International Journal of Education in Mathematics, Science and Technology (IJEMST)

www.ijemst.com

Self-Explanation and Explanatory Feedback in Games: Individual Differences, Gameplay, and Learning

Stephen S. Killingsworth[1], Douglas B. Clark[1], Deanne M. Adams[2]
[1]Vanderbilt University
[2]University of Notre Dame

To cite this article:

Killingsworth, S.S., Clark, D.B., & Adams, D.M. (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.

IJEMST International Journal of Education in Mathematics, Science and Technology

Volume 3, Number 3, July 2015, Page 162-186 ISSN: 2147-611X

Self-Explanation and Explanatory Feedback in Games: Individual Differences, Gameplay, and Learning

Stephen S. Killingsworth[1*], Douglas B. Clark[1], Deanne M. Adams[2]
[1]Vanderbilt University
[2]University of Notre Dame

Abstract

Previous research has demonstrated the efficacy of two explanation-based approaches for increasing learning in educational games. The first involves asking students to explain their answers (self-explanation) and the second involves providing correct explanations (explanatory feedback). This study (1) compared self-explanation and explanatory feedback features embedded into a game designed to teach Newtonian dynamics and (2) investigated relationships between learning and individual differences. The results demonstrated significant learning gains for all conditions. There were no overall differences between conditions, but learning outcomes were better for the self-explanation condition after controlling for the highest level completed by each student. Analyses of individual differences indicated that certain threshold inhibitory control abilities may be necessary to benefit from the self-explanation in games.

Key words: Educational games, Self-explanation, Physics education, Individual differences, Inhibitory control.

Introduction

Research on self-explanation by Chi and others has provided insight into the value of explanation for learning (e.g., Chi, Bassok, Lewis, Reimann, & Glaser 1989; Roy & Chi, 2005; Chi & VanLehn, in press). A recent review of research reports that self-explanation results in average learning gains of 20% to 44% compared to control conditions without self-explanation (Roy & Chi, 2005). This emphasis on explanation is mirrored in research on science education. Work by White and Frederickson (1998, 2000), for example, demonstrates the value of asking students to reflect on their learning during inquiry with physics simulations. Similarly, a growing body of research and scholarship on games and cognition emphasizes informal cycles of prediction, explanation, and refinement at the core of game-play processes (Salen & Zimmerman, 2004, Wright, 2006). We have found, however, that implementing self-explanation in educational games requires careful consideration of the specific affordances and constraints of digital games as a medium and careful evaluation of the relationships between individual abilities, gameplay, and learning outcomes.

Two explanation-based approaches have proved effective for increasing learning in educational games: asking students to explain their answers (self-explanation) and providing students with an explanation (explanatory feedback). The present study includes two versions of self-explanation (partial and full) and one version of explanatory feedback. Given overall similarities observed between the partial and full self-explanation conditions, we collapse across the self-explanation conditions in our analyses. This study explores the following questions:

- What are the relative advantages of self-explanation and explanatory feedback for middle school students playing a game covering challenging concepts in Newtonian dynamics?
- How do students' gameplay behaviors relate to game levels completed and learning outcomes?
- How do students' attentional control abilities relate to gameplay behaviors, game levels completed, motivation, and learning outcomes?

* Corresponding Author: *Stephen S. Killingsworth, s.killingsworth@vanderbilt.edu*

The following section begins by introducing the digital game used in the study. Subsequently, we discuss background research on self-explanation and the rationale for an individual differences approach. The introduction closes with an overview of the present study.

Background: Game Context

The *Fuzzy Chronicles*[†] is a game designed to support middle school students learning about Newtonian dynamics (i.e., Newtonian relationships describing motion physics learning). The *Fuzzy Chronicles* (Figure 1) and other early SURGE games are "conceptually-integrated games" (Clark & Martinez-Garza, 2012), in which the science to be learned is integrated directly into the game mechanics (Clark & Martinez-Garza, 2012). Specifically, interactions with game elements and the conditions for successful play are directly connected to concepts in Newtonian dynamics, rather than being introduced through embedded activities isolated from primary gameplay (i.e., during particular game phases or at locations in a game-world). The latter structure is typical of many virtual worlds designed for science learning[‡].

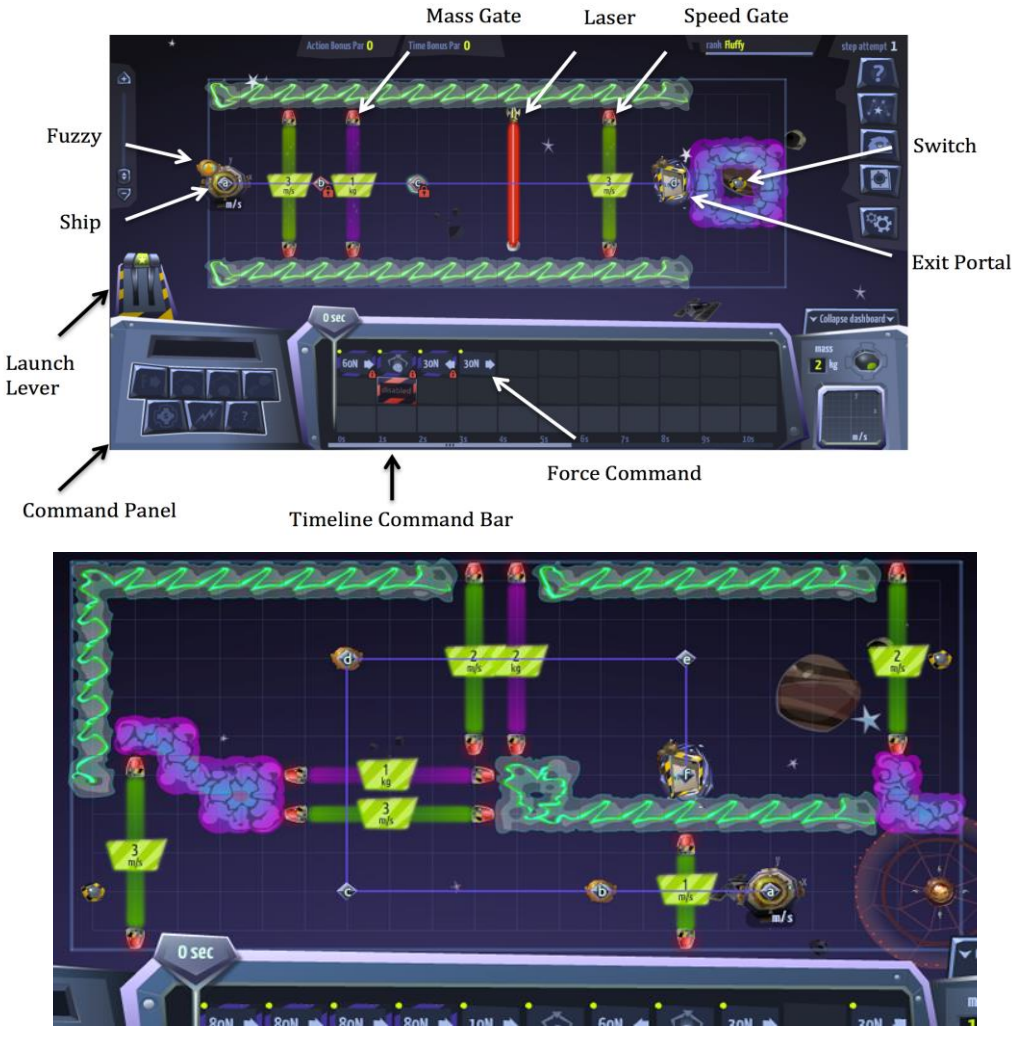

Figure 1. Key parts of a *Fuzzy Chronicles* level with labels (above) and a challenge level (below)

[†] This study employed an early version of The Fuzzy Chronicles. Newer versions of The Fuzzy Chronicles and other SURGE games may be played at www.surgeuniverse.com.

[‡] From our research with the conceptual integration approach used in The Fuzzy Chronicles, we have more recently adopted an approach that we term "disciplinary integration" – see Clark, Sengupta, Brady, Martinez-Garza, & Killingsworth (2015) for a full discussion of the rationale for conceptual integration, disciplinary integration, and the shift from one to the other in our research and designs.

In each level of The *Fuzzy Chronicles*, players must navigate a spaceship around obstacles to reach an exit portal by placing commands on a timeline specifying the magnitude and direction of forces that the ship should apply to achieve the desired path (see Figure 1). A grid is shown within the simulation space of the game (representing distance intervals of one meter). Certain levels also contain Fuzzies (masses of 1 kilogram) that the player may pick up, release, and throw (see Figure 1). The design emphasizes prediction instead of reaction through (a) challenges requiring fewer but higher-impact decisions and (b) requiring players to place force commands before viewing their effects. On each trial the student presses a "launch" lever to view the results of her plan. Students can then revise their plan and re-launch in a new trial.

Background: Self-Explanation in Digital Games

Well-designed games must encourage generative processing (Mayer & Johnson, 2010) to ensure that players make connections between gameplay and formal learning concepts. Unfortunately, few games provide direct supports for generative processing (such as structures for externalizing and reflecting on game-play). More often, articulation and reflection occur outside the game, through discussion among players or participation in online forums (Gee, 2007; Squire, 2005; Steinkuehler & Duncan, 2008). Self-explanation has been proposed as a possible device to encourage generative processing during educational gameplay. According to Roy and Chi (2005), the process of self-explanation can encourage four key forms of cognition, including: (1) recognizing what information is missing while generating inferences, (2) integrating the information taught within a lesson, (3) integrating information from long-term memory with new information, and (4) identifying as well as correcting information. Overall, self-explanation can encourage students to engage in meta-cognitive activities to monitor what they do and do not understand about the material.

Self-explanation has been shown to produce average learning gains of 22% for learning from text, 44% for learning from diagrams, and 20% for learning from multimedia presentations (Roy & Chi, 2005). Despite these successes, implementing self-explanation in educational games has not shown such clear benefits. O'Neil et al. (2014) provide certain possible reasons for this, arguing that in addition to generative processing, self-explanation prompts could also result in extraneous processing by slowing down and distracting the player. O'Neil et al. (2014) also raise the issue that students might respond quickly to avoid deeper processing during self-explanation segments of a game in order to return to the active gameplay quickly.

A study by Moreno and Mayer (2005) demonstrates the need for careful attention to how self-explanation features are implemented in a game environment. The authors conducted a study examining how multiple features (including interactivity and self-explanation) affected learning from a simulation game, *Design-a-Plant* (Lester, Stone, & Stelling, 1999). In the first experiment, students selected answers to questions about the types of roots, stems, and leaves that allowed a plant to survive on a planet. One group of students engaged in self-explanation (asked to provide an explanation for the answer to the plant question) and the other group did not. The authors observed no benefit of self-explanation.

In a second experiment, Moreno and Mayer (2005) factorially manipulated interactivity and self-explanation features. The program provided answers to the questions in the non-interactive condition and students provided answers in the interactive condition. Critically, the authors found an interaction between interactivity and self-explanation for far-transfer measures. Specifically for the non-interactive condition, far-transfer scores were higher for students who engaged in self-explanation. For students with the non-interactive game, far-transfer scores did not differ based on self-explanation. Generally, based on the first two experiments, the authors argued that certain levels of interactivity may already facilitate students' organizing and integrating information at a high level – such that self-explanation does not support further processing. Notably, however, other forms of instruction have also shown limited the benefits of self-explanation because the instructional material sufficiently covers the focal topics (see Matthews and Rittle-Johnson, 2009). Thus, interactivity may be just one of several features of an instructional environment that can minimize the benefits of self-explanation.

In the final experiment by Moreno and Mayer (2005), interactivity and self-explanation were again manipulated factorially. The authors also added a condition in which students choose their own answers, but then received the correct answer before engaging in self-explanation (interactivity+correct-reflection). This allowed the authors to separate the confounded effects of interactivity from the effects of reflecting on a potentially incorrect answer (when explaining one's own answer). The authors observed an overall benefit of self-explanation. The authors also observed findings in the quality of students' explanations, which can be explained by how quickly students were guided to the correct answer in each condition. Students gave the lowest proportion of incorrect explanations in the no-interactivity+correct-reflection condition, in which the program provided the correct

answer immediately. Incorrect explanations were more frequent in the interactivity+correct-reflection, in which the correct answer was provided only after students had potentially provided an incorrect answer. Incorrect explanations were the most frequent in the interactivity+self-reflection condition, in which students had an opportunity both to provide and then potentially reflect upon an incorrect answer.

Altogether, the results of the Moreno and Mayer (2005) study suggest that self-explanation may not always benefit students in an interactive game environment, but further work is necessary to isolate the effects of self-explanation (reflection) from the effects of whether students reflect upon correct or incorrect information. Given the advantages for reflecting on correct information, one way to simplify the self-explanation process that still provides feedback for students' responses is to provide learners with a set of explanation options ("selected-explanation"). Using a game-like environment for instruction on electrical circuits, Johnson and Mayer (2010) found that gains from having students generate their own explanations were equivalent to gains with a base version of the game (without self-explanation), but that having students select an explanation led to higher performance on a transfer level of the game (see also Mayer & Johnson, 2010). As noted by O'Neil et al. (2014), one potential issue is that self-explanation may result in extraneous processing. Students may not be able to explicitly state the correct reasons for giving a particular answer. Providing students with possible explanations as well as feedback can decrease incorrect thinking and reduce extraneous processing. Furthermore, from a design standpoint, it is far simpler to provide quick and effective feedback for selected-explanation responses than for open-ended self-explanation.

Three major recommendations that can be distilled from the literature to date on self-explanation and education games are: 1) students must be asked to reflect upon correct information, 2) self-explanation prompts must take into account the intrinsic processing demands of interacting with a game, 3) providing the students with selection-based self-explanation questions instead of open-ended responses may decrease intrinsic processing load and facilitate feedback.

Background: Cognitive Abilities and Attentional Control

Individual differences in cognitive abilities may have dramatic consequences for whether and how a student benefits from a given instructional practice. In numerous studies, individual differences in abilities have been shown to predict different learning outcomes for given instructional designs (e.g. Fuchs et al., 2014; Höffler & Leutner, 2011; Wiley, Sanchez, & Jaeger, 2014). Thus far, though close examinations of individual differences in strategies and behaviors have received some attention, relationships with measured cognitive abilities have received minimal attention within the literature on self-explanation or within educational games research. Including cognitive ability measures in the present study can help determine who benefits from self-explanation and clarify the relationships between cognitive processing and study outcomes.

The current study focuses on inhibitory control as an individual difference. Inhibitory control is thought to be one important aspect of the set of diverse frontal lobe processes called executive functions (see Miyake et al., 2000). We measure inhibitory control using the Attention Network Test (see Fan et al., 2002), which is thought to measure three ways in which people voluntarily control visual attention to resolve conflicting responses to information (inhibitory control), to select locations of meaningful information in the environment (attentional orienting), and to prepare attention for expected meaningful events and maintain the prepared state (attentional alerting). Inhibitory control (IC) reflects students' abilities to inhibit responses to information that may otherwise interfere with performance. IC may reflect abilities to inhibit pre-potent (or "default") responses and to ignore conflicting information. IC is thought to be one element of the larger set of cognitive processes called executive functions. Executive functions are responsible for regulating thought, action, and emotion based on one's current goals (Blair & Razza, 2007; Miyake et al., 2000).

Given that there is little available research from which to construct particular theories about the relationships between measures in the Attention Network Test (ANT), gameplay, and learning, our analyses of IC in the present study are exploratory. There is good reason, however, to expect that measures in the ANT (and specifically measures of IC) are meaningful for science learning, are relevant for learning from gameplay, and may be specifically relevant for learning from self-explanation. Best et al. (2009) propose that executive functions impact learning outcomes in science (and other domains) either directly or through more complex mental operations and classroom behaviors. Gropen and colleagues propose that as early as preschool, executive functions may be critically important to developing hypothesis testing and abstract reasoning processes supporting conceptual change in science education (Gropen, Clark-Chiarelli, Hosisington, & Ehrlich, 2011). In late elementary and middle school, IC predicts scores on English, mathematics, and science assessments (St.

Clair-Thompson & Gathercole, 2006) and overall semester grades (Visu-Petra et al., 2011). IC has also shown important relationships with fluid intelligence (Unsworth, Spillers, & Brewer, 2009), suggesting general relevance of these processes for academic achievement.

Although we are unaware of any research directly linking self-explanation and IC, one possibility for how these constructs are linked is as follows. First, self-explanation may benefit learning through prompting students to engage in metacognition that they otherwise would not (e.g., McNamara & Magliano, 2009; Roy & Chi, 2005) – though there has been some difficulty measuring individuals' abilities to engage in metacognition (see McNamara & Magliano, 2009). Second, metacognitive processes are likely to be supported by IC (Fernandez-Duque, Baird, & Posner, 2000; Flemming & Dolan, 2012) – perhaps specifically because inhibition may facilitate students' abilities to hold recently encountered content in working memory and to ignore current sensory information and focus on potentially relevant aspects of the current topic. Thus, it is reasonable to propose that individuals with better inhibitory control abilities (1) may be better at using self-explanation techniques and/or (2) may benefit more from engaging in self-explanation.

Because relationships between cognitive abilities and gameplay have received scant attention in educational games research, possible relationships with gameplay are speculative. Nevertheless, we propose that inhibitory control may be relevant for productive play. Gropen et al. (2011) suggest that hypothesis testing and ignoring experiential defaults are important features of science learning that are supported by IC. We expect these benefits should also be relevant for developing abstract principles and rules that guide gameplay. Moreover, games often involve graphical elements that are visually appealing, but not central to the game mechanics or to the learning content. IC may help students ignore these elements and instead identify and compare strategically relevant patterns within or across game levels. In addition to IC potentially being important for gameplay, we suggest that the ability to rapidly scan and integrate multiple sources of information from different regions of the game interface is likely to promote successful play. Because ANT orienting scores reflect abilities to shift attention, higher orienting scores may predict higher game achievement because the students will be able to scan and integrate information more rapidly.

Background: Gameplay-Based Individual Differences

In addition to the relationship between our primary IC measure and gameplay, we expect that student gameplay strategies will impact game performance and learning. In a study with self-explanation in an educational game, Hsu, Tsai, and Wang (2012) divided students into high and low engagement groups[§] based on the ratio of correct to incorrect (or "I don't know") responses to self-explanation prompts. The authors found that high engagement students scored significantly higher than low engagement students on their retention test. In this case, patterns of student behavior during self-explanation questions are interpreted to reflect a particular underlying state of the learner. Although an individual difference measure like this requires further validation to gain broader acceptance, we believe that games and in-game questioning provide a wealth of potential measures of student behavior that may reflect both strategies and other underlying states and traits of the learner.

In addition to the ANT, the second individual difference measure we include in the present study is how many actions students perform per trial during gameplay. Actions here are changes to the timeline (e.g., adding or removing a force command). This measure may have some relationship to students' trial-and-error behavior (as students who perform very few actions before each time they press launch to view the results of their actions are likely to be viewing the results to attempt to guide each move). Moreover, there may be certain parallels between a measure like this and measures of "gaming-the-system" that have been used to detect when students are attempting to avoid effort in intelligent tutoring systems (see Adams et al (2014); Baker, Corbett, Roll, & Koedinger, 2008).

[§] Generally, we advocate that researchers avoid dichotomizing a continuous variable. See Rucker, McShane, & Preacher (2015).

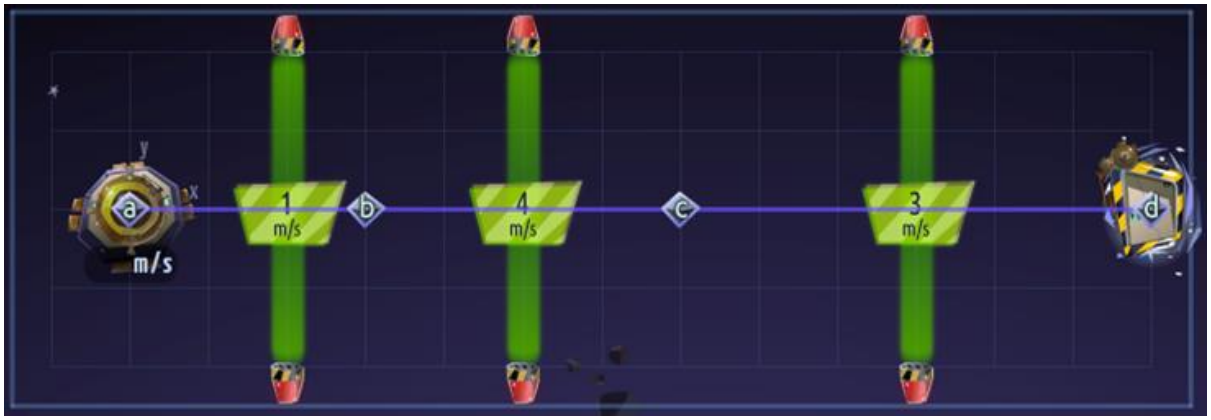

Figure 2. After the introductory dialog in a warp mission, students encounter the critical navigation challenge component of the warp mission.

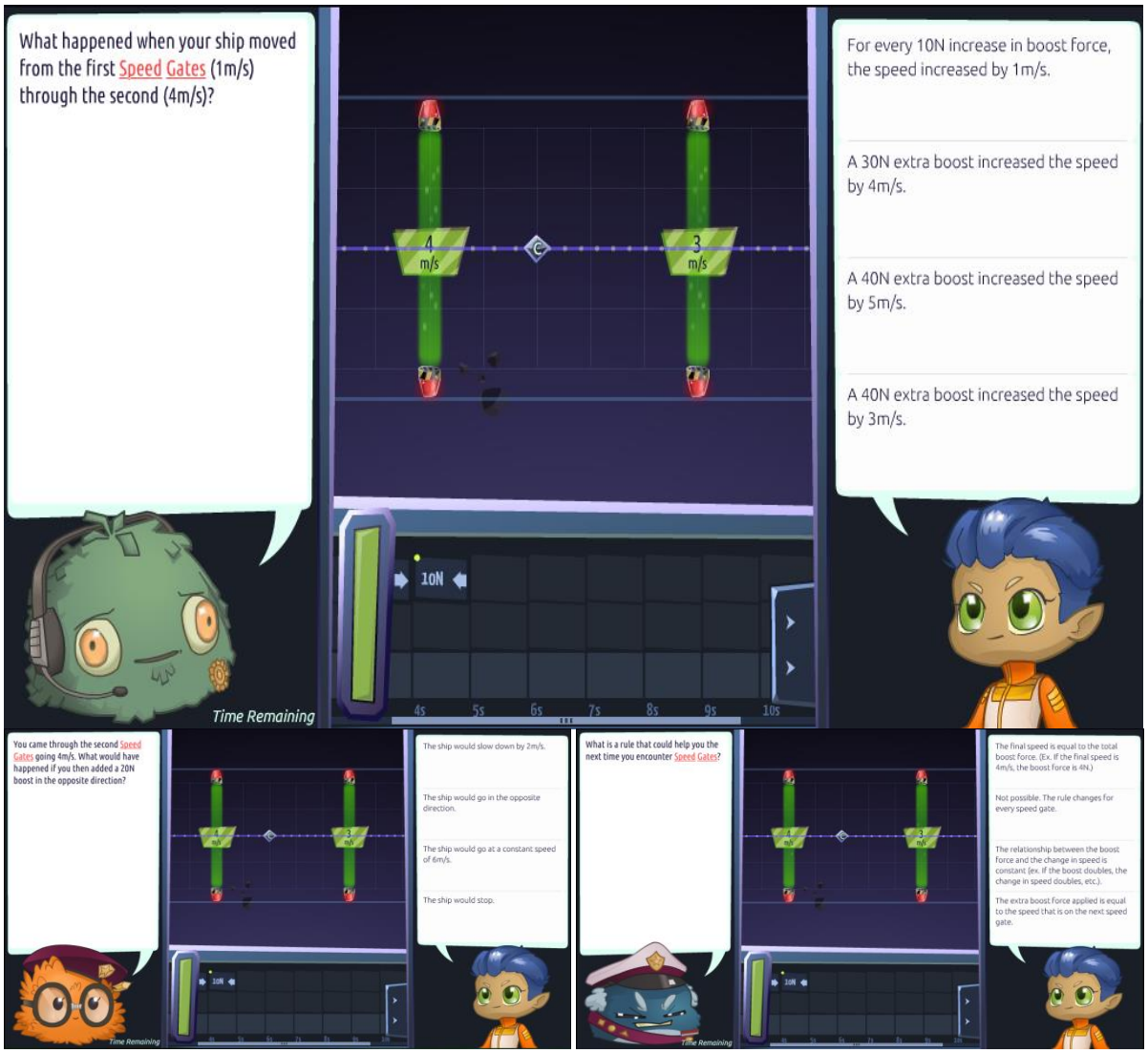

Figure 3. After succeeding in the navigation challenge in a warp mission, students encounter the explanation phase of the warp mission. In the full self-explanation condition, the first of the three questions asks the student to articulate the solution to the navigation challenge in a concrete manner (top), the second question asks the student to characterize the solution with a more abstract/generalizable relationship (bottom left), and the third question asks the student to articulate an even further abstracted a rule of thumb (bottom right).

Rationale for the Present Study on Self-Explanation Prompts

Adams and Clark (2014) examined an early prototype of the explanation functionality in *The Fuzzy Chronicles* using three conditions. Students in the self-explanation condition chose explanations before testing solutions, after incorrect solutions, and upon successfully completing stages (e.g., when asked why they needed to apply an impulse at the beginning of a level, students needed to choose "according to Newton's 1st law the ship will not move unless an unbalanced force acts upon it"). Students in the explanatory feedback condition received tips instead of explanation prompts. Finally, students in the control condition received feedback only about whether or not they succeeded on the level.

Though there were no overall differences between conditions, participants in the control group performed significantly better on Newton's second law questions compared to the self-explanation group. This was most likely due to students in the control group completing more levels as well as reaching levels that included more advanced concepts, but analyses were not presented that included level completions as a predictor/covariate to explicitly evaluate the relationship between level completions and learning outcomes. This highlights the importance of analyses that incorporate gameplay variables (such as level completions) into analyses of learning outcomes. Instructional techniques such as self-explanation prompts may affect gameplay and may affect learning indirectly through gameplay. The present study analyzes such relationships in greater depth. The study by Adams and Clark (2014) also suggested two major ways to improve upon the level design and self-explanation functionality embedded in *The Fuzzy Chronicles*.

Adams and Clark (2014) first suggest that students struggled with the level designs included in their study. The sequence of game levels appeared to build difficult concepts too quickly for the middle school students, and many levels introduced game entities (e.g., asteroids or shields) that were not directly relevant for learning and, thus, increased extraneous processing. Adding the processing demands of self-explanation to already difficult level designs may have occupied resources that might have otherwise supported generative processing. The processing demands of the initial level designs may have also disrupted students' experience of flow (Csikszentmihalyi, 1991) while gaming. During a state of flow participants are completely absorbed in an activity, which can be defined by characteristics such as concentration, time distortion, and sense of control (Kiili, 2005). To address these issues, game levels were simplified to help students focus on the key concepts and on related game mechanics.

The Adams and Clark (2014) study also suggests that the explanation functionality discouraged students from separating levels into smaller chunks to manage cognitive demands. Students often chose to build, test, and refine their solutions for only the initial section of a path before adding actions for subsequent sections of the path. The explanation functionality discouraged this strategy because questions were posed each time a student ran the simulation to view the results of an interim solution (adding an additional cost to repeated attempts). Thus it became apparent that the explanation functionality might better be placed after a student had identified a working solution rather than disrupting the solution process.

Beyond allowing students to manage cognitive demands through solving levels in smaller chunks, placing explanation functionality after students identify working solutions has other potential benefits. Research with cognitive tutors has shown that students will be more likely to engage in minimal processing and "game the system" if they perceive the tutor to be unhelpful (Baker et al., 2008). Postponing self-explanation until after a correct solution is likely to decrease perceiving the self-explanation functionality as unhelpful (as it will no longer interfere with level solutions). Perhaps most critically, postponing self-explanation also focuses students on the correct solution, which may be necessary for effective self-explanation (see Moreno and Mayer, 2005). These lessons from Adams and Clark (2014) underscore the challenges in redesigning and applying the findings of research from one learning context to another, particularly to contexts as rich as digital games. We adopted the following prescriptions in reprogramming explanation functionality in *The Fuzzy Chronicles*.

- First, explanation activities should occur after students correctly complete the level that they will be explaining.
- Second, the explanation functionality should be tied more closely to events in specific level segments so that (a) the students can more easily connect the explanations to the gameplay and (b) the students will be more likely to view the explanations as relevant and useful.
- Third, explanation activities and the game-play context in which they occur should be changed slightly each time they are encountered for a given physics relationship to incentivize understanding rather than simple memorization.
- Fourth, the explanation functionality should feel more a part of dialog with the game characters.

These four recommendations for *The Fuzzy Chronicles* were achieved by the following concrete changes to the game. First, we moved the explanation functionality into a new type of game level, which we call "warp missions." A trial in a warp level begins with some dialog where a game character asks the student for help. Warp levels then present a basic navigation challenge about the focal physics relationship, similar to the navigation challenge in a basic level (Figure 2). Students first need to solve the navigation challenge, and the warp level tracks how many attempts the student makes before solving it. In the full self-explanation condition, the student then encounters three explanation questions. The first question is very concrete and closely tied to the navigation challenge the student has just completed, the second question abstracts the solution to a slightly more generalizable form, and the third question frames the physics relationship in the most generalizable form (Figure 3). After selecting an explanation for each question (correct or incorrect), the game character provides feedback about the explanation. If the explanation was correct, the game character provides additional information, and the dialog moves to the next phase. If the explanation was not correct, the game character asks the student to reconsider the explanation.

After completing each trial of the warp mission, the game calculates a score for the student for that trial based on the number of attempts required in the navigation challenge and the number of attempts for each explanation question. The game computes an overall score for that warp mission based on the scores on the most recent trials for that mission for that student. Students need to earn a mission score above a certain threshold to unlock levels beyond the warp mission. If the student wishes or needs to play the warp mission again, the warp mission randomly selects a slightly new configuration of the navigation challenge and explanation answers so that students are encouraged to focus on the underlying concepts rather than merely applying a solution that was memorized on previous trials.

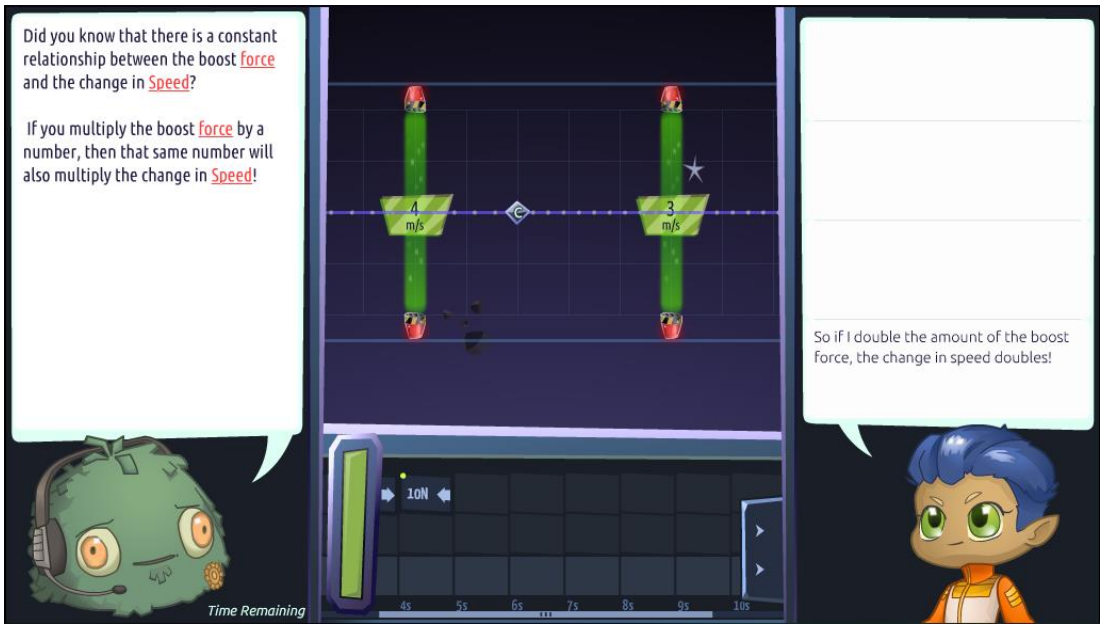

Figure 4. In the explanatory feedback condition, students are simply provided with the rule abstracting the physics relationship. This contains the same information from the self-explanation conditions (the third question of the full condition / the question from the shortened condition).

Comparison Conditions for the Current Study

The current study was structured to include a full self-explanation condition, a partial self-explanation condition, and an explanatory feedback condition. In the partial self-explanation condition, the explanation phase was abbreviated to include only the third self-explanation question focusing on the most generalizable statement of the physics relationship (Figure 3). In the explanatory feedback condition, the generalizable rule about the physics relationship is simply provided to the student (Figure 4). The scoring mechanisms for trials were accordingly adjusted such that the partial explanation condition was weighted evenly across navigation and explanation while the explanatory feedback condition score was based solely on the navigation phase. Because preliminary analyses revealed that the partial and full self-explanation conditions had similar effects on gameplay and learning and had similar relationships with individual difference variables, we collapsed across

the self-explanation conditions to simply compare self-explanation to explanatory feedback. We return to discuss the lack of differences between the full and partial self-explanation conditions in the general discussion.

Research Questions and Hypotheses for the Current Study

The current study compares embedded self-explanation to embedded explanatory feedback in an educational game. Though the study by Moreno and Mayer (2005) suggested that self-explanation will not be effective if interactivity is high, we expect that adopting a selected self-explanation method (like Johnson and Mayer, 2010) along with ensuring that students engage in self-explanation only *after* correctly completing a level will lead to an advantage for self-explanation over explanatory feedback.

- Hypothesis 1. Students in the self-explanation condition will have better learning outcomes than students in the explanatory feedback condition.

Because our game design allows students to progress freely through missions and all students are limited to a fixed number of days of play time, this will inevitably lead students to encounter different amounts of game content. Moreover, because new concepts are introduced throughout the level progression, students who complete more levels will actually encounter concepts that other students may never reach. Given these properties of our intervention, we formed the following two hypotheses.

- Hypothesis 2. Students who have higher pre-test scores will complete more levels.
- Hypothesis 3. Students who complete more levels will have higher post-test scores.

In addition to our interest in the effects of explanation functionality, we investigate how individual differences in the number of actions per trial[**] students take during gameplay affect learning and game play outcomes. As mentioned, we believe that students' gameplay behavior – which may be influenced by strategies, motivation, and other factors – will impact game and learning outcomes. Notably, the number of actions produced per trial is most likely to reflect prior knowledge. A player who is highly knowledgeable may produce more actions in a shorter time because she is aware of the correct solution. This relationship is not guaranteed, however, and player strategies may not track with intuitions. We investigate this metric in detail below (and discuss possible interpretations in the analyses and discussion). We propose that actions per trial will influence level completions, but will not be directly related to learning outcomes.

- Hypothesis 4. Students who produce more actions per trial in standard levels of the game will complete more levels in the game.

In addition to our interest in the effects of explanation functionality, we investigate how students' attentional control abilities impact gameplay, game outcomes, and learning outcomes. To measure students' abilities to direct attention, we included the child-friendly Attention Network Test (Rueda et al., 2004). As discussed above, we expect inhibitory control (measured in the ANT) may be relevant for learning science, gameplay, and perhaps learning from self-explanation. There are several possibilities for how attentional control might influence gameplay or learning from self-explanation. Thus, rather than form specific hypotheses, we put forward the following general hypothesis.

- Hypothesis 5. The relationships between ANT scores, gameplay, and learning will differ between the conditions.

Methods

Subjects

170 students from a middle school in the Southeastern United States participated in this study. The school served a racially diverse, primarily lower middle class population (71% of students qualified for free or reduced

[**] Notably, we include only actions per trial taken on incorrect trials because we wanted to determine how students moved toward a correct solution. This also avoids trials in which students simply knew the solution and immediately implemented it.

lunch). 47 students in our sample were enrolled in an English Language Learning (ELL) program. 1 student was removed from the study because he created two accounts and began playing the game over from the beginning during the study. Data were not analyzed from students (N = 49) who failed to complete one or more measures due primarily to absences. Additionally, data were not analyzed from students who failed to complete more than 4 levels of the game (N = 25) because our conditions were identical before the 4th level. This left 96 students' data for analysis (54 males and 42 females).

The *Fuzzy Chronicles* Game Design

*Game design***.** The game controls and game play mechanics are described in the introduction, above. The game contained 3 types of levels: standard levels, warp levels, and boss levels. Students first played 3-4 standard levels that gave the student a chance to play and explore a physics relationship (e.g., between the amount of force applied and the magnitude of a change in velocity). After the standard levels, students entered a warp level (discussed in depth above). Students then entered one or more challenge levels (or "boss levels") where they needed to apply the relationships in a more challenging combination. After completing the boss levels, the game repeats the full cycle again with another concept. The game contained three core cycles with extra cycles to challenge students who completed quickly.

Game conditions. There were three game conditions: explanatory feedback, partial self-explanation, and full self-explanation (described above). Classrooms were randomly assigned to conditions. As mentioned above, our analyses are collapsed into comparisons between the self-explanation and the explanatory feedback condition.

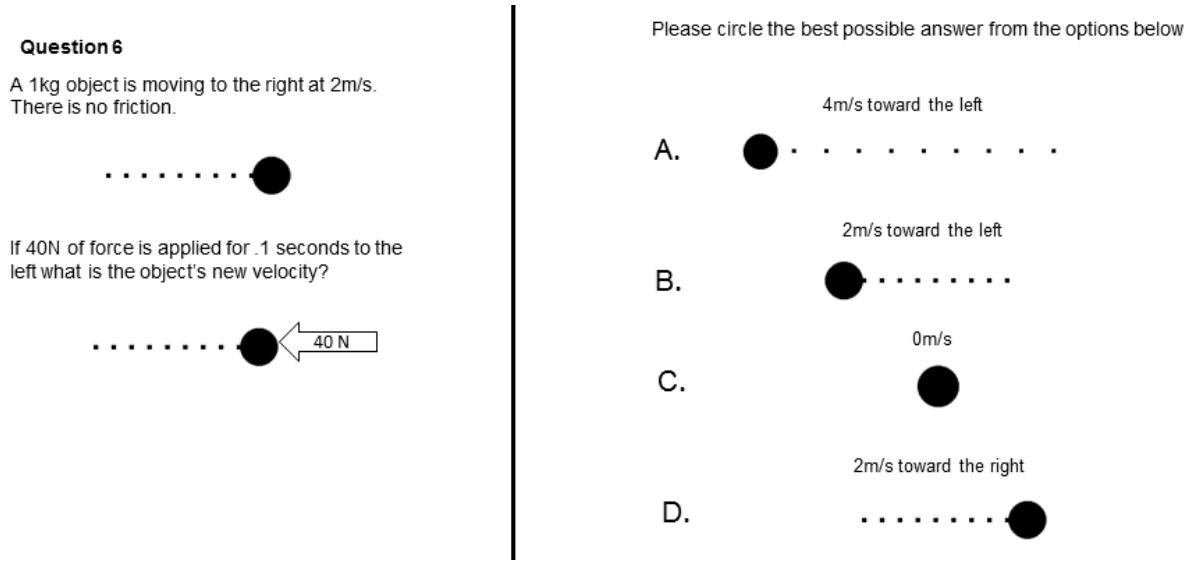

Figure 5. Example "near-transfer" problem from the physics assessment administered before and after gameplay.

Physics assessment. The assessment included 18 total questions. It is important to note that the pre-post questions focus explicitly on solving challenges that are near transfer from the navigation challenges (i.e., solving similar challenges in a non-game context) and not on restating or explicitly articulating the generalizable relationships that are the focus of the explanation functionality. More specifically, the test questions do not focus on rote memorization or restatement of the explanation content. Instead, the relationship of the explanation functionality to the test questions is indirect in the sense that focusing on deeper systemic understanding during explanation phases should support more effective solutions to test questions. Questions were presented in a paper packet with one question and a set of answers per page. The first 6 questions were "near-transfer" items that included graphical representations of objects, forces, and a "dot trace" representation to visualize paths and accelerations (see Figure 5). Each item had 4 possible answers that also had graphical representations (see Figure 5).

The next 6 questions were text-based "far-transfer" items that presented a scenario and had 4 possible text-based responses. These questions presented scenarios and students had to predict the outcome or indicate how an effect could be achieved. The final 6 questions were text-based "explanation" items that were being piloted in

this study but are not analyzed further. These items included general principles as warrants in the answer choices (e.g., "The object continues moving at the same speed. The motion is unchanged because the forces are balanced.").

The Attention Network Test. We adapted the child-friendly version of the ANT (Rueda et al., 2004) for middle school classroom use. On each trial (after a 1500ms ITI), a fixation cross was presented (400 to 1600ms). One of four cue types was then presented (100ms): no cue, a central cue (appearing at the same location as fixation), a double cue (appearing in both possible target locations), or a spatial cue (appearing at the upcoming target location). Cues (asterisks) were about the same area as the target (1.7˚). After a 400ms blank, the target (small fish) was presented either 1.9˚ above or below the prior fixation location. The target was presented either alone (neutral trials) or flanked by distractors (2 fish left and 2 fish right). Participants responded to the direction the central fish was facing (left or right). For left-facing targets, the 'z' key was the correct response. For right-facing targets, the '?' key was correct. On incongruent trials, distractors faced the opposite direction. On congruent trials, all fish faced the same direction.

Feedback was provided as follows. For a correct response: "+10 pts." For an incorrect response: "oops." For a delayed response (>1700ms): "too slow." After 4 practice trials that required a correct answer, participants completed 144 trials split into three 48-trial blocks with elective breaks between blocks up to 1 minute. Total points were visible during breaks. Three "network scores" were calculated from the results of the ANT that measure the contribution of three distinct forms of attentional control – though interactions have been observed by several groups (e.g., Callejas, Lupianez, Funes, & Tudela, 2005; Fan et al., 2002). First, the executive score was calculated for each student as the mean reaction time (RT) for all incongruent trials minus the mean RT for all congruent trials. The executive score is thought to reflect inhibitory control abilities, with smaller scores suggesting smaller differences between distracting and non-distracting trials (or greater inhibitory control).

Second, the orienting score was calculated as the mean RT for central cue trials minus the mean RT for spatial cue trials. The orienting score is thought to index students' abilities to use spatial information to aid attentional selection. An increase in the orienting score means that students are relatively faster at responding when a spatial cue is presented as compared to when only a central ready signal is presented. We interpret larger orienting scores to reflect a greater benefit of using spatial cues to select information. Finally, the alerting score was calculated as the mean RT for no cue trials minus the mean RT for double cue trials. The alerting score may reflect abilities to sustain readiness to attend to information over a variable interval (from fixation) as compared to when given a clear short-term alerting cue. Students have to wait only a short fixed interval for the target after the alerting. With no alerting cue, the onset of the target is ambiguous and at a longer interval from fixation. We interpret larger alerting scores to reflect a greater dependence on temporal cues.

Motivation questionnaire. The questionnaire included 4 Likert scale items with responses from 1 to 5. Responses were labeled as: "Strongly Agree," "Agree," "Neutral," "Disagree," and "Strongly Disagree." The items included were as follows:

1. I liked playing this game.
2. This game was difficult for me to play.
3. I worked hard to understand how to play the game and complete missions.
4. I would like to play this game, or more games like it, again in the future.

Analyses

*Removing ELL students***.** Given an unequal distribution across the three game conditions (with only 1 ELL student in the explanatory feedback condition), differences in game levels and pre-test scores, and potentially different relationships with individual difference measures (which may reflect differences in reading ability), we chose to analyze data from non-ELL students only (N=85).

ANOVA analysis of the ANT. Data from the Attention Network Test were analyzed in an initial ANOVA (see Fan et al., 2002; Rueda et al., 2004). These results are presented in an abbreviated form as they are not of interest to most readers. Analyses of reaction times showed a significant main effect of cue type (no cue > central > double > spatial), a significant main effect of compatibility (incompatible > compatible > neutral), and a significant interaction between these factors. Analyses of accuracy showed a significant main effect of compatibility only (compatible > neutral > incompatible). ANT network scores and reaction times from neutral trials (with no distractors and no cues) were used as predictors in the models below.

Scoring the motivation questionnaire. To analyze motivation questionnaire data, scores were first reverse coded for ease of interpretation (1=Strongly Disagree, 5=Strongly Agree). Preliminary factor analysis confirmed our expectation that two items: item 1 "I liked playing this game" and item 4 "I would like to play this game, or more games like it, again in the future" assessed a common "interest" factor (though the effort item also loaded on this factor). Reliability analysis for the two interest items was very high, $\alpha = .84$. The two items were averaged to form a single "interest" rating used in analyses below.

Regression, moderation, and mediation analyses. Regression analyses below present unstandardized coefficients for slopes as using the symbol *b* and standardized coefficients as β. Other than dummy-coded variables, predictors were mean-centered in regression analyses to permit meaningful interpretations of regression coefficients (Hayes, 2013). All mediation analyses below were conducted using the PROCESS macros (Hayes, 2013). Unlike other approaches, the bias-corrected bootstrap approach to mediation analyses implemented in the PROCESS macros does not suffer from being underpowered or from normality constraints (for discussions of these issues see Hayes, 2013; Preacher & Hayes, 2008).

In certain cases below, the Johnson-Neyman technique was applied to examine regions of significance for a particular interaction. The Johnson-Neyman (JN) technique allows one to make a non-arbitrary choice for identifying regions of a given variable within which another effect is significant (see Hayes, 2013). The JN technique derives the values of a continuous moderator (e.g., ANT executive scores) for which a *t* value, calculated from the ratio of a conditional effect to its standard error (e.g., the ratio of the effect of game condition on level completions to the standard error of this effect), is exactly equal to the critical value (in our case $\alpha = .05$).

Individual differences analyses. In the present experiment, we did not attempt to construct a detailed *a priori* model of how components of the ANT would affect performance across our game conditions. Although our analyses of individual differences are exploratory, we believe, as suggested by several authors (e.g., Cronbach 1957; Hayes, 2013), there can be great value in a degree of exploration within the model-building process.
For individual differences analyses, none of the mean values of predictors differed significantly between the game conditions. In regression analyses, all VIF values were generally below 2 (excepting moderation analyses and where mentioned). Many of the results, however, must be treated with caution given that level completion counts were not normally distributed and given the fairly small sample size for certain analyses.

Extreme univariate outliers for ANT neutral reaction times (N=2) and executive scores (N=1) were removed. The final sample contained 79 students (40 male and 39 female), with an equivalent distribution of males and females across conditions, $\chi^2(1) = 1.07$, $p = .30$. The effects for learning gains observed in the initial analyses (with the full sample) remained significant for the restricted sample in the individual differences analyses. For the individual differences analyses, we also collapsed near- and far-transfer physics measures into a single pre-test and post-test measure.

Results and Discussion

Learning Gains between Game Conditions

A repeated-measures ANOVA was conducted with test administration (pre- vs. post-test) as a within-subjects variable and game condition (explanatory feedback vs. self-explanation games) as a between-subjects variable. Our near-transfer and far-transfer questions were treated as separate measures. The multivariate test showed a significant main effect of test administration, $\lambda = .676$, $F(2, 82) = 19.653$, $p < .0001$, $\eta^2 = .324$, but the multivariate effect of game condition was not significant, $\lambda = .952$, $F(2, 82) = 2.074$, $p = .132$, $\eta^2 = .048$, and neither was interaction between game condition and test administration, $\lambda = .978$, $F(2, 82) = .939$, $p = .395$, $\eta^2 = .022$. Subsequent univariate tests showed that the effect of test administration was significant only for near-transfer questions, $F(1, 83) = 38.440$, $p < .0001$, $\eta^2 = .317$, with students performing better on near-transfer post-test questions ($M = 53.73\%$, $SD = 23.34$) than on pre-test questions ($M = 35.29\%$, $SD = 20.48$). The effect for far-transfer questions did not approach significance, $F(1, 83) = 2.250$, $p = .137$, $\eta^2 = .026$. Finally, none of the univariate effects of game condition or interactions between test administration and game condition approached significance.

Overall, students across conditions demonstrated significant pre-post learning gains. Univariate analyses by question type indicated that only gains on near-transfer questions were significant, but that these gains were

fairly robust. Comparisons between conditions suggested that students in the self-explanation game conditions did not outperform students in the explanatory feedback game.

Motivation Variables between Game Conditions

For the explanatory feedback condition, motivation survey ratings were as follows (means with SDs in parenthesis): interest: 2.79 (0.96), effort: 2.12 (.91), difficulty: 2.82 (.87). For the self-explanation condition, ratings were as follows: interest: 2.81 (1.01), effort: 2.06 (.84), difficulty: 2.88 (1.09). Independent samples t-tests showed no significant differences between the conditions for any rating[††].

Game Play Variables between Explanation Conditions

For the explanatory feedback condition, students completed a total of 14.21 (*SD*=3.68) levels on average and produced an average of .156 (*SD*=.082) actions per trial. For the self-explanation condition, students completed a total of 13.25 (*SD*=3.92) levels on average and produced an average of .162 (*SD*=.081) actions per trial. Independent samples t-tests showed no significant differences between the conditions for either measure. We chose to look at these specific aspects of gameplay in the subsequent analyses on individual differences because (a) level completions are a good index of the total amount of content students encountered and overall success with the game and (b) and average actions per trial give a simple measure of the size of the planned chunks students were testing on each trial.

Knowledge Treatment Effect Adjusted for Levels Completed

Regression models were constructed to analyze differences in learning outcomes between game conditions while controlling for pre-test performance and the highest level students completed. Separate models were constructed for near- and far-transfer questions. Game condition was included as a dummy coded predictor (explanatory feedback game = 0, self-explanation game = 1). For each model, moderation analyses indicated that the predictors had similar effects on post-test scores across game conditions. For the model of near-transfer question post-test performance, game condition was a significant predictor of post-test scores, with students scoring higher in the self-explanation condition than in the explanatory feedback condition, b = 11.129, SE_b = 4.518, p = .016, 95% CI (2.141, 20.118). For the model of far-transfer question post-test performance, game condition was not a significant predictor, b = -.465, SE_b = 5.149, p = .928, 95% CI (-10.711, 9.780). Higher pre-test scores and greater levels completed both predicted significantly higher near- and far-transfer post-test scores, but the relevant statistics are omitted for brevity. The overall models were significant for near-transfer post-test scores, $F(3, 84) = 10.253$, $p < .0001$, $R^2 = .275$, and far-transfer post-test scores, $F(3, 84) = 6.471$, $p < .001$, $R^2 = .193$.

Students were exposed to different amounts of learning material depending upon their success with the game. Though analyses of game play variables suggested that there were no overall differences in the number of levels completed between game conditions, game conditions may have affected students with different abilities differently. That is, even without an overall difference in levels completed, condition differences may exist for particular sub-groups that may mask overall differences in learning outcomes between conditions. The results obtained suggest after controlling for differences in level completions, a small, but significant advantage for the self-explanation game was observed specifically for near-transfer items.

Analysis of Relationships between the ANT, Gameplay, and Learning Outcomes

Regression analysis with physics pre-test scores. Analyses were first conducted with collapsed physics pre-test score (averaged across near- and far-transfer items) as a dependent variable. ANT network scores (executive, orienting, and alerting) and ANT neutral RT (reaction time) were included as predictors in each full model. To ensure the equivalence of predictors across self-explanation and explanatory feedback groups, moderation analyses were conducted with each of the relevant predictors. None of the interaction terms predicted significant variance in pre-test scores. Additionally, game condition was not a significant predictor of pre-test performance,

[††] After controlling for pre-test scores and/or highest level completed there were also no significant differences in motivation ratings between conditions.

b = -.852, SE_b = 3.856, p = .826, 95% CI (-8.531, 6.828), so the model of pre-test scores was collapsed across conditions.

Regression analyses showed that none of the ANT network scores or neutral RT predicted significant variance in pre-test scores. Additionally, the overall model including all predictors did not reach significance, $F(4, 77)$ = .548, p = .701, R^2 = .028. The results suggest that none of the predictors of interest accounted for significant variance in pre-test scores. These results are superficially unimportant to the research question, but analyzing relationships with pre-test scores is meaningful for (1) relating findings to the broader literature on academic achievement and (2) interpreting further relationships between abilities, game play, and learning within the same study.

Regression analysis for mean actions per trial. ANT network scores, ANT neutral RT, physics pre-test score, and game condition were included as predictors. Initially interaction terms were included. Because none of the interaction terms from initial moderation analyses predicted significant variance in pre-test scores, these terms were removed. Additionally, game condition was not a significant predictor of mean actions per trial, b = -.125, SE_b = .785, p = .875, 95% CI (-1.708, 1.458), so the model was collapsed across conditions. Regression analyses showed a significant effect of physics pre-test score, with higher pre-test scores predicting more actions per trial, b = .096, SE_b = .023, p < .001, 95% CI (.049, .142). None of the other predictors explained significant variance in the number of actions students took per trial. The overall model was significant, $F(5, 76)$ = 3.387, p = .008, R^2 = .182. The above analysis shows that only higher combined physics pre-test scores predicted more actions per trial across game types. Generally, the results suggest that our actions per trial measure are not directed by the cognitive abilities we measured in the ANT. The relationship with pre-test scores suggests that students with more prior knowledge tend to make more actions per trial.

Regression analysis with highest level completed. ANT network scores, ANT neutral RT, physics pre-test score, actions per trial, and game condition were included as predictors. Initially, interaction terms were included. Because none of the interaction terms from initial moderation analyses predicted significant variance in pre-test scores, these terms were removed. Additionally, game condition was not a significant predictor of highest level completed, b = -1.022, SE_b = .705, p = .152, 95% CI (-2.427, .383), so the model was collapsed across conditions.

Regression analyses showed a significant effect of average actions per trial, with more actions per trial predicting more level completions, b = .710, SE_b = .103, p < .0001, 95% CI (.505, .915). None of the other predictors explained significant variance in the number levels students completed. The overall model was significant, $F(6, 75)$ = 9.161, p < .0001, R^2 = .423. The above analysis shows that of the predictors entered only the mean number of actions students completed per trial predicted more level completions across game types. These results were surprising as we expected a relationship between pre-test scores and level completions. We return to this point in the general discussion.

Regression analysis with physics post-test scores. For predictors included in the full model, see Table 1. Moderation analyses showed a significant interaction between condition and ANT executive score, b = -.312, SE_b = .123, p = .013, 95% CI (-.557, -.068). Analyses treating condition as a moderator showed that executive scores did not predict differences in post-test scores for the explanatory feedback game, b = .094, SE_b = .086, p = .276, 95% CI (-.077, .266), but that lower executive scores predicted significantly higher post-test scores for the self-explanation game, b = -.218, SE_b = .085, p = .013, 95% CI (-.388, -.048). The JN technique showed that students with ANT executive scores below 46.72ms were predicted to have higher post-test scores in the self-explanation game (see Figure 6). A regression model including the above interaction term showed first that students who scored higher at pre-test scored higher at post-test (see Table 1). The overall model was significant, $F(9, 72)$ = 6.508, p < .0001, R^2 = .449.

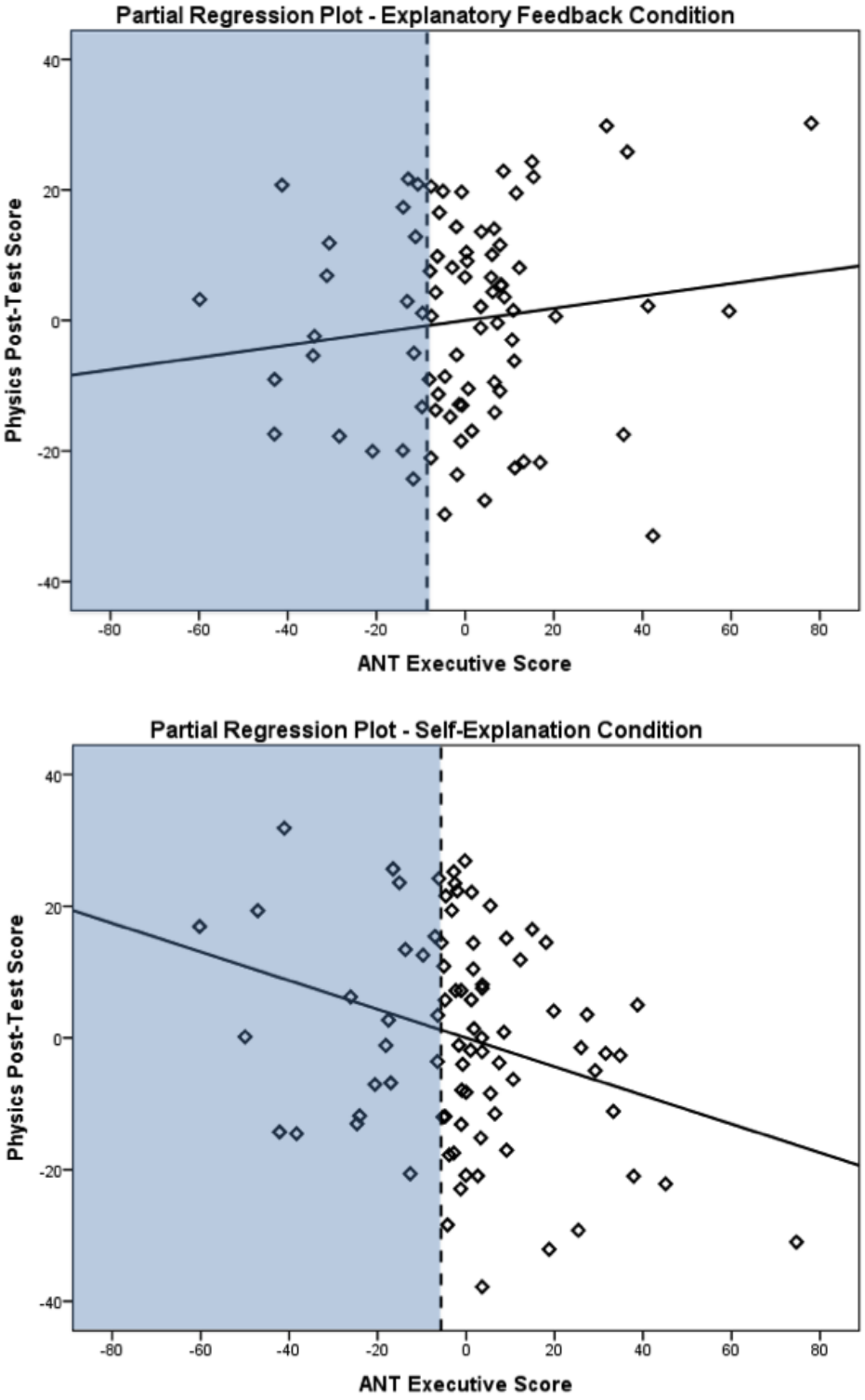

Figure 6. Partial regression plots of the relationship between combined post-test score and ANT Executive Score for the explanatory feedback condition (top) and the self-explanation condition (bottom). The region of significance (in which performance was better for the self-explanation condition than for the explanatory feedback condition) is highlighted on the left of each plot.

Pre-test scores were best predictors of post-test scores. Surprisingly, highest level completed did not predict post-test scores. We examine this further in mediation models below. Most interestingly, IC predicted higher post-test scores specifically for the self-explanation game. This finding suggests that students with better IC (lower ANT executive scores) were at an advantage in learning from the self-explanation game, but that IC had a minimal influence on learning in the explanatory feedback game. Using the JN technique, the results showed

that students with executive scores below 46.72ms (40% of the sample) had better overall learning outcomes for the self-explanation game. In terms of overall learning outcomes, this suggests that a certain threshold level of IC ability is necessary for students to benefit from our self-explanation condition.

Table 1. *Regression analyses for dependent variable: mean combined post-test score.*

Predictor	b (SE_b)	β	p	95% CI
Physics Pre-Test Score	.490 (.126)	.389	***	(.238, .743)
ANT Executive Score	.094 (.086)	.144	.276	(-.077, .266)
ANT Orienting Score	.018 (.048)	.034	.709	(-.078, .114)
ANT Alerting Score	.005 (.057)	.008	.933	(-.109, .119)
ANT Neutral RT	.007 (.027)	.025	.795	(-.047, .061)
Actions per trial	1.261 (.702)	.223	.077	(-.138, 2.660)
Highest Level Completed	.866 (.620)	.163	.167	(-.370, 2.102)
Condition	5.613 (3.785)	.133	.142	(-1.933, 13.159)
Condition X ANT Executive	-.312 (.123)	-.326	.013*	(-.557, -.068)

* $p < .05$, *** $p < .001$. b denotes the unstandardized regression slope coefficient and β denotes the standardized coefficient.

Exploratory Analyses

The analyses above investigate (1) the effects of individual differences (cognitive abilities and gameplay) on our primary outcome variables and (2) moderation effects of how condition affects each of these relationships between primary outcome variables and individual differences. As suggested in the introduction, however, relationships between variables may be mediated such that indirect effects, through influencing another variable, are significant. Additionally, especially when variables are highly correlated, suppression effects may occur in which a significant relationship is masked. We feel the need to highlight the exploratory nature of these analyses because these analyses do not directly follow from our hypotheses, but instead are meant to elucidate findings above. Further analyses might lead several of these results to be combined into a structural equation model that describes the relationships between the various variables more completely. These analyses are meant only to explore certain possible relationships within the OLS path analysis framework implemented in the PROCESS macros.

Suppression of highest level completed effect on physics post-test scores. Because mean actions per trial were a highly significant predictor of the highest level students completed, we were interested in whether the actions per trial variable may be suppressing the effect of highest level completed on physics post-test scores. The regression model for post-test scores above (including the interaction term) was rerun without the actions per trial predictor. This analysis revealed a significant effect of pre-test score and a significant interaction between ANT executive score and game condition, as outlined above. Additionally, the effect of highest level completed was significant, with higher level completions predicting higher post-test scores, $b = 1.548$, $SE_b = .498$, $p = .003$, 95% CI (.556, 2.539). Notably, because these variables suppress one another, the effect of actions per trial is also significant if highest level completed is removed from the model.

Although either predictor can be used to explain variance in post-test scores, we believe that the relationship between highest level completed and post-test score is likely more meaningful than the relationship between actions per trial and post-test score. Because the effect of level completions on post-test scores is significant even when controlling for pre-test score, this suggests that the effect is unlikely merely to reflect prior knowledge. One possibility is that completing more levels and producing more actions per trial reflect how easily students can apply abstract principles to the game (and perhaps achieve better learning from game content). Another possibility is that both mean actions per trial and highest level completed reflect students' fluency with using the game interface and this fluency frees up cognitive resources for students to devote to learning.

The effect of question response accuracy in the self-explanation game. In the self-explanation condition, self-explanation was implemented by students responding to questions in warp missions. Given earlier findings, such as that by Moreno and Mayer (2005), we wanted to investigate the effect of students' responses to these questions more carefully. First, a regression model was constructed with self-explanation response accuracy as the dependent variable. ANT network scores, ANT neutral RT, and physics pre-test score were entered as predictors. None of the predictors were significant. The effect of physics pre-test scores was marginal, $b = .002$, $SE_b = .001$, $p = .065$, 95% CI (.000, .004).

Second, a regression model for physics post-test scores was estimated for the self-explanation condition alone. The model included pre-test scores, ANT network scores, ANT neutral RT, highest level completed, mean actions per trial, and mean self-explanation response accuracy. The model showed significant effects of pre-test score and ANT executive score as expected from the earlier results. Additionally, there was a significant effect of self-explanation response accuracy, b = 65.180, SE_b = 20.813, p = .003, 95% CI (23.148, 107.213). Given the suppression effect noted above and observed VIF values greater than 2, we also conducted the analysis without including mean actions per trial. Interestingly, the effect of self-explanation response accuracy remained significant, b = 69.416, SE_b = 19.775, p = .001, 95% CI (29.509, 109.322), but the effect of highest level completed did not reach significance, b = 1.037, SE_b = .609, p = .096, 95% CI (29.509, 109.322). Highest level completed and self-explanation response accuracy were significantly correlated, r = .413, p = .003.

Finally, a parallel multiple mediation model was constructed using OLS path analysis and with combined physics post-test scores as a dependent variable. The model included a direct path between physics pre-test score and post-test score and indirect paths that connected pre-test score to each potential mediator (mean actions per trial, highest level completed, and mean self-explanation response accuracy) and then to post-test score (see Figure 7). The analysis controlled for the effects of other measured variables (ANT network scores) and was based on 10,000 bootstrap samples.

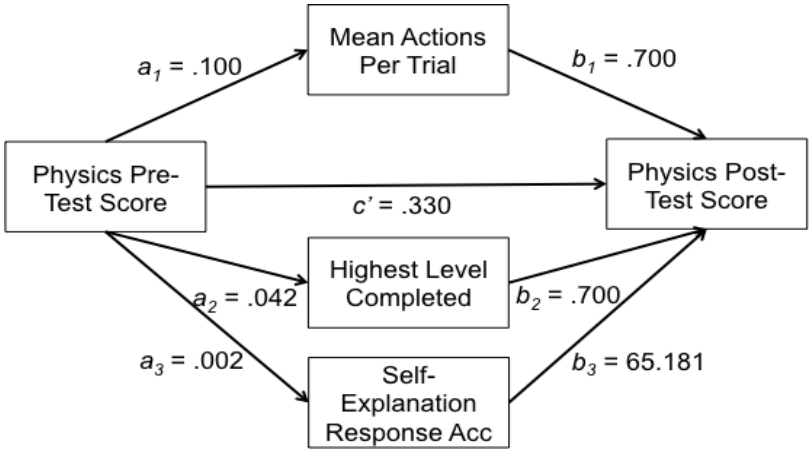

Figure 7. Visual representation of the parallel multiple mediation model evaluated for potential mediators between combined physics pre-test score and post-test score.

The direct effect of physics pre-test score on post-test scores (*c'* in Figure 7) remained significant after accounting for the indirect paths through our potential mediators (*c'* = .330, p = .036). Additionally, the analysis showed that the bias-corrected bootstrap confidence intervals for the specific indirect effect of mean actions per trial (a_1*b_1) as a mediator contained zero (-.119, .300) and so did the confidence interval for the specific indirect effect (a_2*b_2) of highest level completed (-.023, .198). The confidence interval for the specific indirect effect of self-explanation response accuracy (a_3*b_3), however, did not contain zero (.014, .311), suggesting a partial mediation effect. The confidence interval for the total indirect effect on self-explanation questions also did not contain zero (.026, .462). Finally, as is clear from earlier analyses, the total effect of pre-test score (i.e., the effect without including mediators model) was significant (c = .560, p < .001). Given the suppression effects observed above, these analyses were repeated with just one of the mean actions per trial or highest level completed variables and the results were unchanged.

Altogether, the results suggest an important role for correct responses to self-explanation questions in driving learning. Pre-test performance was marginally relevant for how well students performed on in-game self-explanation questions and was relevant for post-test performance. Additionally, (a) accuracy of self-explanation responses was a significant predictor of post-test scores after controlling for pre-test scores and (b) the indirect effect of pre-test knowledge on responses to in-game self-explanation questions accounted for how well students performed on post-test questions. We return to discuss the role of self-explanation responses further in the general discussion.

The effect of inhibitory control for the self-explanation game. A parallel multiple mediation model was constructed using OLS path analysis with combined physics post-test scores as a dependent variable. The model included a direct path between ANT executive score and post-test score and indirect paths that connected ANT executive score to each potential mediator (mean actions per trial, highest level completed, pre-test score, and

self-explanation response accuracy) and then to post-test score (see Figure 8). The analysis controlled for the effects of other measured variables and was based on 10,000 bootstrap samples.

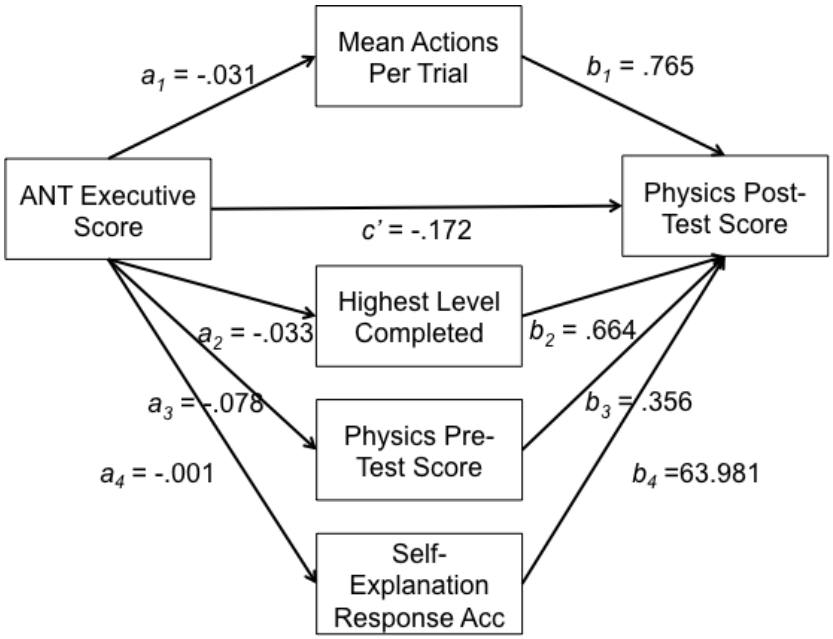

Figure 8. Visual representation of the parallel multiple mediation model evaluated for potential mediators between ANT executive score and combined physics post-test score.

The direct effect remained significant (c' = -.172, p = .040) and the total effect was also significant (c = -.298, p = .005). Each of the bias-corrected bootstrap confidence intervals for the specific indirect effects contained zero: for mean actions per trial (a_1*b_1): (-.133, .045), highest level completed (a_1*b_1): (-.146, .020), pre-test score (a_3*b_3): (-.164, .017), and self-explanation response accuracy (a_4*b_4): (-.206, .020). The confidence interval for the total indirect effect on self-explanation questions also contained zero (-.333, .013). Given the suppression effects observed above, these analyses were repeated with just one of the mean actions per trial or highest level completed variables and the results were unchanged.

The above analyses are important first because they indicate that the direct effect of IC on learning outcomes in the self-explanation game cannot be accounted for by indirect effects through mean actions per trial, level completions, self-explanation response accuracy, or pre-test scores. Altogether, we interpret these and prior findings to suggest that IC directly affects how students learn from the self-explanation game but does not affect the gameplay variables we analyzed.

Regression Analysis with Motivation Ratings

The same predictors were included as in the analyses for highest level completed above. Highest level completed was also included as a predictor. Separate analyses were conducted for each motivation rating. For interest ratings, moderation analyses showed no significant differences in the effects of these predictors across conditions. Additionally, the effect of condition was not significant, b = .127, SE_b = .222, p = .569, 95% CI (-.316, .571), so this predictor was not included. Physics pre-test score was a significant predictor, with students that had higher initial physics assessment scores providing higher interest ratings, b = .021, SE_b = .006, p = .002, 95% CI (.008, .033). Students with higher level completions also provided significantly higher interest ratings, b = .062, SE_b = .027, p = .024, 95% CI (.009, .116).

For effort ratings, moderation analyses showed no significant differences in the effects of these predictors across conditions. Additionally, the effect of condition was not significant, b = -.096, SE_b = .200, p = .633, 95% CI (-.495, .303), so this predictor was not included. Physics pre-test score was the only significant predictor, with students that had higher initial physics assessment scores providing higher effort ratings, b = .016, SE_b = .006, p = .010, 95% CI (.004, .028). Finally, for difficulty ratings, there were no significant moderation effects and none of the predictors were significant. Altogether, we found that students with higher pre-test scores had higher ratings of interest and effort. Additionally, students with more level completions provided higher interest ratings.

General Discussion

Overview of Findings for Experimental Hypotheses

Here we provide an overview of the findings relevant to each of our earlier experimental hypotheses. Some findings are discussed in greater detail following the overview. Although evaluating these hypotheses is valuable, we believe the most interesting findings are the results of analyses of individual differences discussed after this section.

- Hypothesis 1 was not generally confirmed. Nevertheless, there was an advantage of self-explanation over explanatory feedback for near-transfer items after controlling for level completions.
- Hypothesis 2 was not confirmed. Higher physics pre-test scores did not significantly predict more level completions.
- Hypothesis 3 was partly confirmed. More level completions predicted higher physics post-test scores but only when a suppressive predictor (actions per trial) was removed.
- Hypothesis 4 was confirmed. Students who made more actions per trial during standard gameplay completed more levels.
- Hypothesis 5 was confirmed. The relationships between abilities, gameplay, and learning differed between the conditions. Most importantly, inhibitory control predicted learning from the self-explanation game, but not from the explanatory feedback game.

The Effects of Self-Explanation on Learning Outcomes

Because we observed no differences between our full self-explanation and partial self-explanation conditions, we collapsed across the self-explanation conditions to examine differences between self-explanation and explanatory feedback overall. Though our full and partial self-explanation designs did not differ, others have observed differences based on different self-explanation designs (see O'Neil et al., 2014). One possible explanation is that our full self-explanation condition simply did not promote more reflection than did the partial condition. Alternatively, the benefits of prompted self-explanation may diminish with each question presented.

There were also no overall differences in learning gains between the self-explanation and explanatory feedback conditions. Once we controlled for the total number of levels completed, however, students had better near-transfer post-test scores in the self-explanation condition. Because new levels introduced new learning content in our study, this suggests that individual differences in students' gameplay may have masked an overall condition difference. Generally, this result highlights the importance of considering student gameplay behavior and strategies when investigating learning from games. Our approach of identifying quantifiable metrics of play and examining regression models is certainly not the only useful approach – for example, qualitative classifications of play may reveal even deeper relationships with learning outcomes. Regardless, when conducting a study with traditional pre/post assessment we urge researchers not just to examine learning outcomes and play separately, but to investigate deeper connections between them.

In addition to overall condition differences, we observed important results specifically for the self-explanation game. First, pre-test knowledge contributed marginally to accuracy in answering in-game questions. Second, accuracy in answering in-game questions was a significant predictor of post-test scores (and level completions were reduced to marginal significance). In line with the findings of Moreno and Mayer (2005), the present results suggest an important role for correct responses to self-explanation questions. The results might even be taken to suggest equal or greater importance for accurate responses to self-explanation questions than for level completions. Notably, however, progress on levels and accuracy on responses to in-game self-explanation questions were correlated and were thoroughly interwoven in our game, which makes this finding less straightforward.

Finally, though significant variance was still attributable to a direct relationship between pre- and post-test scores, mediation analyses showed that the indirect relationship (from pre-test scores, to in-game question scores, to post-test scores) was significant, but other indirect effects (through level completions or actions per trial) were not. More specifically, how students applied their prior knowledge to answering in-game questions explained some of their improvement at post-test even after accounting for the direct effect of prior knowledge

on post-test scores. The results suggest, however, that the most notable indirect relationship between pre-test scores and post-test scores was through influencing students' accuracy on in-game self-explanation questions.

Relationships between the ANT, Self-Explanation, Gameplay, and Learning

Lower executive scores (higher IC) predicted higher post-test scores for the self-explanation game, but there was no relationship between these variables for the explanatory feedback game. Mediation analyses suggested that there was no evidence that level completions, actions per trial, or accuracy in responding to self-explanation questions mediated the direct effect of IC on post-test scores. Finally, analyses of regions of significance using the Johnson-Neyman technique showed that students below a particular executive score (medium-to-high IC) had better post-test outcomes with the self-explanation game. Also, notably, there were no regions in the distribution of ANT executive scores (IC) for which students had better learning outcomes with the explanatory feedback game. The simplest interpretation of these findings is that though no students seem to benefit from our explanatory feedback manipulation over our self-explanation manipulation, students need to be above a certain threshold for IC abilities in order to take advantage of self-explanation.

Specifically, we identify three potential benefits of self-explanation based on suggestions by Roy and Chi (2005) that may be moderated by IC. First, IC may be necessary for integrating information across gameplay episodes and for easily relating these episodes to self-explanation activities. Second, IC may help students engage in metacognition during self-explanation activities so that they actively identify and correct potentially incorrect hypotheses that were formed during gameplay. Third, students with better IC may actively maintain potentially relevant information from self-explanation activities in working memory to facilitate subsequent gameplay. As mentioned in the introduction, prior research suggests that self-explanation may promote metacognition and reflection (e.g., McNamara & Magliano, 2009), and students' abilities to engage in metacognition and reflection may be connected to their inhibitory control abilities (Fernandez-Duque et al., 2000; Flemming & Dolan, 2012). Thus, IC may be necessary to obtain certain benefits claimed for self-explanation.

It is possible that our findings with IC are broadly applicable to self-explanation (beyond just educational games). IC may have value for predicting who will benefit from self-explanation in any intervention because a certain level of inhibitory control ability may be necessary to engage in relevant metacognition and abstraction. An alternative possibility is that isolated measures of IC correlate with other aspects of executive function or even working memory (see Miyake et al., 2000), and these correlated variables are actually what is relevant for predicting success with self-explanation. One useful way to investigate these possibilities would be to conduct a non-game-based self-explanation study that compares structural equation models based on several measures of inhibitory control and of potential alternative explanatory constructs (e.g., working memory).

Alternatively, it is possible that our findings are applicable only to self-explanation within games (or even within our particular game). Differences in IC may not be relevant to learning from the self-explanation process itself, but instead may be relevant to how self-explanation affects gameplay and thinking during gameplay. Gropen et al. (2011) suggest that IC may be relevant for students to suppress experiential interpretations of experiences in the sciences and for students to employ analytic and abstract explanations. Thus, self-explanation may provide conceptual scaffolds that can be used to reach a deeper understanding of gameplay, but this reflection may be applied successfully only if students have sufficient abilities for inhibitory control. Under this interpretation, IC may be especially relevant for self-explanation in games because it determines how well students can connect gameplay to self-explanation activities. This relationship could also contribute to why self-explanation has shown inconsistent results in studies with educational games (Adams and Clark, 2014; Johnson and Mayer, 2010; Moreno and Mayer, 2005; O'Neil et al., 2014).

Gameplay Metrics in the *Fuzzy Chronicles*

The two gameplay metrics we adopted in this study were the mean number of levels a student completed and the mean number of actions a student produced per trial during gameplay. As we expected, greater numbers of actions per trial predicted greater level completions and greater level completions predicted higher post-test scores (after removing actions per trial from the analysis of post-test scores). Our metric of the number of actions produced per trial was highly correlated with the number of levels students completed. This may suggest that students who understood how to solve puzzles in the game placed more actions on each trial because they

were better able to predict how multiple actions would influence the ship's motion. Placing more actions per trial may have, in turn, led students to complete levels more quickly.

In addition to these expected findings, we observed certain unexpected results. We were surprised that pre-test scores did not predict level completions. This suggests that the differences in prior knowledge between students in our sample did not substantially contribute to how many levels they were able to complete. Interestingly, however, pre-test scores did predict the number of actions per trial that students produced. Given the relationship between pre-test scores and actions per trial and the relationship between actions per trial and level completions, one might expect that the number of actions per trial was masking (or mediating) the effect of pre-test scores on level completions, but follow-up analyses suggested that this was not the case. The results might suggest that students with higher pre-test scores tended to test larger potential solution chunks at a time and perhaps chunks that were better grounded in prior knowledge. These students, however, may have also spent more time thinking between levels or reading text in levels that led to the absence of a relationship between pre-test scores and level completions. Any proposal to explain these findings is merely speculative. The results suggest that gameplay metrics can provide additional insight into relationships with cognitive abilities and with learning outcomes. Developing a more complete picture of how these components relate to one another in future studies, however, may require more sophisticated game metrics, more subtle analyses of relationships between various gameplay metrics, and supplementing quantitative metrics with qualitative analyses of play.

Limitations and Considerations

Benefits and Difficulties of Using Motivation and Engagement Measures

Our analyses of our measures of motivation generally suggest that higher prior knowledge leads to greater interest and perceived expenditure of effort. Completing more levels also led to higher levels of interest. This may reflect that as students complete more levels, they experience increases in self-efficacy that lead to greater interest in the game. Importantly, in the analyses above we collect motivation/engagement measures at the end of our study, treating our post-intervention measures of game motivation as outcomes of play. This is a common practice in educational games research, but bears further scrutiny. Given our measures were collected at the end of the study, it would be problematic to incorporate the measures as potential mediators mediation models, but incorporating the measures as predictors in regression models also raises obvious questions about causal relationships that complicate interpretation. An alternative approach is to measure motivation at the outset of a study (after a brief exposure to a game) using an instrument such as the questionnaire on current motivation that has been used to measure contextualized achievement motivation with non-game learning materials (Rheinberg, Vollmeyer, & Burns, 2001; Freund, Kuhn, & Hollig, 2011).

Beyond the issue of when one collects measures, there is also an inherent complexity in interpreting relationships between motivation/engagement and other measured variables. This particular complexity is revealed through our finding that higher pre-test scores predict higher game interest ratings. First, it is worth noting that our ratings cannot differentiate between possible contributions of topic interest in physics, topic interest in video games generally, and situational interest in the specific gameplay experience (see Krapp, Hidi, & Renninger, 1992 for a discussion of situational and topic interest). Second, the relationship between interest and pre-test scores can have several interpretations. One interpretation of our observed interest-knowledge relationship is that pre-test scores reflect prior knowledge that allows students to make sense of game content and to interpret game experiences in a meaningful way. This interpretation relates to the idea that relevance is generally important for supporting motivation to learn (see Pintrich, 2003). Alternatively, our results may reflect a restricted segment of a wider inverted U shape relationship between game interest and prior knowledge (see Tobias, 1994 for a discussion) as our sample was not likely to include any students with high prior knowledge.

Another alternative to the "relevance explanation," above, is that pre-test scores may not have impacted students' interest, but instead interest ratings merely reflect students' estimates of their improvement on the post-test (and thus of how much they learned). Finally, it is possible that the causal direction of the relationship is reversed. Game interest/effort may reflect broader domain interest/effort; and students' interest in the subject area might support greater effort on the pre-test. Despite the difficulties in measuring and analyzing relationships with motivation, this is an important direction for educational games research – especially considering the often-delivered, but seldom-investigated, claim that games are valuable learning tools because they support motivation and engagement. It is also worth noting that our measures of interest and other variables were not especially sophisticated in the present study and we generally urge researchers to adopt previously validated instruments to measure motivational constructs.

Spurious Relationships and the Problem with "Prior Knowledge"

In individual differences analyses, it is always important to consider that a significant relationship with an outcome of interest can always reflect a spurious relationship in which another correlated but unmeasured variable is actually the causally-relevant predictor. Pre-test scores receive very little scrutiny in this regard. Many researchers are (and should be) interested in assessing students' initial domain knowledge. Performance on a pre-test, however, reflects more than simply domain knowledge. Jonassen and Gabrowksi (1993) define prior knowledge as the "knowledge, skills, or ability that students bring to the learning process" (p. 417), but often pre-test scores are implicitly taken to reflect prior knowledge in the restricted sense of domain knowledge (for other issues relating to prior knowledge, see Dochy, Segers, & Buehl, 1999).

Certainly some variance in a student's score on an assessment will reflect domain knowledge, but students like those in our study have yet to receive relevant formal instruction. In such cases, we find it more reasonable to assume that better pre-test scores reflect students' naïve intuitions (e.g., beliefs about the physical world, diSessa, 1993), generic strategies for making sense of problems, specific knowledge from other areas that informs their reasoning, and even individual differences in relevant cognitive and motivational variables. Furthermore, even for adults who may have been exposed to relevant formal instruction, pre-test scores may still reflect other individual differences such as spatial abilities. We urge researchers to include multiple individual difference measures along with pre-test scores to help isolate the contribution of domain knowledge and other abilities (that may support performance on a pre-test) to explaining treatment interactions.

From Abilities to Theoretically-Grounded Aptitude Complexes

In the present study, our discussions of individual differences have focused on constructs we call abilities. We name inhibitory control and orienting as abilities, and one could consider many other individual differences in cognition to be abilities as long as they have some predictive value for learning or performance, but there are two important points to consider when investigating cognitive abilities. The first consideration is the established theoretical basis for the particular ability construct and the second is the complexity of the contextualized effects of abilities within real-world learning.

Regarding the first point, there is a long history of cognitive research that continues to strive for identifying and characterizing basic cognitive processes (e.g., working memory, inhibitory control) or basic domain-relevant skills (e.g., phonemic awareness) that contribute to higher-order action and thought. When selecting a cognitive ability of interest, it is beneficial to consider basic processes for several reasons. First, one has access to an abundance of existing research on these processes to guide predictions during model building and to guide interpretation when reporting results. Second, if these processes are indeed basic components of more-complex acts of cognition, then this serves to organize research around a common set of abilities and may mitigate concerns over spurious relationships to some degree (as, at the very least, it is unlikely that some more basic process accounts for a relationship between one's ability measure and other outcome variables). Finally, we think measures of basic cognitive abilities are more promising than measures such as general intelligence because cognitive ability measures support specific theories about the relevant cognition underlying learning and performance.

The second point regarding investigations of cognitive abilities is that abilities are contextualized. This point was repeatedly made by Richard Snow (e.g., Snow, 1994) in his various discussions of what he called "aptitude complexes." Snow did not envision student performance on a given task as simply an outcome of individual cognitive abilities and the instructional design, but instead he saw performance as a multi-factor relationship among variables such as cognitive abilities, motivational factors, tasks, teachers, and prior knowledge (see also Roeser et al., 2002 for an informative discussion and review). For example, cognition and motivation may have interdependent effects. An individual's cognitive abilities may impact learning only if the individual is sufficiently motivated to learn. Additionally, even if a student is highly motivated, cognitive deficits may hinder a student's success. Though there is likely some reciprocity between motivation and cognition (as students with greater cognitive abilities are likely to develop greater self-efficacy), the relationship is undoubtedly more complex. In addition to relationships among student characteristics, there are also relationships among aptitudes and prior knowledge – as suggested by a study of spatial abilities and physics assessment performance that suggests the relevance of spatial abilities for assessment performance diminishes with increasing knowledge (Kozhevnikov and Thornton, 2006).

Conclusions

Generally, we found that self-explanation games promoted better near-transfer learning outcomes than an explanatory feedback game after controlling for the number of levels that students completed. Moreover, accuracy in responding to self-explanation questions was an important predictor of learning in the self-explanation game. Finally, we found that students' inhibitory control abilities were relevant for learning from our self-explanation game, but not for learning from our explanatory feedback game. Together, these results suggest that self-explanation can benefit learning beyond explanatory feedback in educational games, but researchers must consider how learning is influenced by individual abilities and gameplay behavior.

More generally, this research underscores the value of considering the relationships between abilities, gameplay, and learning when evaluating the effects of an educational game design. When an educational game design presents numerous new concepts to students, more level completions in the game are likely to produce better learning outcomes, but game behavior and individual abilities (1) may differentially impact learning and play outcomes between different game designs and (2) may impact learning both directly and indirectly through their effects on level completions or other gameplay variables. Moderation and mediation analyses are two valuable methods for building more sophisticated models of the learning process in educational games. Moreover, these methods can help build a framework for differentiated game interventions that avoids the simplifying assumption that the treatment with the greatest mean learning outcome is necessarily the best for every student.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, and the National Science Foundation through grants R305A110782 and 1119290 to Vanderbilt University. IES supported software development and data collection. NSF supported analysis and manuscript authoring. 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. & Clark D. B. (2014). Integrating self-explanation functionality into a complex game environment: Keeping gaming in motion. *Computers and Education*, *73*, 149-159.

Adams, D.M., McLaren, B.M., Durkin, K., Mayer, R.E., Rittle-Johnson, B., Isotani, S., & van Velsen, M. (2014). Using erroneous examples to improve mathematics learning with a web-based tutoring system. *Computers in Human Behavior, 36*, 401-411. http://dx.doi.org/10.1016/j.chb.2014.03.053

Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, *18*(3), 287-314. http://dx.doi.org /10.1007/s11257-007-9045-6

Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review*, *29*(3), 180-200. http://dx.doi.org/10.1016/j.dr.2009.05.002

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. Child development, 78(2), 647-663. http://dx.doi.org/10.1111/j.1467-8624.2007.01019.x

Callejas, A., Lupianez, J., Funes, M. J., & Tudela, P. (2005). Modulations among the alerting, orienting and executive control networks. *Experimental Brain Research*, *167*(1), 27-37. http://dx.doi.org/10.1007/s00221-005-2365-z

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. http://dx.doi.org/10.1207/s15516709cog1302_1

Chi, M. T. H., & VanLehn, K. A. (in press). The content of physics self-explanations. *Journal of the Learning Sciences.*

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

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. http://dx.doi.org/10.1186/s40594-014-0014-4

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist, 12*(11), 671. http://dx.doi.org/10.1037/10049-022

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience*. New York: Harper Perennial.

DiSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and instruction*, *10*(2-3), 105-225.

Dochy, F., Segers, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*(2), 145-186. http://dx.doi.org/10.3102/00346543069002145

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14*(3), 340–7. http://dx.doi.org/10.1162/089892902317361886

Fernandez-Duque, D., Baird, J., & Posner, M.. (2000). Awareness and metacognition. *Consciousness and Cognition, 9*, 324–326.

Fleming S.M., Dolan R.J. (2012). The neural basis of accurate metacognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, 367:1338–1349.

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. Personality and Individual Differences, *51*(5), 629-634. http://dx.doi.org/10.1016/j.paid.2011.05.033

Fuchs, L. S., Schumacher, R. F., Sterba, S. K., Long, J., Namkung, J., Malone, A., Hamlett, C. L., Jordan, N. C., Gersten, R., Siegler, R. S., & Changas, P. (2014). Does working memory moderate the effects of fraction intervention? An aptitude-treatment interaction. *Journal of Educational Psychology, 106*(2), 499-514. http://dx.doi.org/10.1037/a0034341

Gee, J. P. (2007). *What video games have to teach us about learning and literacy*, Revised and Updated Edition (2nd ed.). Palgrave Macmillan.\

Gropen, J., Clark-Chiarelli, N., Hoisington, C., & Ehrlich, S. B. (2011). The importance of executive function in early science education. Child Development Perspectives. http://dx.doi.org/10.1111/j.1750-8606.2011.00201.x

Hayes, A. F. (2013). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. New York, NY: The Guilford Press.

Höffler, T. N., & Leutner, D. (2011). The role of spatial ability in learning from instructional animations–Evidence for an ability-as-compensator hypothesis. *Computers in Human Behavior, 27*(1), 209-216. http://dx.doi.org/10.1016/j.chb.2010.07.042

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.

Jonassen, D. H. & Gabrowski, B. L. (1993). *Handbook of individual differences, learning and instruction*. Part VII, Prior knowledge. Hillsdale: Lawrence Erlbaum Associates.

Kozhevnikov, M. & Thornton, R. (2006). Real-time data display, spatial visualization ability, and learning force and motion concepts. *Journal of Science Education and Technology, 15*, 113-134.

Kiili, K. (2005). *On educational game design: Building blocks of flow experience.* Tampere, Finland: Tampere University of Technology Press.

Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning and development. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 1-26). Hillsdale, NJ: Erlbaum.

Lester, J. C., Stone, B. A., & Stelling, G. D. (1999). Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. In*Computational Models of Mixed-Initiative Interaction* (pp. 185-228). Springer Netherlands. http://dx.doi.org/10.1007/978-94-017-1118-0_5

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of experimental child psychology,* 104, 1–21. doi:10.1016/j.jecp.2008.08.004

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.

McNamara, D.S., & Magliano, J.P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). New York, NY: Elsevier Science.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A.H., Howerter, A. & Wager, T.D. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49-100. http://dx.doi.org/10.1006/cogp.1999.0734

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. http://dx.doi.org/10.1037/0022-0663.97.1.117

O'Neil, H. F., Chung, G. K., Kerr, D., Vendlinski, 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. http://dx.doi.org/10.1016/j.chb.2013.07.025

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology, 95*(4), 667. http://dx.doi.org/10.1037/0022-0663.95.4.667

Preacher, K. J. & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879-891. http://dx.doi.org/10.3758/BRM.40.3.879

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). QCM: A questionnaire to assess current motivation in learning situations. *Diagnostica*, *47*(2), 57-66.

Roeser, R. W. Shavelson, R. J., Kupermintz, H. Lau, S. Ayala, C., Haydel, A., Schultz, S., Gallagher, L., & Quihuis, G. (2002). The concept of aptitude and multidimensional validity revisited. *Educational Assessment, 8*(2), 191-205. http://dx.doi.org/ 10.1207/S15326977EA0802_06

Roy, M., & Chi, M. T. (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.

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia, 42*(8), 1029–40. http://dx.doi.org/10.1016/j.neuropsychologia.2003.12.012

Rucker, D. D., McShane, B. B., & Preacher, K. J. (2015). A researcher's guide to regression, discretization, and median splits of continuous varaibles. *Journal of Consumer Psychology* http://dx.doi.org/10.1016/j.jcps.2015.04.004

Salen, K., & Zimmerman, E. (2004). Rules of play: game design fundamentals. 2004.

Snow, R. E. (1994). Abilities in academic tasks. In R. J. Sternberg & R. K. Wagner (Eds.) *Mind in context: Interactionist perspectives on human intelligence* (pp. 3-37). Cambridge: Cambridge University Press.

Squire, K. (2005). Changing the game: What happens when video games enter the classroom. *Innovate: Journal of online education*, *1*(6).

St Clair-Thompson H.L, & Gathercole S.E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*,*59*, 745–759

Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds.*Journal of Science Education and Technology*, *17*(6), 530-543. http://dx.doi.org/10.1007/s10956-008-9120-8

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research*, *64*(1), 37-54. http://dx.doi.org/10.3102/00346543064001037

Visu-Petra, L., Cheie, L., Benga, O., & Miclea, M. (2011). Cognitive control goes to school: The impact of executive function on academic performance. *Procedia Social and Behavioral Sciences*, *11*, 240-244. http://dx.doi.org/10.1016/j.sbspro.2011.01.069

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2009). Examining the relations among working memory capacity, attention control, and fluid intelligence from a dual-component framework. *Psychology Science Quarterly, 51*(4), 388-402.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and instruction*, *16*(1), 3-118. http://dx.doi.org/10.1207/s1532690xci1601_2

White, B. Y., & Frederiksen, J. R. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. In J. A. Minstrell & E. H. Van Zee (Eds.), *Inquiring into Inquiry Learning and Teaching in Science* (pp. 331-370). American Association for the Advancement of Science.

Wiley, J., Sanchez, C. A., Jaeger, A. J. (2014). The individual differences in working memory capacity principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 615-636). New York, NY: Cambridge University Press.

Wright, W. (2006). Dream machines. *Wired 14*(04). http://archive.wired.com/wired/archive/14.04/wright.html