ABSTRACT

Research suggests that self-explanation functionality can effectively support learning in the context of digital games. Research also highlights challenges, however, in balancing and integrating the demands and abstraction of self-explanation functionality with the demands and structure of the game. These challenges are particularly true for games that are, themselves, cognitively more complex. The current study presents an approach that adapts the abstraction of self-explanation prompts based on a player's performance. The results demonstrate that students in this condition (a) scored significantly higher on the post-test than students whose self-explanation prompts were not adaptively adjusted and were always abstract and (b) scored higher, but not significantly so, than students who did not receive the self-explanation functionality. Analyses of gameplay metrics suggest that trade-offs in terms of progress through the game may explain some aspects of these post-test comparisons. Analyses also demonstrate that both self-explanation conditions significantly outperformed the navigation-only comparison condition on a gameplay metric that suggests deeper model-based thinking.

© 2016 Elsevier Ltd. All rights reserved.

### 1. Introduction

Prompting students to engage in self-explanation can enhance learning by encouraging students to engage in meta-cognitive activities to monitor what they do and do not understand (Chi & VanLehn, 1991; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Roy & Chi, 2005). Previous research demonstrated, however, that incorporating questions to prompt self-explanation in a digital game can disrupt cognitive and gameplay flow in a manner that results in fewer game levels completed and no significant increases in performance (Adams & Clark, 2014). O'Neil et al. (2014) similarly concluded that, while self-explanation prompts can support generative processing, self-explanation prompts can also result in (a) extraneous processing by slowing down the player and distracting the player from learning or (b) minimal processing due to the player ignoring the prompts and trying to return to gameplay as quickly as possible. These findings contrasted with other previous findings stating that self-explanation and explanatory prompts can enhance learning in digital games (Hsu, Tsai, & Wang, 2012; Johnson & Mayer, 2010; Mayer & Johnson, 2010; Moreno & Mayer, 2005).

Together, previous research suggested that self-explanation functionality can be effective in games for learning, but this previous research also illuminated the challenges in balancing and integrating the demands and abstraction of self-

\* Corresponding author. Box 230, 230 Appleton Place, Nashville, TN 37203-5721, USA.

E-mail address: [doug.clark@vanderbilt.edu](mailto:doug.clark@vanderbilt.edu) (D.B. Clark).

explanation functionality with the demands of the game, particularly for games that are themselves cognitively more complex. Accordingly, the current study presents an approach that adapts the abstraction of the self-explanation prompts based on a player's performance. More specifically, the purpose of this study is to explore whether adapting the abstraction of self-explanation prompts based on students' performance might minimize cognitive and/or gameplay disruption and thus enhance learning and engagement. This study also explores how such prompts might differentially affect game play behavior and how variations in game play behavior might relate to variations in outcomes on conceptual assessments.

In support of these goals, the following structure is employed in presenting the study: (1) a theoretical framing is first outlined in terms of digital educational games and research on self-explanation, (2) methods employed in the study are outlined, (3) overall results are analyzed within and across experimental conditions in terms of the pre-test, post-test, engagement, and gameplay metrics, (4) results are analyzed within the adaptive self-explanation condition comparing performance across levels of abstraction within that condition, and (5) results, implications, caveats, and conclusions are discussed.

## 2. Background

Digital games are influential and ubiquitous presences in the lives of young learners. Digital games can be defined as interactive digital environments that (a) are based on a set of agreed rules and constraints, (b) are directed toward a clear goal that is often set as challenge, and (c) constantly provide feedback in terms of either scores or changes in the game world to allow players to monitor their progress toward the goal (Wouters, van Nimwegen, van Oostendorp, & van der Spek, 2013). A 2008 study by the Pew Internet and American Life Project found that 97% of teens aged 12–17 played computer, web, portable, or console games, and 50% of them reported daily or nearly daily play (Lenhart et al., 2008). A 2012 Pew study reported that digital games generated \$25 billion in sales in 2010 (Anderson & Rainie, 2012).

With this massive interest in digital games, aspects of recreational games have spread into education. The current study defines “educational games” as games intended to support learning of academic content. The games could be referred to as “serious games” in the sense “that the objective of the computer game is not to entertain the player, which would be an added value, but to use the entertaining quality for training, education, health, public policy, and strategic communication objectives” (Wouters et al., 2013, p. 2). Investigation into the use of games for education has grown from a small niche area to a major focus of research over the past 15 years (e.g., Gee, 2007). Reports by the National Research Council and others (e.g., Honey & Hilton, 2010; Martinez-Garza, Clark, & Nelson, 2013; Young et al., 2012) have acknowledged this potential, but also acknowledged the unevenness of systematic evidence for games as learning tools. Furthermore, although several recent meta-analyses have demonstrated that games can successfully support learning, these reviews have underscored the importance of design choices to the efficacy of educational games (Clark, Tanner-Smith, & Killingsworth, 2016; Wouters et al., 2013). At the broadest level, the purpose of the current study is to explore a design approach for increasing the efficacy of digital educational games through the integration of research on self-explanation and learning.

### 2.1. Self-explanation and learning

Self-explanation can be defined as a content-relevant articulation expressed by the student as the student engages in a learning activity (Chi, 2000). Prompting for self-explanation can take many forms, including verbal prompts (Chi, DeLeeuw, Chiu, & LaVanher, 1994), prompts generated by computer tutors (Alevan & Koedinger, 2002; Hausmann & Chi, 2002), or prompts embedded in the actual learning materials (Hausmann & VanLehn, 2007). The hypothesis underlying self-explanation research is that prompting students to self-explain their actions and their thinking behind key steps in a learning activity causes higher learning gains than having students study the material without such prompting. Self-explanation is viewed as “a domain general constructive activity that engages students in active learning and insures that learners attended to the material in a meaningful way while effectively monitoring their evolving understanding” (Roy & Chi, 2005, p. 272).

Research on self-explanation by Chi and others has explored the value of self-explanation for learning (e.g., Chi & VanLehn, 1991; Chi et al., 1989; Roy & Chi, 2005). Roy and Chi's (2005) review of research on students' self-explanations reported that self-explanation resulted in average learning gains of 22% for learning from text, 44% for learning from diagrams, and 20% for learning from multimedia presentations. According to Roy and Chi (2005), the process of self-explanation encourages four key cognitive mechanisms: (a) recognizing what information is missing while generating inferences, (b) integrating the information taught within the lesson, (c) integrating information from long term memory with new information, and (d) identifying as well as correcting incorrect information. Overall, self-explanation can encourage students to engage in meta-cognitive activities to monitor what they do and do not understand.

### 2.2. Self-explanation and educational games

Although research has shown that self-explanation can be effective for learning from texts, diagrams, and multimedia, research on educational games has demonstrated fewer consistent benefits. While some studies suggested efficacy in game-like settings (Johnson & Mayer, 2010; Mayer & Johnson, 2010; Moreno & Mayer, 2005), other studies demonstrated that self-explanation prompts can also result in either (a) extraneous processing by slowing down and distracting the player from

learning or (b) minimal processing due to learners ignoring the prompts in order to return to gameplay as quickly as possible (Adams & Clark, 2014; O'Neil et al., 2014).

Research has attempted to explore the design structures and contextual conditions under which self-explanation functionality is effective within games. A series of experiments conducted by Moreno and Mayer (2005) with the simulation/game *Design-a-Plant* (Lester, Stone, & Stelling, 1999) demonstrated that self-explanation facilitated additional learning only when the student is reflecting upon correct information and is not already engaged by the game in active cognitive processing. The first experiment showed no significant benefit for oral self-explanations, possibly because the game's interactive elements already engaged learners cognitively. The second experiment showed an interaction between self-explanation and interactivity. Students in the non-interactive version performed better on transfer and retention tasks after engaging in self-explanation while students in the interactive version performed better on retention questions when they did not reflect on their answers. Moreno and Mayer proposed as an explanation that, because the answers were provided by the system, the non-interactive condition guaranteed that students were reflecting only on correct answers. Students in the interactive condition, however, were potentially reflecting on answers that were incorrect because they were generated by the students themselves. In a third experiment, interactive students and non-interactive students were asked to reflect on correct answers provided by the program or themselves. The results showed that performance on far-transfer tasks was best for the learners who reflected on the program's correct responses regardless of interactivity. This suggested that self-explanation is most effective when students are asked to think about correct solutions rather than general reflections.

One possible way to reduce the cognitive demands and potential distraction of self-explanation prompts while ensuring that students are focusing on correct solutions involves providing learners with a set of explanation options. Using a game-like environment to teach about electric circuits, Mayer and Johnson (2010) found that students who were asked to select from a list of possible explanations to justify their answers performed significantly higher on the final transfer task level of the game (compared to students in the base version of the game). This result showed that even though the students did not generate the explanations themselves, adding the self-explanation questions prompted them to make connections between the game and the underlying rules of the electrical circuit system. The study also showed that there were no significant differences between students who received the self-explanation prompts, explanatory feedback, or a combination of the two. All three conditions facilitated performance on the final level of the game as well as improved performance, in terms of accuracy, on the first nine levels of the game.

In a second series of studies with the electric circuits games, Johnson and Mayer (2010) found that asking students to generate their own explanations did not result in significant improvement over the base version of the game. Students who selected an explanation from a list, however, showed significantly higher performance. Providing students with possible explanations as well as feedback can therefore decrease both incorrect thinking and extraneous processing. This result was fortunate from a pragmatic standpoint of design because open-ended self-explanation responses crafted by students may be difficult for game software to analyze in a responsive and feedback-rich manner.

Adams and Clark (2014) and O'Neil et al. (2014) have illustrated, however, that self-explanation questions may result in minimal processing even when students are provided with possible explanations. Adams and Clark (2014) highlighted the importance of minimizing the number of prompts and tightly coupling prompts with students' actions to optimize processing demands and to avoid disrupting the flow of the game. Adams and Clark (2014) and O'Neil et al. (2014) also noted that players sometimes actively avoided engaging in deeper processing during the self-explanation phase by randomly selecting each answer option until the game indicated that the correct answer has been selected. This decision circumvented the need to think or pay attention to the self-explanation functionality. Exploiting elements of the game interface to minimize thinking is known as "gaming the system" in a negative sense in the cognitive tutor literature (Baker et al., 2008). Adams and Clark (2014) findings demonstrated that students showing higher circumventing behavior achieved smaller learning gains regardless of learning condition. Hsu, Tsai, and Wang's (2012) study demonstrated a similar pattern. In Hsu et al.'s study, participants within the self-explanation group were divided into high- and low-engagement groups based on the ratio of correct and incorrect/"I don't know" responses to self-explanation prompts. While there were no significant differences between Hsu et al.'s self-explanation and control conditions on the retention test overall, high-engagement students in the self-explanation condition scored significantly higher than the students in the control condition and the low-engagement students in the self-explanation condition. This suggested that educational game designers must design a balance that engages students in thinking deeply about the material without overloading the students.

Lastly, designers of self-explanation functionality for digital games must consider prompt abstraction. O'Neil et al. (2014) provided three different types of self-explanation questions designed to target different types of processing in their fraction game, *Save Patch*. The first type of question targeted essential/intrinsic processing and attempted to attract the learner's attention to the units of distance that were used in the game. The second type of question was designed to help the learner draw connections between the game elements and the corresponding mathematical elements, thereby encouraging more intrinsic/essential processing. The third type of question was intended to promote generative processing by asking the learner to think more abstractly about movement mechanics in the game environment. Students in the self-explanation group were given all three question types after each game level and were asked to make two answer selections from one or two of the questions. Although the students showed no difference in learning outcomes between the base game and the self-explanation group, the results showed that students within the self-explanation group who answered more questions that drew connections between the game mechanics and the math concepts (question type 2) showed better learning process scores on game features (e.g., fewer deaths or higher levels reached in the game). The authors noted that although abstract

prompts were effective for older learners with the electric circuits game (Johnson & Mayer, 2010; Mayer & Johnson, 2010), younger children may find abstract rule statements too complicated.

O'Neil et al.'s (2014) results suggested that self-explanation prompts should not be too simple/concrete (question type 1) because students may only engage in minimal processing. O'Neil et al. also argued, however, that self-explanation prompts should not be too complex for the targeted learners (question type 3) because that may result in extraneous processing. O'Neil et al.'s results also showed a significant correlation between gameplay measures and performance on the learning outcomes, suggesting that students' abilities to engage effectively in the game impacted their learning outcomes. Adams and Clark (2014) similarly proposed that students in their study may have found the self-explanation prompts unhelpful because the prompts were too abstract. Adams and Clark (2014) also suggested that students are more likely to engage in deep processing when the self-explanation functionality focuses on students' specific challenges in the game.

While O'Neil et al. (2014) and Adams and Clark (2014) highlighted the challenges of incorporating self-explanation prompts that focus on abstract connections and relationships, these abstract connections and relationships are highly desirable in terms of learning goals. This tension raises the question of how a game design might leverage and scaffold abstract connections through the self-explanation functionality without unduly increasing processing demands.

After Adams and Clark (2014), the first study exploring this line of research by our research group compared three conditions for providing high-abstraction self-explanation prompts (Killingsworth, Clark, & Adams, 2015). Killingsworth et al. (2015) compared three conditions: (a) explanatory feedback where the high-abstraction relationship was provided directly to the student in the game dialogue as the explanation for the navigation phase just completed, (b) high-abstraction self-explanation where the student selected self-explanation answers from a high-abstraction prompt for the navigation phase just created, and (c) full self-explanation where the student selected self-explanation answers for first a low-abstraction prompt, then a medium-abstraction prompt, and finally a high-abstraction prompt after the navigation phase.

Results from Killingsworth et al. (2015) demonstrated that the high-abstraction self-explanation and full self-explanation conditions did not differ significantly. Therefore, there was a collapsing of the two conditions into a single "self-explanation" category for comparison with the explanatory-feedback condition. Overall, Killingsworth et al. (2015) determined that the self-explanation conditions promoted better near-transfer learning outcomes than the explanatory-feedback condition, but only after controlling for the number of levels that students had completed. Additionally, students' inhibitory control abilities were correlated with learning outcomes for the self-explanation condition but not for the explanatory feedback condition. Together, these results suggested that (a) self-explanation can benefit learning beyond explanatory feedback in educational games, but (b) the benefits of self-explanation are mediated by students' executive cognitive abilities. The findings underscored the challenges highlighted by O'Neil et al. (2014) and Adams and Clark (2014), including the increased processing demands inherent in adding the self-explanation functionality as well as the danger of overloading students with the self-explanation functionality.

### 2.3. Adaptivity in educational games

The current study explores an approach to scaffold abstract connections through the self-explanation functionality without unduly increasing processing demands by adaptively increasing the abstraction of self-explanation prompts as players increase in proficiency. Prior research indicates that adaptive feedback and prompts can improve learning. Chi, VanLehn, Litman, and Jordan (2011) found, for example, that adaptive learning systems for teaching introductory college physics improved learners' learning gains significantly. Reif and Scott (1999) studied the differences between human tutoring, step-based tutoring, and no tutoring for a physics domain. They found that gain scores between human tutors and step-based tutors were not significantly different, but were significantly better than having no tutor. VanLehn (2011) found intelligent computerized tutors that adapted content according to learner performance outperformed human tutors. Atkinson, Renkl, and Merrill (2003) demonstrated the efficacy of combined fading with the introduction of prompts designed to encourage learners to identify the underlying principle illustrated in each worked-out solution step. Azevedo, Cromley, and Seibert (2004) found that students who used adaptive scaffolds to learn a complex science domain learned better than students learning using fixed scaffolding or no scaffolding.

Research also shows that there are many feasible approaches to adaptive feedback. Ringenberg and VanLehn (2006), for example, found that adaptively providing students with worked out examples for physics problems was as effective as providing them with adaptive hints. Desmarais and Baker (2012) reviewed the learner models that have played the largest roles in the success of such learning environments and the latest advances in the modeling and assessment of learner skills. That said, different contexts afford different approaches, and some domains may be easier to support through adaptivity than others. Effect sizes for adaptive tutoring systems, for example, were the largest for STEM domains compared to humanities domains (Cohen, Kulik, & Kulik, 1982).

Adaptivity has been demonstrated as effective in digital games for learning. van Oostendorp, van der Spek, and Linssen (2014) found that dynamically adapting the challenge level of a serious game about triage significantly increased learning among participants without affecting reported engagement. Sampayo-Vargas, Cope, He, and Byrne (2013) similarly found in a game focusing on Spanish cognates that adaptively adjusting the difficulty of the game led to significant increases in student learning without impacting engagement. Soflano, Connolly, and Hainey (2015) demonstrated the value of adaptivity within educational games in the sense that (a) learning styles identified using questionnaires were not always consistent with the learning styles demonstrated during gameplay and (b) learning styles fluctuated during the learning process in games. Lee,

Rowe, Mott, and Lester (2014) and Rowe and Lester (2015) demonstrated supervised machine learning techniques (e.g., dynamic Bayesian networks) can effectively drive adaptation decisions to support significant improvements in student learning gains and problem-solving. Other research on adaptivity in educational games has found more mixed results. Conati and Manske (2009), for example, evaluated the impact of adaptive feedback on the effectiveness of a pedagogical agent in an educational computer game. They compared a version of the game with no agent to two versions with agents using different student models to guide the agent's interventions. Conati and Manske found no difference in student learning across the three conditions.

Digital games comprise a wide range of mechanics and formats and provide a wide range of options for integrating adaptivity as a result. Lopes and Bidarra (2011) surveyed adaptivity in digital games (both recreational games as well as serious games) and outlined several ways in which games can focus adaptivity: (1) the game world itself and the objects within it can be varied so that the world can be made easier or harder depending on proficiency, (2) the game mechanics in terms of how actions are implemented can be made easier or harder, (3) the attributes and algorithms of the nonplayer characters can be enhanced, (4) the sequence and pacing of the game narratives can be adjusted, and (5) game scenarios can be adapted within the narrative in terms of what a player encounters within a level of the game. The present study builds on existing research by exploring an approach within Lopes and Bidarra's fifth category of game scenarios in the sense of adaptively increasing the conceptual abstractness of self-explanation prompts as a player's proficiency increases.

#### 2.4. The current study

Based on the findings across the prior research on self-explanation in digital games discussed in the previous sections, six major design principles can be distilled concerning effective integration of self-explanation functionality into educational games:

1. Students should be asked to reflect upon correct information.
2. To decrease intrinsic processing load and facilitate feedback, students should be provided with self-explanation "answers" from which to choose as opposed to being required to write open-ended responses.
3. Self-explanations prompts should take into account the intrinsic processing demands of the game itself, which may explain why self-explanation prompts have proven most effective in simpler game-like environments.
4. Self-explanation prompts should be tightly coupled with specific student actions and specific game challenges.
5. Students' levels of engagement can impact the efficacy of the self-explanation functionality.
6. The nature, focus, and abstractness of the self-explanation functionality can affect the students' levels of processing and learning outcomes.

In addition to these six design principles derived from prior research, the current study explores a seventh hypothesized design principle based on research on adaptive functionality.

7. Adaptively adjusting the abstraction of self-explanation functionality might ameliorate the negative aspects of self-explanation functionality observed in prior research while supporting the positive aspects.

The current study explores this seventh hypothetical design principle by comparing three conditions. The comparison is conducted in the context of self-explanation functionality designed in accordance with the first six design principles. In the navigation-only condition, players maneuver through navigational challenges without any self-explanation prompts. In the navigation + abstract condition, these navigational challenges are paired with a self-explanation prompt that focuses on abstract connections between the navigational challenges and Newtonian relationships. In the navigation + adaptive condition, the navigational challenges are paired with self-explanation prompts that adaptively increase from low abstraction (in which the prompts focus concretely on navigational moves) to high abstraction (in which the prompts focus more abstractly on the navigational moves in terms of overarching Newtonian relationships). These three conditions were designed to test three core predictions:

1. Students in the navigation + adaptive condition would demonstrate greater pretest-posttest gains than students in the navigation + abstract condition or the navigation-only condition.
2. Students in the navigation-only condition would progress significantly further in the game than the navigation + abstract condition because the navigation-only students would not require time to interact with the self-explanation functionality, the navigation-only students would not progress significantly further than the navigation + adaptive students because the adaptive design would make the self-explanation functionality less disruptive and more accessible.
3. Students in the navigation + adaptive condition would display patterns in their gameplay behavior indicating deeper conceptual sophistication than the students in the other conditions.

### 3. Methods

#### 3.1. Participants

The current study was conducted with the 7th grade classrooms of two teachers in two different middle schools in the southeastern United States. A total of 210 students, including 124 from the first school and 86 from the second school, assented to participate in the study. Both schools were racially and culturally diverse. In terms of socio-economics, 43% of the students at the first school and 83% of the students at the second school were eligible for free or reduced lunch. Students were randomly assigned to one of three game conditions within each classroom (i.e., each classroom included students in all three conditions). Students were removed from the final data analyses if they did not complete the entire pretest and posttest or if their data were compromised in some other way. A total of 170 students were included in the final analyses, including 97 students from the first school and 73 students from the second school.

The students represented a broad range of video game playing experience outside of school. In the survey administered to the students, students reported playing video games an average of 4.6 h per week with a standard deviation of 3.7 and a range of 0–10 h. These students therefore appear representative of the wide range of experience with digital games that students typically bring to the classroom.

Five days were spent in each school constituting one week of class time. In all classes, students were instructed on how to navigate through initial steps in the game and received help as needed. Students were also encouraged to talk about the game and share strategies with their peers if they got stuck. Thirty minutes before the end of the last class on the final day, students were asked to stop playing and were guided through the post-test and a short engagement survey. Because experimental procedures were similar across schools and students were assigned across all conditions in each classroom, students were aggregated across classrooms and schools for the analyses.

#### 3.2. The Fuzzy Chronicles game and experimental conditions

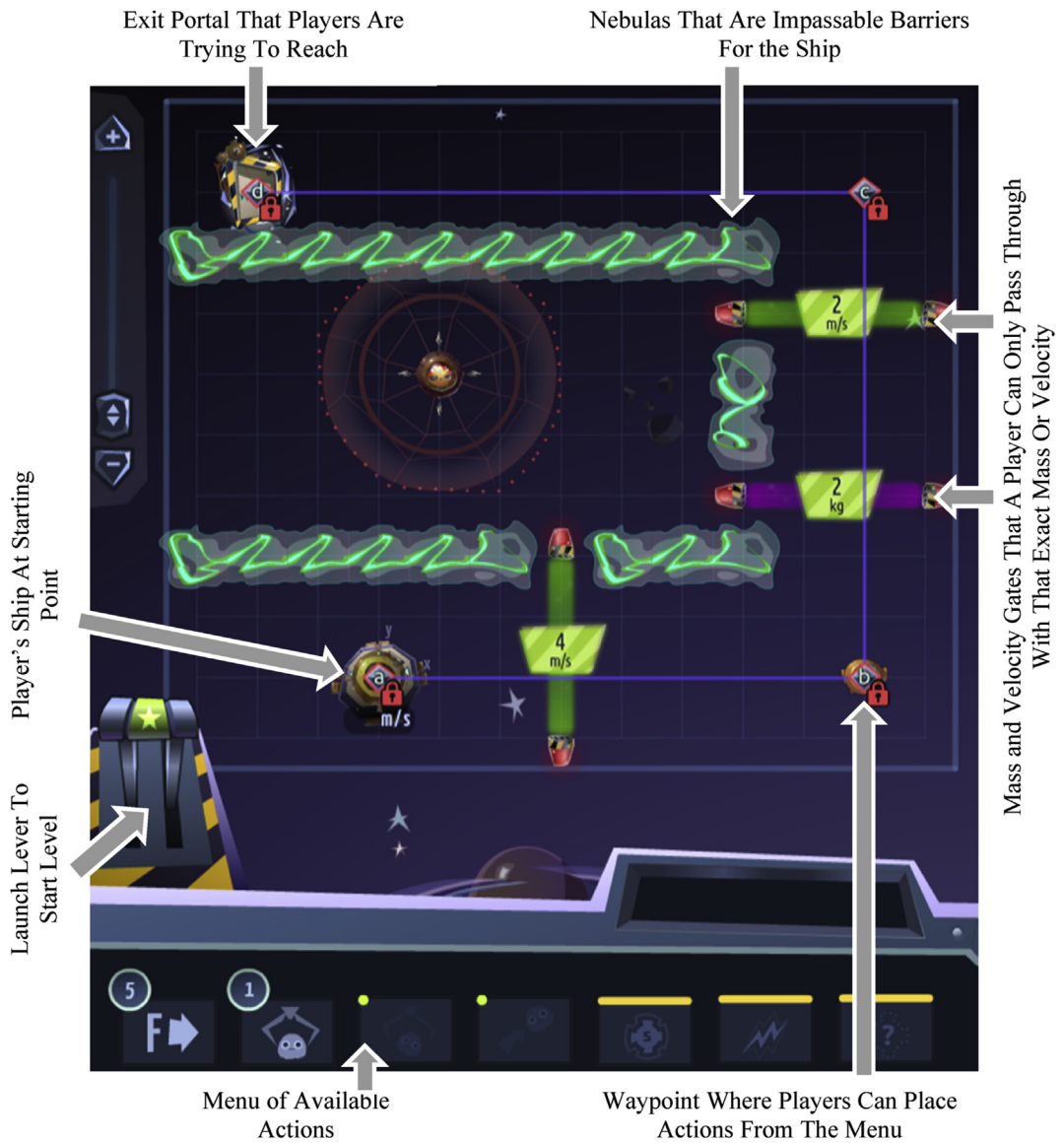
*Fuzzy Chronicles* is a digital game that was developed to support students learning about Newtonian dynamics (Clark, 2012, Clark, Sengupta, Brady, Martinez-Garza, & Killingsworth, 2015, Clark et al., 2016). Students play as the space navigator Surge who must pilot a spaceship through a two-dimensional spatial grid (see Figs. 1 and 2) by placing actions at waypoints along a trajectory. Game play is divided into levels, each constituting a separate navigational challenge. Overall, players complete levels of the game that require them to consider the appropriate magnitude and direction of forces to propel their ships in a desired direction at a desired speed. *Speed gates* in a level require the ship to travel at a specific speed to pass through safely (Fig. 1). A player may also pick up *fuzzies*, which add mass to her/his ship. *Mass gates* allow only ships with a certain amount of mass to pass through safely. Players may also throw fuzzies. Throwing a fuzzy launches the fuzzy in one direction while applying an equal and opposite force to the ship.

Players create a flight plan by placing actions on the map at waypoints (Fig. 2). Actions include forces of various magnitudes and directions as well as other commands to pick up, release, or throw fuzzies. When the player is ready, she/he clicks the launch lever. Surge's ship is then launched. Actions are executed if and when Surge's ship passes through the waypoint where the actions were placed. The plan may or may not be successful, and actions may or may not be executed depending on whether the ship actually reaches the waypoints where the actions were placed. If the player is successful, the next level in the game is unlocked and she/he can proceed. If the user is not successful, the player returns to the planning phase in order to add, change, move, or delete any of the actions. Therefore, players cannot skip ahead.

Players navigate from one level to the next on a *star map* that contains differently colored regions designating progressively more challenging combinations of Newtonian concepts. This study focuses on the first two regions, red and blue, because those are the regions that most students can feasibly complete in one week of class time. Levels in the red region focus on relationships between force and changes in velocity. Players alter the direction and magnitudes of forces on their ships in order to manipulate their speed and direction (i.e., velocity) across the map to reach the exit portal. Levels in the blue region focus on relationships between force, mass, and changes in velocity. Blue levels are similar to red levels with the addition of an emphasis on mass. Players pick up fuzzies along their trajectory. Fuzzies have a mass of 1 kg each and players must adjust their force calculations to account for their increased mass.

The red and blue regions each include four or five *introductory levels*, which introduce the new ideas and mechanics of the region, as well as two *boss levels* at the end, which are more challenging and require the player to integrate across the mechanics and ideas of the region. In addition to the introductory levels and boss levels in a region, a *warp mission* (Fig. 3) is included in each region to separate the introductory levels for the region from the culminating challenge, or boss, levels for the region.

The warp mission, which is located in between the introductory and boss levels, is intended to support the player in mastering the basic ideas of the region before progressing to the boss levels of that region (Fig. 3). Warp mission levels differ from introductory and boss levels in several important ways. First, warp mission levels involve multiple variants of the mission that must be completed before a player can unlock and progress to the boss levels. More specifically, missions for a warp mission level must be completed multiple times until the player has raised her/his overall average mission score for that level from a beginning score of 0 to a score of 80, which indicates a high level of proficiency. Players receive a new, but similar, variant of the warp mission on each attempt. Warp mission levels are thus different from introductory and boss levels because



**Fig. 1.** Anatomy of an EPIGAME level. Players place and modify actions from the menu at various waypoints to create a travel plan to reach the Exit Portal. When the plan is ready, the player pulls the launch lever. Actions are activated when and if the ship reaches the waypoint where they have been placed.

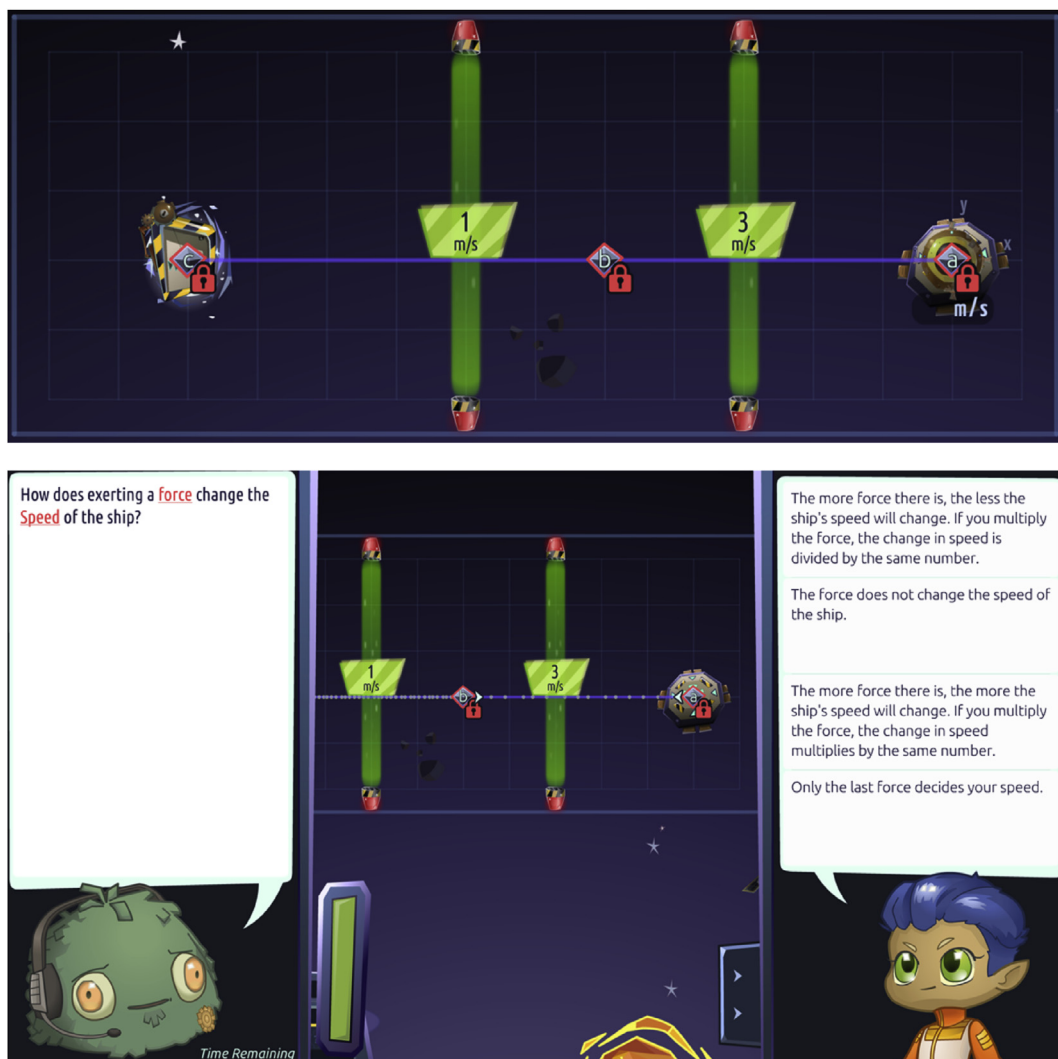


**Fig. 2.** Clicking on a waypoint on the map shows the commands currently at that waypoint (left). Force commands can be added and edited in terms of magnitude and direction (right). Other commands can be added in terms of picking up, dropping, or throwing masses (fuzzies), which, in turn, affect the mass and/or velocity of the player's ship.

introductory and boss levels (a) need to be successfully completed only once to unlock the subsequent level and (b) do not change from one attempt to the next.

Essentially, each time the player attempts a warp mission level, the player receives one of several variants of the mission that involve the same conceptual and navigational challenge but a different layout and orientation. The player's score for the variant is based on the number of attempts needed to solve that variant and its associated explanation prompts, with more attempts resulting in a lower score. The new overall score for the ongoing warp mission level is the average of the player's score on the most recent variant and the player's current overall score for the level. In this way, the overall score for a warp mission level is essentially a running average of scores on the variants of that warp mission level. Students therefore (a) repeatedly encounter variants of the mission until the running average of those scores exceeds a certain threshold and (b) receive feedback on their answers to the self-explanation prompts.

The current study compares three game conditions. The researchers hereafter refer to these conditions simply as nav+adaptive, nav+abstract, and nav-only for brevity. These conditions differed only in terms of the type of the self-explanation functionality, if any, included in the warp mission levels. As described earlier, the nav-only condition does not pose self-explanation prompts after the navigation phase. Nav+abstract and nav+adaptive conditions pose self-explanation prompts targeted at the concepts just encountered in the navigation phase. In the nav+adaptive condition, these explanation prompts begin as "low abstraction" and then increase to "medium abstraction" and eventually to "high abstraction" as the player's running average of scores for that level increases. Low-abstraction prompts and answers focus on articulating the



**Fig. 3.** After the introductory dialogue in a warp mission, students encounter the navigation phase of the warp mission (top). After succeeding in the navigation phase in a warp mission, students encounter the explanation phase of the warp mission (bottom).



concrete specific actions that the student chooses in the navigation phase. An example of a low-abstraction prompt and answers would include:

Prompt: What force did you exert for 0.1 seconds to speed up from 2m/s to 4m/s?

Answer: I exerted a 20N force to speed up by 2m/s.

High-abstraction prompts and answers focus on underlying rules and relationships. An example of a high-abstraction prompt would include:

Prompt: How does exerting a force change the speed of the ship?

Answer: The more force there is, the more the ship's speed will change. If you multiply the force, the change in speed multiplies by the same number.

In the nav+abstract condition, the self-explanation prompts are always “high abstraction.”

### 3.3. Physics assessment

A twenty-one question, multiple choice pre-post test was created to assess physics understanding. The pre-test and post-test were identical. Test questions asked students to determine changes in velocity from the application of various forces as well as to determine the magnitude and direction of forces required to achieve various changes in velocity (Fig. 4 presents an example question).

A Cronbach alpha analysis of test item reliability demonstrated good levels of reliability for the twenty-one items combined (0.834). Individual proposed subsets of questions also demonstrated fair reliability, with questions about force and velocity change focal to the Red levels scoring 0.716 and questions about mass relationships focal to the Blue levels scoring 0.738. Item difficulties were estimated for each of the questions with a two-parameter item response theory model using the *ltm* package in R (Rizopoulos, 2006). A third classroom implementation included six of the questions, and those data were also included for this analysis. In an item response framework, item difficulties were calculated on a latent scale with a practical range of  $-4$  to  $4$ , where an estimated difficulty of  $0$  indicates an average level of difficulty for the sample population. Based on the estimated item difficulties for these runs, it was determined that the average difficulty estimate for the 7 blue mission items (1.79) was higher than the average difficulty estimate of the 7 red mission items (1.48). This finding was not surprising because blue mission questions deal with mass in addition to the concepts present in the red missions. A differential item functioning (DIF) analysis of all post-test items was conducted, and it found no significant difference in item functioning across each of the three experimental conditions.

## 4. Results

Results are first presented within and across experimental conditions in terms of the pre-test, post-test, engagement, and gameplay metrics. Results are then presented within the adaptive condition comparing performance on the low-abstraction, medium-abstraction, and high-abstraction explanation prompts.

### 4.1. Pre-test and post-test scores

A one-way ANOVA of pre-test scores was conducted for the three conditions (Table 1). As expected, there were no significant differences between the three conditions on the pre-test ( $F(2, 132) = 2.248, p = 0.110$ ). Next, a one-way ANOVA of post-test scores was conducted. The one-way ANOVA showed that the difference in post-test scores between nav-only, nav+abstract, and nav+adaptive conditions was significant,  $F(2,132) = 3.215, p = 0.043$ . Tukey's HSD test showed that the post-test scores of students in the nav+adaptive condition were significantly higher than those of the students in the nav+abstract condition.

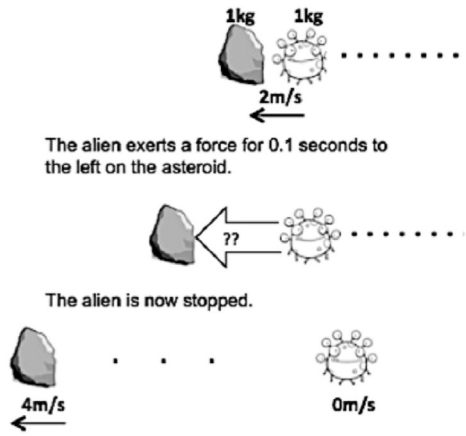
Paired t-tests between pretest and post-test scores for each condition at each school were then run (Table 2). All of these paired t-tests showed significant improvement ( $p < 0.001$ ) from pretest to post-test (nav-only:  $t(43) = -8.898$ , nav+adaptive:  $t(37) = -6.832$ , nav+abstract:  $t(52) = -6.568$ ). The effect sizes for each condition were calculated using Cohen's  $d$  corrected for paired T-tests (nav-only: 1.309, nav+adaptive: 1.151, nav+abstract: 0.904). These findings demonstrated that students in all conditions scored significantly higher on the post-test than the pre-test with a large effect size.

### 4.2. Highest level completed

ANOVAs were conducted to compare the experimental conditions for progress in the game (Table 1). Students could move onward to a subsequent game level only after successfully completing the game level preceding it. For this reason, the highest game level a student completed measured how far the student progressed in the game. A one-way ANOVA showed that the difference in highest level completed between nav-only, nav+abstract, and nav+adaptive conditions was statistically significant,  $F(2,132) = 10.658, p < 0.001$ . Tukey's HSD test showed that the highest level completed was significantly higher in the nav+adaptive and nav-only conditions than in the nav+abstract condition.

**Question 14**

A 1kg alien is holding onto a 1kg asteroid in space. They are moving to the left at 2m/s.



The alien exerts a force for 0.1 seconds to the left on the asteroid.

The alien is now stopped.

How much force was used by the alien to push the asteroid?

Please circle the best possible answer from the options below

A. 20N

B. 4N

C. 2N

D. 40N

Fig. 4. Example assessment question.

**Table 1**  
ANOVAs comparing overall performance metrics by condition.

Performance metric	Nav-only		Nav+Abstract		Nav+Adaptive		F	p	Significant tukey HSDs & effect sizes
	M	SD	M	SD	M	SD			
Pre-test score	3.864	2.216	3.925	2.623	4.974	3.132	$F(2,132) = 2.248$	0.110	–
Post-test score	8.182	3.768	6.981	3.815	8.974	3.745	$F(2,132) = 3.215$	0.043	NA adap > NAbstr Cohen d = 0.53
Highest level completed	19.77	4.377	15.60	4.753	18.08	4.181	$F(2,132) = 10.658$	0.000	NOnly > NAbstr Cohen d = 0.91 NA adap > NAbstr Cohen d = 0.55
Total engagement	9.568	3.323	10.038	3.281	9.421	3.072	$F(2,132) = 0.465$	0.629	–
Average actions used per attempt	5.301	2.422	9.938	2.902	9.203	2.34	$F(2,132) = 42.009$	<0.001	NA adap > NOnly Cohen d = 1.64 NAbstr > NOnly Cohen d = 1.74
Average attempts per level	4.771	4.571	4.963	3.975	3.706	2.417	$F(2,132) = 1.305$	0.275	–

#### 4.3. Engagement

Students completed an engagement survey where they responded to four statements that would indicate their level of engagement in the game on a scale of 1 = Strongly Disagree to 5 = Strongly Agree. Students' scores for their surveys were summed and reported here as engagement (Table 1). Cronbach Alpha for the four items was 0.875. A one-way ANOVA of engagement among the three conditions found that there was no significant difference in engagement across the three conditions,  $F(2, 132) = 0.465$ ,  $p = 0.629$ .

Pearson correlations between total engagement and post-test score and between total engagement and highest level completed for students in each condition were conducted (Tables 3 and 4). None of the correlations between engagement and post-test score were significant (nav-only:  $r = -0.142$ ,  $p = 0.357$ ; nav+abstract:  $r = -0.233$ ,  $p = 0.093$ ; nav+adaptive:  $r = -0.196$ ,  $p = 0.237$ ). This suggests that student engagement was not related to post-test score in any condition. Total engagement, however, was significantly negatively correlated with highest level completed in the nav+abstract ( $r = -0.402$ ,  $p = 0.003$ ) and nav+adaptive ( $r = -0.693$ ,  $p < 0.001$ ) conditions, and the correlation was nearly significant in the nav-only

**Table 2**

Pre-test post-test comparisons within each condition.

Condition	Pretest (SD)	Post test (SD)	p	Effect size (Cohen's d)
Nav-Only	3.864 (2.216)	8.182 (3.768)	<0.001	Post>Pre: 1.309
Nav+Adaptive	4.974 (3.132)	8.974 (3.745)	<0.001	Post>Pre: 1.151
Nav+Abstract	3.925 (2.623)	6.981 (3.815)	<0.001	Post>Pre: 0.904

condition ( $r = -0.276$ ,  $p = 0.07$ ). This suggests that engagement as self-reported by the students was negatively correlated with their highest level completed, particularly for the conditions with the explanation prompts.

#### 4.4. Actions used and attempts per level

Next, two gameplay performance metrics were examined: (a) how many changes students made to their navigation plans on each attempt of a level on average and (b) the average number of attempts used per level completed (i.e., total attempts divided by the total number of levels completed in the game). These metrics provided insight into the degree to which players thought about or understood the underlying relationships in the game (i.e., as opposed to guessing, brute-forcing, or iteratively nudging their navigation plans toward successful configurations).

Average actions per attempt were calculated as the average number of additions, deletions, or changes a player made to her/his navigation plan each time the player attempted or re-attempted a game level. A one-way ANOVA showed that the difference in actions used per attempt between nav-only, nav+abstract, and nav+adaptive conditions was statistically significant,  $F(2, 132) = 42.009$ ,  $p < 0.001$ . Tukey's HSD test showed that students in the nav+adaptive and nav+abstract conditions used significantly more actions for each attempt on average than those in the nav-only condition.

Average attempts per level were calculated as the total number of attempts divided by the highest level in the game successfully completed by a player. A smaller average meant that a player attempted most levels only a few times before succeeding, a higher average number of attempts meant that the player attempted each level more frequently. A one-way ANOVA showed that the difference in the average number of attempts needed to complete each non-warp level between nav-only, nav+abstract, and nav+adaptive conditions was not statistically significant,  $F(2, 132) = 1.305$ ,  $p = 0.275$ .

#### 4.5. Abstraction of prompts within the nav+adaptive condition

Finally, student scores were analyzed within the nav+adaptive condition on the low-abstraction, medium-abstraction, and high-abstraction self-explanation prompts to compare student performance across the levels of abstraction (Table 5). Repeated measures ANOVA ( $F(2, 115) = 10.153$ ,  $p = 0.002$ ) showed that students scored significantly differently across levels of abstraction. This difference was significant for the comparison of high versus low and the comparison of high versus medium, using Bonferroni's correction for multiple comparisons, indicating that students performed better on the high-abstraction prompts than the less abstract ones (High vs. Medium = 9.291,  $p = 0.001$ ; High vs. Low = 5.787,  $p < 0.001$ ). There was no significant difference between the scores for the low-abstraction and medium-abstraction prompts (Medium vs. Low = -3.505,  $p = 0.358$ ).

## 5. Discussion

Students improved significantly from pretest to posttest with a large effect size in all conditions. Prior research demonstrated that taking the pre-test by itself does not lead to increased outcomes on the post-test (Killingsworth, Adams, & Clark, 2016; Martinez-Garza & Clark, submitted). The increased post-test scores can therefore be interpreted as representing learning from the game as opposed to merely a testing effect.

It was anticipated that the nav+adaptive condition, in which students received self-explanation prompts that adaptively increased in abstraction, would demonstrate significantly higher post-test scores than the nav+abstract and nav-only conditions, in which students respectively received only high-abstraction self-explanation prompts or no self-explanation prompts. While the post-test results of the nav+adaptive condition were indeed higher than the results of the nav-only and nav+abstract conditions, this difference was significant only for the nav+abstract condition. This indicates that while the nav+adaptive approach was a significant improvement over the nav+abstract approach, the nav+adaptive was not significantly more effective than the nav-only approach in terms of the post-test results.

Student progression through the game might explain some portion of these patterns. It was expected that (a) the nav-only group would complete significantly more levels in the game than the nav+adaptive or nav+abstract groups because the nav-only group would not have any self-explanation prompts to complete along the way and (b) the nav+adaptive group would complete significantly more levels in the game than the nav+abstract group because the adaptive approach would make the self-explanation prompts more accessible for the nav+adaptive group. These hypotheses proved correct, but only the difference between nav-only and nav+abstract and the difference between nav+adaptive and nav+abstract were significant. While the nav-only students completed more levels in the game than the nav+adaptive students, this difference was not

**Table 3**

Pearson correlations between engagement and post-test score.

Condition	Engagement		Post-test		Pearson correlation	p
	M	SD	M	SD		
Nav-Only	9.568	3.323	8.182	3.768	-0.142	0.357
Nav+Abstract	10.038	3.281	6.981	3.815	-0.233	0.093
Nav+Adaptive	9.421	3.072	8.974	3.745	-0.196	0.237

**Table 4**

Pearson correlations between engagement and highest level completed.

Condition	Engagement		Highest level completed		Pearson correlation	p
	M	SD	M	SD		
Nav-Only	9.568	3.323	19.77	4.377	-0.276	0.070
Nav+Abstract	10.038	3.281	15.60	4.753	-0.402	0.003
Nav+Adaptive	9.421	3.072	18.08	4.181	-0.693	<0.001

**Table 5**

Repeated measures ANOVA of the low, medium, and high-abstraction explanation prompts within the nav-adaptive condition.

	Low-abstraction		Med-abstraction		High-abstraction		F	p	Multiple comparisons (bonferroni corrected)
	M	SD	M	SD	M	SD			
Nav+Adaptive	107.4	6.42	103.9	6.30	113.2	6.69	F(2, 115) = 10.153	0.002	High > Low p < 0.001 High > Medium p = 0.001

significant, suggesting that the time required for the adaptive self-explanation functionality was partially offset by time saved by the nav+adaptive students in the rest of the game.

Perhaps the adaptive self-explanation functionality bolstered learning enough for students to glean useful content insights that allowed them to progress faster through the game. Alternatively, perhaps the adaptive nature of the nav+adaptive condition allowed students to move more quickly through the warp missions, thereby allowing them more time to work on other levels in a manner similar to the advantages of the nav-only condition. Future research should explore these alternatives to understand the underlying relative contributions and mechanisms. Current findings make clear, however, that the nav+adaptive students progressed significantly further than the nav+abstract students. This progress, although not as extensive as the progress of the nav-only students, combined with whatever other pedagogical advantages of the adaptive prompts to support significantly higher performance on post-tests for the nav+adaptive students relative to the nav+abstract students. Relative to the nav-only students, the nav+adaptive students may have scored higher but not significantly higher because the nav-only students' somewhat further progress in the game diluted some of the pedagogical advantage of the self-explanation functionality in terms of the post-test.

Turning to engagement, it was predicted that students in the nav-only condition would report the highest levels of engagement on the survey because it was expected that the self-explanation prompts of the other two conditions would distract from game play. Additionally, it was also expected that the nav+adaptive condition would report higher engagement than the nav+abstract condition because it was assumed that the adaptation of the self-explanation prompts in the nav+adaptive condition would cause the self-explanation functionality to flow more smoothly and naturally for students. Despite these predictions, there were no significant differences in total engagement across conditions. This represents progress from the studies with earlier designs of *Fuzzy Chronicles*, where the designs of the self-explanation prompts were too intrusive and disrupted gameplay (e.g., Adams & Clark, 2014). These findings support the design decisions underpinning the current version of the game: (a) focusing the self-explanation prompts primarily in the warp missions rather than including self-explanation prompts with every level, (b) presenting only one self-explanation prompt after each warp mission navigation phase, (c) creating short and focused navigation phases for each warp mission variant so that the connection between the self-explanation prompt and the navigation phase was clear and obvious, (d) presenting the self-explanation prompt only after the successful completion of the navigation phase so that students reflected only on successful solutions and remained undisturbed in their gameplay, and (e) integrating the self-explanation prompts more deeply into the progression mechanics of the game such that students could not ignore the prompts by randomly clicking on responses.

Exploring the role of engagement more deeply, there were no significant correlations between engagement and post-test score for any condition, but there were significant negative correlations (for nav+adaptive and nav+abstract), and a trend toward significant negative correlation (for nav-only), between engagement and highest level completed. This indicates that as engagement increased, the highest level completed in the game decreased. This seems counterintuitive at first, but might suggest that the students who progressed the furthest through the game tended to be those who were the most focused on the game as a school assignment to maximize.

In terms of other game performance metrics, the average number of actions used per attempt and the average number of attempts per completed game level were counted. It was expected that the nav+adaptive condition would utilize significantly more actions per attempt but fewer attempts per level than the nav+abstract or nav-only conditions, and it was anticipated that the nav+abstract condition would utilize significantly more actions and fewer attempts than the nav-only condition. It was assumed that students who were conceptually best prepared for the game, which was hypothesized to be the nav+adaptive students, would be able to engage in more sophisticated planning for each attempt. It was also anticipated that those students would use a higher number of actions per attempt, and that their navigation plans would also be more successful, ultimately requiring fewer attempts per completed level.

These hypothesized differences parallel the distinctions between model-based reasoning and constraint-based thinking that Parnafes and Disessa (2004) observed in students working with a game-like environment called *NumberSpeed*. Students exercising constraint-based thinking do not try to understand and predict the whole system of relationships in a challenge. Instead, they simply make a few modifications, observe the outcomes of these local modifications, and then make a few more modifications in response. In this way, they “nudge” the plan closer toward a successful solution. Model-based reasoning, however, involves considering the underlying relationships of the challenges in a deep and comprehensive way in order to plan a solution that accounts for those relationships and addresses the challenge more globally as a system. This distinction in strategies is frequently observed in classrooms using SURGE games. It was therefore predicted that the students in the nav+adaptive condition would evidence higher degrees of model-based thinking because it was hypothesized that the nav+adaptive condition would support a deeper understanding of the underlying relationships. It was anticipated that the nav+adaptive condition students would therefore (a) demonstrate higher number of actions used per attempt because students would design more complete solutions as opposed to piecemeal nudging and (b) demonstrate fewer attempts on average per level because of the relatively deeper understanding involved in the nav+adaptive condition compared to the other two conditions.

The findings supported the prediction concerning the average number of actions used per attempt. The nav+adaptive and nav+abstract students utilized a significantly greater number of actions per attempt than the nav-only students, and there was no significant difference in the number of actions used between nav+adaptive and nav+abstract. In terms of attempts per level, the nav+adaptive condition did have the lowest number of attempts per level, but the differences across conditions were not significant.

Finally, student warp mission scores were analyzed within the nav+adaptive condition on the low-abstraction, medium-abstraction, and high-abstraction self-explanation prompts. O'Neil et al. (2014) and Adams and Clark (2014) demonstrated the challenges of incorporating self-explanation prompts that focus on abstract connections and relationships. O'Neil et al. (2014), in particular, uncovered important relationships concerning learning outcomes and the types of prompts with which students choose to engage. Specifically, O'Neil and colleagues argued that self-explanation prompts should not be too simple lest students engage only in minimal processing, nor too abstract as to cause confusion and extraneous processing. Thus, abstraction was desirable albeit with the caveat of potentially overwhelming students.

In interpreting the findings for this study, it is important to remember that students first needed to master the low-abstraction prompts before encountering the medium-abstraction prompts and, likewise, needed to master the medium-abstraction prompts before encountering the high-abstraction prompts. It is also important to remember that the navigation phases accompanying each self-explanation prompt were isomorphic for the low, medium, and high-abstraction prompts. As stated previously, the score for a warp mission variant is the average of the scores on the navigation and explanation phases of that mission variant.

The paired t-tests demonstrate that students in the nav+adaptive condition earned significantly higher scores on variants with high-abstraction prompts than on variants with medium-abstraction or low-abstraction prompts. This was unexpected because it was anticipated that prompts with higher levels of abstraction would be more difficult than prompts with lower levels of abstraction (e.g., O'Neil et al., 2014). Specifically, it was expected that the low-abstraction scores would be significantly higher than the medium-abstraction scores, and that the low and medium abstraction scores would be significantly higher than the high-abstraction scores. There are, however, several possible explanations for the results.

An optimistic explanation might be that the less abstract prompts effectively scaffolded students' understandings for the more abstract prompts. From this perspective, students were prepared for the high-abstraction variants by the time they reached them. A more general explanation, however, might be that students also became better at answering prompts in general. A third explanation might be that students improved on the navigation components of the variants, which also influenced score. All three of these factors likely contribute. Regardless of the relative contributions, however, students' higher scores on the high-abstraction variants suggest that adaptively increasing the abstraction of the self-explanation prompts was an effective approach to providing high-abstraction prompts in an accessible manner.

### 5.1. Caveats and limitations

It is reasonable to argue whether or not the self-explanation functionality in the current study should be considered solely as self-explanation functionality. As described, students needed to achieve a certain score on the warp mission before being allowed to move on to the subsequent challenge levels for that section of the game. The goals involved (a) minimizing incentives for “gaming the system” in a negative sense to circumvent the self-explanation functionality (Baker et al., 2008) and (b) integrating the self-explanation prompts more fully into the “game atoms” and structures of the game itself rather than

inserting them awkwardly outside the “atoms” of the game as an artificial appendage (Echeverria, Barrios, Nussbaum, Amestica, & Leclerc, 2012).

For this study, each time a player attempted a warp mission, she/he received one of several variants of that mission that involved the same challenge conceptually and navigationally but in a different layout and orientation. The overall score for the warp mission level was the running average of the player's most recent score on the most recent variant and the player's previous overall score for the warp mission level prior to that variant (i.e., essentially a running average for that warp mission level). Students therefore (a) repeatedly encountered variants of the mission until the running average of their variant scores exceeded a certain threshold and (b) received feedback on their answers to the prompts. This is very different from Chi's initial approach to self-explanation, in which students received no feedback and were merely encouraged to answer the prompt with some form of explanation (e.g., Chi et al., 1994).

It was decided to integrate the self-explanation functionality in the current study into coherent game atoms such that students could not simply ignore the prompts by randomly clicking on responses. Following from the findings of relevant research, additional principles in the design of the self-explanation functionality were added: (a) confining the self-explanation prompts to the warp missions rather than including self-explanation prompts with every level, (b) presenting only one self-explanation prompt after each warp mission navigation phase, (c) creating short and focused navigation phases for each warp mission variant so that the connection between the self-explanation prompt and the navigation phase is clear and obvious, and (d) presenting the self-explanation prompt only after the successful completion of the navigation phase so that students reflect on only successful solutions and do not have their gameplay disrupted. In this manner, the approach builds upon the evolving research on integrating self-explanation prompts into digital games. It is possible, however, that new terminology is needed to differentiate the approach from the original intended meaning of self-explanation.

A second question to consider involves the potential generalizability of the approach. Many game genres are reported in the literature, such as action, puzzle, role playing, and many others. Each game genre has different game elements and functionalities. Furthermore, the disciplinary focus in the current study, Newtonian mechanics, is only one of a huge range of possible foci. The self-explanation functionality approach explored in the current study is only of value to the extent that it generalizes to some extent across some subset of games genres and/or disciplinary topics. In terms of game genres, generalizability appears promising mechanically because the approach is relatively modular; it focuses on the addition of dialogue interactions after key short levels in the game progression. Therefore, the approach is mechanically feasible across many genres including action, puzzle, role-playing, and others. The much bigger challenge for implementation, however, involves integrating these dialogue interactions in a manner that does not disrupt the flow of gameplay for the player. As outlined earlier, research on *Fuzzy Chronicles* with an earlier version of the self-explanation functionality by Adams and Clark (2014) demonstrated that the design of the self-explanation functionality in that earlier version was too overbearing and intrusive in terms of its timing, frequency, and structure. As a result, that version of the self-explanation functionality disrupted gameplay flow, gameplay progress, and learning. Thus generalizability across genres is mechanically very feasible, but requires careful design in terms of its structure, timing, and frequency as well as its narrative integration.

The second aspect of generalizability involves generalizability across disciplinary topics and the learning goals the game designers intend for the game. The self-explanation functionality explored in the current study and in previous research (e.g., Hsu et al., 2012; Moreno & Mayer, 2005) assumes that the learning goals for the game involve having the players come to understand and articulate relationships that they are exploring through their gameplay. Furthermore, the self-explanation approach outlined assumes that the designers can articulate normative and non-normative relationships that players are likely to perceive so that appropriate dialogue prompts and options can be crafted within the system such that players will find the dialogue options relevant and meaningful in terms of their own perceptions as they reflect on their gameplay. Thus in terms of disciplinary topics, the self-explanation approach outlined would appear to generalize beyond Newtonian mechanics to other domains that are well-structured in terms of their underlying causal, logical, or mathematical relationships, but further research would be required to explore generalizability to less structured domains or less structured learning goals.

## 6. Conclusion

Students in all conditions improved from pre-test to post-test with a large effect size. This suggests that the overall design of the *Fuzzy Chronicles* game itself is improving over the development and refinement iterations compared to earlier studies and pilot studies with *Fuzzy Chronicles* (Adams & Clark, 2014, Clark et al., 2015). At the most basic level, this outcome demonstrates the importance of design beyond medium in developing digital games for learning as highlighted in an earlier meta-analysis of digital games, design, and learning (Clark, Tanner-Smith, & Killingsworth, 2016). In addition to overall refinement of level and concept sequences, control schemes, and user interface, the current version of *Fuzzy Chronicles* instantiated the features demonstrated as critical for self-explanation functionality in games based on the findings of prior research on games and self-explanation (e.g., Adams & Clark, 2014; Johnson & Mayer, 2010; Mayer & Johnson, 2010; Moreno & Mayer, 2005).

In addition, the current version of *Fuzzy Chronicles* integrated the self-explanation prompts into the structure of the game such that they affected score and progress in the game. As discussed in the Caveats and Limitations section, this structure arguably transforms the functionality beyond the original conceptions of research on self-explanation (Chi et al., 1989, 1994). The functionality addresses two concerns, however, by (a) minimizing incentives for “gaming the system” in a negative sense to circumvent the self-explanation functionality (Baker et al., 2008) and (b) integrating the self-explanation more fully into

the “game atoms” and structures of the game itself rather than inserting them awkwardly outside the “atoms” of the game as an artificial appendage (Echeverria et al., 2012). Importantly, the engagement survey demonstrated no significant overall differences across conditions, which suggests relatively successful integration from that perspective.

In terms of the focal research question, an approach to adaptively increasing the abstraction of the self-explanation prompts encountered by students in the nav+adaptive condition was explored. As discussed, the high-abstraction prompts focused on key Newtonian relationships underlying gameplay, while the low-abstraction prompts focused concretely on the student's actions that led to a successful solution of the specific navigational challenge just completed. The findings demonstrate that this approach resulted in significantly higher post-test scores in the nav+adaptive condition relative to the nav+abstract condition, which included the same number of prompts overall. Furthermore, the analyses of students' scores in the nav+adaptive condition show significantly higher scores on average on the high-abstraction prompts than on the low or medium-abstraction prompts. This suggests that the adaptive structure, beginning with low-abstraction prompts and then increasing abstraction as student readiness increases, effectively scaffolds students to mastering the high-abstraction prompts.

Adams and Clark (2014) and O'Neil et al. (2014) concluded that, along with generative processing, self-explanation prompts can result in (a) extraneous processing by slowing down the player and distracting them from learning or (b) minimal processing if learners ignore the prompts and return to gameplay as quickly as possible. O'Neil et al. further concluded that high levels of abstraction in the self-explanation functionality may increase processing demands too substantially. The adaptive features: (a) resulted in significantly greater post-test scores than the non-adaptive high-abstraction functionality, (b) scaffolded students smoothly up to the high-abstraction prompts by focusing on the underlying Newtonian relationships, and (c) may have minimized adverse impact of self-explanation prompts in a manner allowing students to progress significantly further in the game than students in the non-adaptive nav+abstract condition.

While the post-test results of the nav+adaptive condition were higher than in the nav-only and nav+abstract conditions, this difference was significant only with regard to the nav+abstract condition. Thus, while the nav+adaptive approach was a significant improvement over the nav+abstract approach, the nav+adaptive was not significantly better than the nav-only approach in terms of the post-test results. This might be connected to the finding that the nav-only students still progressed further, although not significantly further, in the game than the nav+adaptive condition.

Looking beyond the post-test scores, however, the nav+adaptive and nav+abstract conditions elicited different game play behaviors than the nav-only condition. Specifically, students in the conditions with the self-explanation functionality used a higher average number of actions per attempt than did students in the nav-only condition. This suggests that the students in the conditions with the self-explanation functionality may have engaged in higher degrees of model-based thinking than the students in the nav-only condition (Parnafes & Disessa, 2004). If the self-explanation functionality is indeed fostering higher-degrees of model-based thinking, it might be assumed that the self-explanation functionality is providing additional advantages uncaptured by the current post-test format.

## Acknowledgements

Core development of the digital game at the heart of this study was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A110782 to Vanderbilt University. Data collection and analysis was supported through that grant and grant 1119290 from the National Science Foundation 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

- Adams, D. M., & Clark, D. B. (2014). Integrating self-explanation functionality into a complex game environment: Keeping gaming in motion. *Computers & Education*, 73, 149–159.
- Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive science*, 26(2), 147–179.
- Anderson, J., & Rainie, L. (2012). Gamification: Experts expect 'game layers' to expand in the future, with positive and negative results. *Pew Internet & American Life Project*. Available at: <http://www.pewinternet.org/2012/05/18/the-future-of-gamification/>.
- Atkinson, R. K., Renkl, A., & Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, 95(4), 774.
- Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, 29(3), 344–370.
- Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in gaming the system behavior in interactive learning environments. *Journal of Interactive Learning Research*, 19(2), 185–224.
- Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology*, 5, 161–238.
- Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145–182.
- Chi, M. T. H., DeLeeuw, N., Chiu, M., & LaVanher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Chi, M. T. H., & VanLehn, K. A. (1991). The content of physics self-explanations. *Journal of the Learning Sciences*, 1(1), 69–105.
- Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction*, 21(1–2), 137–180.

- Clark, D. B. (2012). *Designing games to help players articulate productive mental models*. Keynote commissioned for the Cyberlearning Research Summit 2012 hosted by SRI International, the National Geographic Society, and the Lawrence Hall of Science with funding from the National Science Foundation and the Bill and Melinda Gates Foundation. Washington, DC <http://www.youtu.be/xlMfk5rP9yl>.
- Clark, D. B., Sengupta, P., Brady, C. E., Martinez-Garza, M. M., & Killingsworth, S. S. (2015). Disciplinary integration of digital games for science learning. *International Journal of STEM Education*, 2(1), 1–21.
- Clark, D. B., Tanner-Smith, E., & Killingsworth, S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, 86(1), 79–122. <http://dx.doi.org/10.3102/003465431558206>. <http://rer.sagepub.com/content/early/2015/10/20/0034654315582065.full.pdf+html>.
- Clark, D. B., Virk, S., Sengupta, P., Brady, C., Martinez-Garza, M., Krinks, K., et al. (2016a). SURGE's evolution deeper into formal representations: The siren's call of popular game-play mechanics. *International Journal of Designs for Learning*, 7(1), 107–146. <https://scholarworks.iu.edu/journals/index.php/ijdl/article/view/19359>.
- Cohen, P. A., Kulik, J. A., & Kulik, C. L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19(2), 237–248.
- Conati, C., & Manske, M. (2009). Evaluating adaptive feedback in an educational computer game. In *International workshop on intelligent virtual agents* (pp. 146–158). Springer Berlin Heidelberg.
- Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1–2), 9–38.
- Echeverria, A., Barrios, E., Nussbaum, M., Amestica, M., & Leclerc, S. (2012). The atomic intrinsic integration approach: A structured methodology for the design of games for the conceptual understanding of physics. *Computers & Education*, 59(2), 806–816.
- Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). New York, NY: Palgrave Macmillan.
- Hausmann, R. G., & Chi, M. H. (2002). Can a computer interface support self-explaining. *Cognitive Technology*, 7(1), 4–14.
- Hausmann, R. G., & VanLehn, K. (2007). Explaining self-explaining: A contrast between content and generation. *Frontiers in Artificial Intelligence and Applications*, 158, 417.
- Honey, M. A., & Hilton, M. (2010). *Learning science through computer games and simulations*. National Research Council. Washington, DC: National Academy Press.
- Hsu, C.-Y., Tsai, C.-C., & Wang, H. Y. (2012). Facilitating third graders' acquisition of scientific concepts through digital game-based learning: The effects of self-explanation principles. *The Asia-Pacific Education Researcher*, 21(1), 71–82.
- Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, 26, 1246–1252.
- Killingsworth, S., Adams, D., & Clark, D. B. (2016). Learning, cognition, and experimental control in an educational physics game. In R. Lamb, & D. McMahon (Eds.), *Educational and learning games: New research* (pp. 31–54). New York, NY: NOVA Publishing.
- Killingsworth, S., Clark, D. B., & Adams, D. (2015). Self-explanation and explanatory feedback in games: Individual differences, gameplay, and learning. *International Journal of Education in Mathematics, Science and Technology*, 3(3), 162–186. [http://ijemst.com/issues/3\\_3\\_1\\_Killingsworth\\_Clark\\_Adams.pdf](http://ijemst.com/issues/3_3_1_Killingsworth_Clark_Adams.pdf).
- Lee, S. Y., Rowe, J. P., Mott, B. W., & Lester, J. C. (2014). A supervised learning framework for modeling director agent strategies in educational interactive narrative. *IEEE Transactions on Computational Intelligence and AI in Games*, 6(2), 203–215.
- Lenhart, A., Kahne, J., Middaugh, E., Macgill, A., Evans, C., & Vitak, J. (2008). Teens, video games, and civics. *Pew Internet & American Life Project*, 64. Retrieved from <http://www.pewinternet.org/Reports/2008/Teens-Video-Games-and-Civics.aspx>.
- Lester, J. C., Stone, B., & Stelling, G. (1999). Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction*, 9, 1–44.
- Lopes, R., & Bidarra, R. (2011). Adaptivity challenges in games and simulations: A survey. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(2), 85–99.
- Martinez-Garza, M., Clark, D. B., & Nelson, B. (2013). Digital games and the US National Research Council's science proficiency goals. *Studies in Science Education*, 49, 170–208. <http://dx.doi.org/10.1080/03057267.2013.839372>.
- Martinez-Garza M. and Clark D.B. (submitted). Data-mining epistemic stances from raw game-play data.
- Mayer, R. E., & Johnson, C. I. (2010). Adding instructional features that promote learning in a game-like environment. *Journal of Educational Computing Research*, 42(3), 241–265.
- Moreno, R., & Mayer, R. E. (2005). Role of guidance, reflection, and interactivity in an agent-based multimedia game. *Journal of Educational Psychology*, 97(1), 117–128.
- van Oostendorp, H., van der Spek, E. D., & Linssen, J. (2014). Adapting the complexity level of a serious game to the proficiency of players. *European Alliance for Innovation Endorsed Transactions on Serious Games*, 1(2), 1–8.
- O'Neil, H. F., Chung, G. K. W. K., Kerr, D., Vendlinks, T. P., Buschang, R. E., & Mayer, R. E. (2014). Adding self-explanation prompts to an educational computer game. *Computers in Human Behavior*, 30, 23–28.
- Parnafes, O., & Disessa, A. (2004). Relations between types of reasoning and computational representations. *International Journal of Computers for Mathematical Learning*, 9(3), 251–280.
- Reif, F., & Scott, L. A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. *American Journal of Physics*, 67(9), 819–831.
- Ringenberg, M. A., & VanLehn, K. (2006). Scaffolding problem solving with annotated, worked-out examples to promote deep learning. In *International conference on intelligent tutoring systems* (pp. 625–634). Springer Berlin Heidelberg.
- Rizopoulos, D. (2006). Itm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. <http://www.jstatsoft.org/v17/i05/>.
- Rowe, J. P., & Lester, J. C. (2015). Improving student problem solving in narrative-centered learning environments: A modular reinforcement learning framework. In *International conference on artificial intelligence in education* (pp. 419–428). Springer International Publishing.
- Roy, M., & Chi, M. T. H. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 271–286). New York: Cambridge University Press.
- Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers & Education*, 69, 452–462.
- Soflano, M., Connolly, T. M., & Hainey, T. (2015). Learning style analysis in adaptive GBL application to teach SQL. *Computers & Education*, 86, 105–119.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
- Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, 105, 249–265. <http://dx.doi.org/10.1037/a0031311>.
- Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B., et al. (2012). Our princess is in another castle: A review of trends in serious gaming for education. *Review of Educational Research*, 82, 61–89. <http://dx.doi.org/10.3102/0034654312436980>.