Productive Responses to Failure for Future Learning

Alison Lee

Submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
under the Executive Committee of the
Graduate School of Arts and Sciences

Columbia University
2017

ABSTRACT

Productive Responses to Failure for Future Learning

Alison Lee

For failure experiences to be productive for future performance or learning, students must be both willing to persist in the face of failure, and effective in gleaning information from their errors. While there have been extensive advances in understanding the motivational dispositions that drive resilience and persistence in the face of failure, less has been done to investigate what strategies and learning behaviors students can undertake to make those failure experiences productive. This dissertation investigates what kinds of behaviors expert learners (in the form of graduate students) employ when encountering failure that predict future performance (Study 1), and whether such effective behaviors can be provoked in less sophisticated learners (in the form of high school students) that would subsequently lead to deeper learning (Study 2). Study 1 showed that experiencing and responding to failures in an educational electrical circuit puzzle game prior to formal instruction led to deeper learning, and that one particular strategy, "information-seeking and fixing", was predictive of higher performance. This strategy was decomposed into three metacognitive components: error specification, where the subject made the realization that a knowledge gap or misunderstanding led to the failure; knowledge gap resolution, where the subject sought information to resolve the knowledge gap; and application, where subjects took their newly acquired information to fix their prior error. In Study 2, two types of prompts were added to the educational game: one that provoked students through these metacognitive steps of error specification, information seeking, and fixing, labelled the "Metacognitive Failure Response" (MFR) condition; and a second prompt that provoked students to make a global judgment of knowing, labelled the "Global Awareness" (GA) condition. The results indicated that although there were no significant condition differences between the three groups (MFR, GA, and control condition where participants received no prompt at all), more time spent on the MFR prompt predicted deeper and more robust learning. In contrast, more time spent on the “Global Awareness” prompt did not predict deeper learning, suggesting that individual factors (such as conscientiousness) did not alone account for the benefits of time spent on the MFR prompt on learning. These results suggest that while MFR participants who carefully attended to the metacognitive prompts to specify the source of their errors and seek information experienced learning benefits, not all MFR participants sufficiently attended to the prompts enough to experience learning gains. Altogether, this body of research suggests that using this “error specification, info-seeking, fixing” strategy can be effective for making failure productive, but other instructional techniques beyond system-delivered prompts must be employed for full adoption of this metacognitive response to failure. Implications for teaching students to respond effectively to failure, for games in the classroom, and for design and engineering processes are discussed.

Table of Contents

List of Tables

List of Figures

Acknowledgements

This dissertation is dedicated to the three most important women in my life: my mother, who sacrificed all she had and more, so that we could pursue our happiness; my sister, who taught me what unequivocal love, support, and honesty could do to lift a soul and render strength; and my aunt, who groomed me from the start to reach for the stars and take no prisoners, expectations be damned. You are the mirrors through which I see myself most clearly, and I'll always be grateful and humbled for your presence in my life.

To the other two-thirds of the Δ, Jenna Marks and Mel Cesarano: you guys give me LIFE. We would've never gotten here without our friendship to carry us through, and I am all the intellectually and emotionally richer for it. To think it all started with a pocket square! Ben Friedman, Laura Malkiewich, Michael Swart, Sorachai Kornkasem, Ilya Lyashevsky – your intellect, and, above all, your friendship, helped shape the researcher and the person I've become. Together, you've made these past five years the most intellectually stimulating and personally rewarding I've ever experienced – what a cohort to learn and grow with!

I am grateful to Dr. John Black, who saw enough promise in my early work to fund and support me through my Ph.D – thank you for investing in and guiding me from the start, when I had such a ways to go! A special thank you to the Ben and Grace Wood Fellowship, who supported my doctoral work. I am also grateful to Dr. Catherine Chase, who challenged me to be a better thinker and more rigorous researcher, and Dr. Ryan Baker, who never failed to be a supportive and encouraging mentor. I'm so lucky to have studied under all of you.

With special mention to Lilia Klubuk and Paul Wargo, I am also thankful for my high school alma mater, who enthusiastically supported my research. It was a joy and a surreal experience to be collecting data at the very same institution that helped shaped the academic I would become.

Finally, to my partner and love of my life, Ben Dunn: thank you for your love and support. You are my rock, biggest cheerleader, and my closest confidante. We've come so far; we've an even longer road ahead! To the next great adventure…

Theoretical Framework

Introduction

When failure is discussed in both the public and academic space, the emphasis is largely on affective responses to failure. From Cinderella stories like JK Rowling and Steve Jobs, to the seminal works on mindset (Dweck, 2006) and grit (Duckworth, Peterson, Matthews, & Kelly, 2007), the question often asked is how we can encourage people to persist in the face of failure, given that failure is often on the pathway to success. However, little has been done to identify when and how failure can be useful – that is, what are the kinds of "necessary and sufficient" conditions of the task, the learner, and the instructional method to make failure productive?

Some researchers have investigated how failure can be beneficial for future understanding or outcomes. Manu Kapur (2006) introduced a framework for interventions called "productive failure", where he posited that students who have the opportunity to grapple and fail within an open-ended environment more deeply understand the formal concepts from the direct instruction that follows. Loibl and Rummel (2014) further asserted that the affordance of this type of failure is that it provides a global metacognitive cue to the learner – the realization that one doesn't quite understand the system fully, a gap in knowledge – that can then be resolved in formal instruction later on. However, these researchers also fail to answer a fundamental question about what happens *in the moment of failure* – what kinds of information, and what kinds of actions, can people leverage to optimize their understanding? For example, are there certain kinds of reflection that should be happening in the failure space, or perhaps kinds of actions that need to be taken in response to failure? What about the role of acknowledging and pinpointing what caused the failure, which may be a vital part of noticing and understanding the features of a concept?

To elucidate the questions surrounding the utility of failure for learning, I outline several areas of the learning sciences pertinent to them. First, I discuss theories of transfer, particularly the theory of Preparation for Future Learning (PFL), where experiences in novel environments can inform learning that follows. In particular, I discuss the implications of productive failure that occur in PFL activities, and how they suggest that actions and intuitions that develop during failure and exploration might be key to robust learning later on. I will then discuss the field of metacognition, and how prior work on metacognitive judgments and strategy selection in learning spaces informs the questions surrounding the utility of failure for future learning. I also elucidate the ways in which failure at large can be effective for learning, and the gaps in research that fail to specify what or how people make use of information provided during failure. Finally, I argue why games in particular are useful both as a place to study failure, and as an effective intervention for preparation for future learning, drawing from the realm of games research and embodied cognition. Together the confluences of these fields form a theoretical framework to ask the following questions: Is experiencing failure during exploration of a problem space (in a game) prior to formal instruction (PFL) beneficial for learning? Are there specific kinds of responses to failure that are more effective for learning than others?

Transfer and Learning

Transfer, or the ability to take what is learned or experienced in one context to use in another, is a fundamental goal of learning. However, a century of investigations and epistemological debate has yielded ambivalent conclusions about precisely what constitutes and yields transfer of knowledge (Barnett & Ceci, 2002; Detterman, Sternberg, & Norwood, 1993; Klahr & Chen, 2011; Singley, Anderson, & Cambridge, 1989; Thorndike & Woodworth, 1901). Recently, investigations in the field of the learning sciences have led to novel approaches in the instruction, definition, and utility of transfer for learning and skill development. In the Preparation for Future Learning paradigm, transfer is defined as the use of prior experiences to inform and improve later formal learning, as opposed to a one-to-one mapping of prior content to a novel context (Bransford & Schwartz, 1999) The idea is that we oftentimes use prior knowledge to notice and frame new information, and that these "knowing with" kinds of prior experiences can greatly shape and improve understanding of the new context. For example, consider Broudy's example of asking people to address the problem of repopulating eagles, illustrated by Schwartz, Bransford, and Sears (2005). Schwartz et al. (2005) demonstrated that when considering people's proposals for eagle repopulation through the lens of PFL, adults used their prior experience and knowledge about repopulation of other animals and ecosystems to ask more effective questions and strategize more appropriately, even if their solutions were not especially sophisticated. This predilection to frame and infer the important aspects of a problem because of prior experiences can lead to faster, deeper learning than if one did not have these experiences; in other words, having relevant, familiar prior experiences can *prepare* you to ask the right questions, notice the important components, and learn more deeply from novel content (i.e. future learning). As such, transfer in PFL is treated as both an *experience* that can foster deeper learning (i.e. transferring out), and as a *measure* of learning (i.e. transferring in). Through this lens, we can look at transfer as a process, where accessing prior experiences and information is a practice to be cultivated for more effective learning, rather than only looking at transfer as an indicator that learning has occurred. While PFL may seem to be a common-sense approach to education, PFL as a framework for pedagogy is a relatively new approach. A variety of studies in recent years have shed light on what kinds of activities could provide the most effective experiences to prepare students for future learning.

Productive failure as preparation for future learning. PFL studies highlight the utility of activities that allow students to explore and grapple with relevant content prior to instruction. Oftentimes, these activities are explicitly designed such that students are thrust into a problem-solving environment that compels them to wrestle with underlying principles of the concepts. Consider, for example, Schwartz, Chase, Oppezzo, and Chin's (2011) work on using invention with contrasting cases as preparation for future learning. In their study, students had to invent a formula that captured the concept of a ratio, using cleverly designed cases that deliberately highlighted and contrasted underlying factors of ratio structures (for example, space and number of items). Their work demonstrated that even when students aren't always ultimately successful in their invention of the ratio formula, their experiences with invention prior to formal learning led to better transfer outcomes on two dimensions: better formal learning of ratio structures in physics, *and* better application of this ratio structure to other domains. This finding demonstrates that students who explored the underlying principles of ratios through invention learned, abstracted and applied these concepts better than those who took the "traditional" route of learning first and then practicing. Their conclusions suggest that a critical mechanism of this PFL activity, inventing with contrasting cases, is fostering an "appreciation of the deep structure" of the concept such that students readily called upon their experiences with this deep structure when learning about the formal concept later on. However, they do not discuss what specific mechanisms of invention-with-contrasting-cases led to greater noticing of the deep structure. Thus, there are some questions not yet answered: what is the role of iteration, failure, strategy, and realizations about insufficient solutions in noticing these deep structures?

Manu Kapur's (2008) work with PFL attempted to address some of these questions by isolating failure as a vital component of preparing students for future learning. Kapur used the

PFL framework to design an intervention using either well-structured, scaffolded problems or ill-structured problems prior to formal learning, His work revealed that despite students in the ill-structured condition struggling with defining, analyzing, and solving their problems (in other words, failing to generate explicit understanding of the concepts or effective solutions), these experiences were more conducive to learning later on. This phenomenon, which he called "productive failure", demonstrated that success in the traditional sense (that is, success in clearly defining concepts and generating effective solutions) may not necessarily lead to greater learning; in fact, designing problem solving tasks that scaffolds and directs learning towards "success" may unwittingly undermine the effortful cognition that could benefit formal learning later on. Instead, environments and tasks that permit students to fail and grapple with concepts rather than "succeed" can be more beneficial to future learning that follows those failures. Instruction that "teach to the failure", or address how students' incorrect solutions are actually instantiations of the concept, are crucial to transfer and deeper conceptual understanding. However, Kapur also fails to elucidate what specifically in the productive failure space led to greater learning. Does failure in itself call attention to deep features? Or does failure afford opportunities to engage in cognition that then leads to deep feature noticing? How much failure is sufficient for PFL?

Metacognition

One possible explanation for how failure might lead to later success is the role of metacognition, or the ability to judge and monitor one's own states of knowing, and employ strategies to improve understanding. Metacognition is a critical part of learning because it permits learners to identify, more deeply understand and effectively address gaps in knowing (Flavell, 1979). Flavell's seminal theory on metacognition splits metacognitive activities into two types:

metacognitive knowledge, and metacognitive experiences (or regulation). The former, metacognitive knowledge, is the information the person knows or has acquired about the selthe task, and the strategies one can employ to go about solving that task. For example, a student who is taking a quiz might have prior judgments on what she knows, how hard the questions are, and what the most effective ways to solve certain quiz questions are. Applied to the space of failure in games, a player might have metacognitive knowledge about themselves (How adept of a player am I? Do I know enough to beat this level?), of the task (How difficult is this game? How is this different or similar to other games I've played before?), and of the strategies she might employ (What is the best approach to the problem, given what I know and how hard this is?). Metacognitive experiences, on the other hand, are the phenomenological acts of enacting metacognitive knowledge – that is, the moment when that student does make a judgment about herself or the task, and the strategies she employs in the moment. In metacognitive experiences, one can engage in two kinds of metacognitive activity: metacognitive monitoring, and metacognitive control (Son & Schwartz, 2002). Metacognitive monitoring is the on-line or ongoing appraisal of one's own understanding or performance, using one's own metacognitive knowledge to make judgments (correctly or otherwise) about how one is doing. Metacognitive control is the regulation of one's strategy or behavior based on one's monitoring, such as deciding to approach a problem differently after realizing that an earlier approach is ineffective. Throughout a metacognitive experience, the student may make predictions about her performance (metacognitive monitoring), and use such information to allocate effort and attention towards the intended goal (metacognitive control). To extend the earlier game example, in the moment of failure, the player may have a metacognitive experience, where she employs her metacognitive knowledge to evaluate her own performance, and make a decision about what to do next – for example, to try and fix her earlier solution. Yet, it's also possible that the player might simply employ this action – fixing her solution – because it has worked for her in the past, and not because of some conscious and deliberate appraisals about herself or the task (Borkowski & Muthukrishna, 1992; Davidson, Deuser, & Sternberg, 1994). As such, it is especially difficult to determine in metacognitive research whether a participant's actions in a task are reflective of metacognitive activity without explicitly asking them to report metacognitive intent.

Further advancements break the field of metacognition down into various dimensions: meta-memory, meta-comprehension, and meta-strategic knowledge. For the purposes of this paper, we will focus on the topic of meta-strategic knowledge, particularly in the domain of problem-solving and STEM learning. Siegler's (1994) seminal work on children's strategy use shows that young children vary widely in their strategy use (for example, using several strategies on the same problem, or different strategies on different problems). Strategy selection for these children often followed predictable "overlapping wave" patterns, where they first used a variety of strategies, and then repeated useful ones, discovered new strategies, or abandoned others to hone in on the most effective approaches. Furthermore, effective strategy use often required users to inhibit less advanced (and perhaps more habitual) strategies in order to employ more sophisticated (and less familiar) strategies (Kuhn & Pease, 2010). Kuhn & Pease (2010) argue that the process of "constructing, implementing, and monitoring" a more sophisticated strategy, which requires ongoing metacognitive monitoring and control, is distinctive from inhibiting a less effective one. Applied to the context of failure in games, this means that players could inhibit less effective prior approaches to the game space, but not generate and test new, more effective strategies towards completing the level. But a question still remains unanswered: are there specific kinds of strategies that are more related to later learning afterwards, not just success on the immediate problem? In other words, if strategy selection is a vital part of problem solving, and insights from problem-solving experiences can be transferred to later learning (PFL), are there specific kinds of strategies during the problem-solving phase that are particularly good for transfer?

Metacognition and transfer. Metacognitive processes are commonly discussed as a critical component of teaching students to transfer because the act of self-monitoring helps facilitate the recognition of when the information or strategy might be relevant in other contexts (Adey & Shayer, 1993; Belmont, Butterfield, & Ferretti, 1982; Perkins & Salomon, 1992). Each moment of failure affords an opportunity to make a metacognitive judgment about what knowledge component is lacking. Metacognition could presumably occur in two places in transfer – during the "transfer out" component, where students can monitor and reflect what kinds of information they're processing right now might be useful in the future, or during the "transfer in" component, where students could review what prior strategies and knowledge could improve performance in the current context. In preparation for future learning activities, the utility of metacognition primarily lies in the "transfer out" phase, where during the exploration phase students reflect on what features of the problem space are important to pay attention to. It can be argued that students who act in more reflective ways during the PFL activity, whether those metacognitive behaviors are enacted by natural predilection or provoked by the environment, would attend more carefully to deep features and therefore will be more prepared to learn from future learning activities. Furthermore, metacognition is especially valuable in failure spaces, because the most "productive" affordance of failure is to address head-on what those gaps between expected and actual outcomes are, and what actions should be taken to resolve them (Loibl & Rummel, 2014). Presumably, what makes productive failure good for preparing students for future learning is contingent on students' abilities to reflect on their incorrect solutions, address gaps in knowledge, and select strategies and actions in response to these appraisals. It is through these metacognitive monitoring and control mechanisms that cue students to identify and engage with deep features of the concepts. This is another key investigation: how can we affirm the role that metacognition plays in the efficacy of productive failure activities for PFL? Does metacognition globally impact the efficacy of PFL activities? Is the utility of productive failure activities contingent on students' metacognitive behaviors?

Failure

The topic of failure as beneficial for later success is not a novel one. The fields of engineering and design, for example, have long accounted for the possibility and benefits of failure in the design process. In engineering, failure analysis engineers oversee the evaluation of what specific errors or failed components in a product caused the failure to inform future designs, while top design firms like IDEO tout the "expectation of failure" as a normalized part of the change process. In both of these contexts, failure is an expected and well-documented phenomenon that affords the opportunity to provide critical information about the quality of the current product, explore or test the limitations or parameters of the system, and develop further insight and inferences to inform future products. Implicit in these approaches is the idea that the developers – engineers and designers in this case, but also anyone in the role of problem-solving, like students and teachers – can appropriately recognize and use the information produced by a failure to improve future performance. Yet for novice designers and learners, this implicit process is not so intuitive – you must possess the resiliency to look at that failure as an opportunity rather than a marker of (in)ability; you must be invested in the end product enough to want to use such information to improve your solution; and you must enact or develop the kinds of metacognitive monitoring and control skills required to interpret the information provided from the failure and act accordingly. The most pertinent requirement of the three is the question of what skills and behaviors must one enact in response to failure to make use of the information afforded, so as to improve future insight and understanding. In short, how does one make failure productive? Are there ways of designing tasks that promote productive failure, as Kapur's work suggests, rather than just plain failure (that is, if there is even in fact a difference between regular failure and productive failure)?

Loibl and Rummel (2014) addressed some of these questions by asserting that productive failure improves learning by calling students' attention to the gaps in understanding when they confront a failure. In their work, they demonstrated that attempting to solve problems before formal learning can lead to a global awareness of knowledge gaps - that is, acknowledging that some component of their understanding is incomplete without specification. This awareness is a kind of global metacognitive judgment that arises from students' inability to solve the problems (a failure), that are then fully specified and addressed in teacher instruction. Consider the differences between their approach to the "benefits of failure" and that of engineers and designers' approaches: Loibl and Rummel argue that it is not important for students to successfully specify where their understanding breaks down or is lacking, while engineers and designers insist on that specification in order for future products to improve. The process and goals of their approaches also differ: PFL at large is interested in the acquisition of deep conceptual understanding, while designers and engineers emphasize knowledge gleaned from failure *in the service of* an end product. Yet, it would seem that the specification of knowledge gaps and errors, such as those made by engineers and designers, would be a critical contribution towards deep conceptual understanding. While Loibl and Rummel discuss global knowledge gaps (global metacognitive awareness) as a mediator for failure to positively impact learning, they do not explicitly discuss moment-to-moment response-to-failure behaviors that can also be productive for deep conceptual understanding. As such, a critical question posed by this research is whether the specification of one's own gap in understanding or errors (metacognitive monitoring), as well as the actions that follow such specifications (metacognitive control), are a vital component of deeper learning later on.

Research on impasse-driven learning (K VanLehn & Springer, 1988) highlighted that when learning to use a procedural skill, students may employ simplistic "repair" or "help-seeking" strategies, and that these strategies can then be integrated into the larger sequence of procedural approaches they employ, thus expanding their proficiency on the skill. Impasse was defined as the moment at which a student could not go further in their problem solving because of a gap in prior experience or lack of procedural knowledge necessary to complete that task step. Teachers, learning materials, or intelligent systems can provide timely help information to learners that can help get them past an impasse and provide additional steps for approaching a problem. Thus, VanLehn and Singer (1988) argued, procedural learning only occurs at impasses because students must recognize that there is a limitation in their current capability that prevents success, and remedy it through information seeking. This suggests that there is an optimal - indeed, perhaps only productive – way to respond to an impasse (or failure). Yet there are also two limitations to VanLehn's theory: first, he references only procedural skill development – specifically, problem-solving skills- through this framework, and does not explicitly discuss how declarative knowledge about the underlying system of the problem is developed through impasse. Secondly, while he elegantly discusses the strategies and conditions that students may use to help-seek to get through an impasse, he does not describe these strategies in the context of metacognitive judgments of one's own knowledge and the specification of what knowledge is missing. In other words, how do procedural impasses relate to declarative understanding of the underlying content, and what role does error or impasse specification play in this type of learning?

Failure and Motivation. While failure can be beneficial to learning because it evokes an element of metacognition – that is, it forces students to realize that they don't know something as well as they thought they did – it can also be detrimental to student motivation. Students are often intimidated by failure, particularly in school tasks where failure often involves high-stakes consequences, such as failing a quiz or getting a low score on your homework. There is a bevy of motivational constructs related to failure, and whether students are willing to persist through them. Student self-efficacy (or their sense of competency – see Bandura, 1994) and perceived difficulty of academic tasks (Darnon, Butera, Mugny, Quiamzade, & Hulleman, 2009) may impact whether students expect to be successful at the task or not, and therefore impact whether they are willing to put forth the effort to try or persist. Goal orientation (Pintrich, 2000) and mindset (Dweck, 2006) can also significantly impact students' willingness to persist through failure, because failure carries different connotations for students with different goals and mindsets. For example, Belenky & Nokes-Malach (2012) found that students with mastery goals, or goals that center on understanding and skill development, rather than performance goals, which center on demonstration of competency (Dweck & Leggett, 1988), benefit more from PFL activities because their mastery approach goals allow them to shift their attention to deep features of the task, rather than fixating on merely performing well. In contrast, those who have performance goals might be demotivated by the failure, because it did not demonstrate their skill successfully. Similarly, those with fixed mindsets may treat failure as an indication of a fixed ineptitude they don't have the capacity to change, while those with growth mindsets may treat the failure as an opportunity to improve with time and effort (Dweck, 2006). Furthermore, common school tasks do not often permit or encourage efforts to respond to those failures - that is, they don't provide the tools, encouragement or opportunities for students to review their incorrect solutions, appraise where knowledge gaps occur, seek to close such gaps, and fix their solutions. What curricular tools might provide low-stakes, engaging problem-solving environments that encourage student iteration, permit for metacognitive behaviors, circumvent motivational concerns about student beliefs and goal orientations, and allow for exploration of academic content in meaningful, goal directed ways?

Games and Learning

One possible way to address the issues of student metacognitive ability and motivation is to couch the productive failure tasks in a game. In fact, games are particularly well suited for investigating questions about productive failure as a key component for effective PFL because they provide the right motivational benefits for engaging in failure; because they provide a space for exploring and manipulating content in a situated and realistic way; and because they offer affordances for learning in problem space through game mechanics that allow for a variety of metacognitive responses. Failure is a critical component of games, where the process of failing (a level, a fight, a boss, a puzzle) is inherent in the game design in order for it to be compelling and entertaining. People appear to be incredibly productive when encountering failure in games, where they use the failure experience to inform future decision-making and understanding of the problem space (Juul, 2013). These kinds of metacognitive behaviors - reflecting, judging the goodness of one's performance, coordinating strategies, planning next actions to address what went wrong previously - are ones we strive for students to employ, but are enacted so naturally in game environments. Furthermore, game spaces seem to promote resilient behaviors in the face of failure - perhaps because the failures do not have high stakes (outside of the game), and therefore does not negatively impact motivation. On the contrary, despite deliberate designs for inducing failure, games seem to encourage engagement and persistence, even (and perhaps especially) when the player is frustrated and confused. Therefore, the game space is a valuable space for us to investigate what cognitive mechanisms are at play in failure that are good for future learning, while alleviating the concerns about motivation and the high-stakes nature of failure in school tasks.

While game spaces offer the opportunity and incentive to engage in effortful behaviors in response to failure, they also offer a wide variety of cognitive and motivational benefits for student learning in general. Situating exploration, problem-solving, and systems manipulation in a game can be a powerful method for generating intuitions about a particular concept or system (Garris, Ahlers, & Driskell, 2002; Honey, Hilton, & Washington, 2011). Situated cognition theorists posit that all learning naturally occurs in situ, and that situated grounded experiences are the most effective ways for students to explore and deeply understand concepts and develop skills (Brown, Collins, & Duguid, 1989). Games can provide these experiences, particularly for content that is difficult to directly experience in real life, such as science systems that are invisible to the naked eye, or happening at temporal and spatial scales well beyond human scope (Halverson, Shaffer, Squire, & Steinkuehler, 2006; Honey et al., 2011). These authentic environments allow players to systematically build and experience “cycles of expertise” in realistic contexts by systematically presenting and scaffolding skill development, first in isolation, and then interwoven with other previously learned skills, to produce a host of flexible and dynamic skillsets, and to build the capacity for metacognitive strategy selection and appraisals of problem spaces(Gee, 2005). Embodied cognition theorists extend this further by asserting that learning is most effective when enabling the body's sensorimotor and perceptual faculties – that is, when the learning is embodied by the agent within the contextual environment. Game spaces allow such embodiment by allowing players to explore with surrogate agents within a constrained environment, directly interacting with, perceiving, and manipulating objects and forces within the space through that playable agent (Clark, 2003; Fadjo, Hallman Jr, Harris, & Black, 2009). Prior work on embodiment for math and computational thinking instruction through video game environments demonstrated that surrogate embodiment can have powerful implications for future educational game design (Fadjo et al., 2009). Many commercial games, like Legend of Zelda, Goldeneye 007, and Super Mario, are classic examples of how surrogate embodiment in game spaces can facilitate problem-solving and spatial reasoning skill development. In fact, Arena (2012) found that playing one of two commercial games, Civilization IV and Call of Duty 2, not only prepared students to learn more about World War II than just through instruction alone, but also specifically increased students' understanding of global strategic elements and tactical strategic elements, respectively. Furthermore, game spaces are deliberately designed to constrain users to specific goals and system parameters structures (Black, Khan, Huang, & In, 2014; Garris et al., 2002; Malone, 1981; Reese, 2007). For example, conservation of momentum, gravitational force, and mass are all key components in the game Portal, and must be explored, implicitly understood, and mastered in order for the player to proceed. These constraints and system parameters allow students to focus on the key components of the environment that are important for understanding and problem-solving. These grounding experiences can prepare students to better learn from formal content later on (Black et al., 2014; Hammer & Black, 2009). It's also possible that these prior game experiences are then later accessed during the formal learning – that is, the student imagines the game space and the manipulations that they encountered while learning about the formal concepts, in order to bridge the two experiences together (Black, Segal, Vitale, & Fadjo, 2012). This undoubtedly is also a key component of deep learning and transfer.

However, not all games are created equal, and not all games necessarily elicit productive failure. In order for a game to adequately prepare students for future learning, the game content and mechanics need to be aligned carefully with the target learning, such that the interactions students engage with in the game permit for them to directly experience and manipulate the deep features of the concepts. For example, Math Blasters would be a poor example of a game that prepares students for future learning, because the game mechanics do not actually allow students to grapple with the underlying concepts of mathematical operators. On the other hand, Civilization has been highly touted as a great game for learning because it allows students to directly control and manipulate the factors that leads to the success and downfall of a civilization, such that these experiences could inform their learning of formal civics concepts later on (like trade, war, territory, diplomacy, and resources). As such, the selection of a game that illustrates and permits student interactions with underlying features and structures of a concept or system is vital to using games as preparation for future learning. Furthermore, the game must allow players the agency to explore, enact strategies, and respond to consequences in the environment that allows for meaningful and effortful play. Agency and affordances for choices are important not only for student-driven learning and skill development, but also for promoting motivation.

Games, above all, are touted for their educational potential because of their promise for motivating players to engage in effortful behaviors. Motivation is especially critical for metacognitive and strategic behaviors, because as mentioned previously, these behaviors can be effortful, difficult, and not obviously tied to one's performance or goals. As previously discussed, games have the potential to motivate players to engage in these behaviors as part of gameplay. Game scholars argue that video games are motivating because they promote agency and self-efficacy, provide an optimal balance between player skills and level difficulty that induces flow, use narratives and character development to induce emotional investment, and allow for socially situated practices such as collaboration or competition(Gee, 2005). These motivational factors are also closely linked to strategic responses to failure. Choice and control, critical parts of gameplay, promote intrinsic motivation (Malone, 1981), and also enable players to employ reflection and strategy selection to approach problem contexts in open-ended and user-driven ways. The use of gating (where players are not allowed to continue in the game until they've mastered a prior skill or level), scaffolding (given through simple levels and visual or auditory cues, such as arrows to direct attention or pings to indicate proximity to goal states), tutorials (often given as an introductory level, through a non-playable character explanation, or through overlays on to game levels themselves), and a progression of increasingly challenging levels (to develop the aforementioned "cycles of expertise") are common ways games place players in a state of flow (Csikszentmihalyi, 2000). Unsurprisingly, these also serve Vygotsky's Zone of Proximal Development, where learning is optimized when the learning environment and content is just outside of learner's initial abilities but within the scope of their *potential* learning, given experience and their use of pedagogical tools. Narratives, character development, and socially situated practices such as collaboration and competition also all serve to motivate players to continue engaging in the game, persisting even when encountering setbacks or failure.

Games and Failure

What, though, specifically about games makes it so conducive for investigating failure? Failure is a ubiquitous part of games (Blumberg, Rosenthal, & Randall, 2008; Juul, 2013), where the experience of failure is central to the enjoyment and advancement of gameplay, despite in-the-moment frustration. In fact, researchers have argued that impasse-driven learning is at the heart of successful gaming experiences, where games are deliberately designed to induce impasses that catalyze shifts in game strategies and techniques, thereby expanding gamers' "skill toolboxes" (Blumberg et al., 2008; Gee, 2005). Juul (2013) argues that games capitalize on the experiences of failure to capture and sustain attention because players know that with skill improvement they can overcome these failures, thereby producing feelings of self-efficacy, enjoyment, and satisfaction. This is closely related to the relationship between flow and zone of proximal development: teachers want students to persist precisely because the experience of failing and re-trying can provide insights and opportunities for skill development that might eventually lead to success, but success is only possible if the problems the students face are within their ability, and only if the students believe that their failure experiences actually lead to skill improvement. In other words, the tension that Juul points out is what induces flow for gamers, the frustration in-the-moment of failure and the enjoyment that arises when that frustration is resolved with success, is only possible if the game provides an experience that is within the players' zone of proximal development – the "sweet spot" of skill and challenge. Thus, opportunities for failure, flow, and optimization of problem difficulty with student ability is also an important consideration for the selection of an appropriate educational game. A final consideration for why games are conducive for studying failure is that they provide a low-stakes environment for student to engage in failure. Academic activities can be intimidating, because student performance on these tasks are used for student grades and may be subject to scrutiny by the students' peers, teachers, and family. As such, these high-stakes tasks provide very little incentive for students to allow themselves to engage in risky, exploratory behaviors, or might lead them to disengaging from the activity altogether. Games typically don't carry these implications in a classroom setting; students often treat their identities as "gamers" (and that set of motivations, confidence, and dispositions) very differently from their identities as "students". Because failure is such a naturalistic part of gameplay, players are not intimidated by failure in these environments and therefore are more willing to engage in a wider range of exploratory, strategic behaviors.

The assertion that failure in games are compelling because it points out an inadequacy that we must resolve (Blumberg et al., 2008; Juul, 2013; K VanLehn & Springer, 1988) – the central thesis of impasse-driven learning - is rooted in metacognition. Take, for example, a player that is stuck on a level in *Little Big Planet*. The player's frustration stems from the increasingly clear realization that he is inadequate – that a part of his understanding of the problem space is missing, or that he has not yet mastered a skill necessary for that level. He then tries varying kinds of actions: he repeats the same approach several times, paying attention to how far the character jumps, what cues are in the environment that he can use to pull himself up, the tools afforded to him in the environment; the timing of his moves. Eventually, he switches tactics several times, and finally manages to get his agent to the top platform and move on. Within this sequence of actions, we see several metacognitive steps arising: he makes the appraisal of his own skill and performance; he analyzes and attempts to specify what part of his performance or strategy is lacking; and he chooses strategies and behaviors to enact based on these judgements (i.e. persisting with one approach, or trying new ones). These metacognitive behaviors are crucial to Juul's analysis of failure in games: players must systematically approach, glean information from, and enact behaviors in response to failure effectively in order for them to experience satisfaction from eventual success. In other words, these productive responses to failure are crucial to successful gameplay, and to the enjoyment of games. The fact that gamers naturally enact these productive responses to failure are precisely why games are such an optimal environment for studying what kinds of response-to-failure behaviors are most conducive to later learning.

Games are also an optimal environment for studying responses to failure because it allows for logging of actions taken during the problem-solving process. Using log data to track student behaviors and strategies are becoming the new standard for studying in-situ cognition; researchers have used log data to study everything from knowledge states (Corbett & Anderson, 1994), affect (Baker, D'Mello, Rodrigo, & Graesser, 2010), help-seeking (Roll, Aleven, McLaren, & Koedinger, 2011), to game and learning strategies (Rowe, Asbell-Clarke, et al., 2015). Using timestamped data and clickstream actions that players take, we can engineer an innumerable amount of features to capture game behaviors and problem-solving strategies. For example, using log data of student help-seeking actions, Aleven et al. (2006) developed a metacognitive computational model of help-seeking behaviors to detect when students were engaging in productive or counterproductive metacognitive behaviors when solving geometry problems in an intelligent tutoring system. They were able to identify when a student was seeking help effectively, avoiding help purposely, or abusing help options (like bottom-out hints) using log data in real-time. Thus, there is empirical precedence for using log data to detect and operationalize metacognitive actions. Another example of log-data use, this time in the realm of games, is using stealth assessments in games to measure implicit learning (Rowe, Baker, &

Asbell-Clarke, 2015; Shute, Ventura, & Kim, 2013). Rowe et al. (2015)triangulated log data from the game, videos that were coded for strategy use, and post-test data to validate detectors of implicit understanding of Newton's Laws in a particle simulation game. In particular, they used log data to determine when students were employing specific strategies that implied understanding of Newton's Laws, and mapped them on to learning data (such as the number and location of clicks) to investigate the relationship between student learning and strategy use. We can use a similar approach to detect when particular strategies – this time, in response to failure – are related to metacognitive judgments and later performance.

Theoretical Framework

Failure and success might seem to be diametrically opposed, but research and methodologies in learning theory, STEM, games research, and design highlight that failure is oftentimes critical for later success. Furthermore, the Preparation for Future Learning (PFL) and Productive Failure paradigms indicate that exploring, grappling, and failing in open-ended problem-solving tasks prior to formal instruction improves one's capacity and preparation for understanding the formal content later on. Yet, little is known about what specifically in the experience of failure is, in fact, conducive for later success – whether that success is a solution (i.e. a product of some design process, or a solution to a problem) or deeper understanding (i.e. conceptual knowledge). Is experiencing failure unto itself sufficient for later success, or is there something else that must happen in the space of failure that makes it conducive for later success?

One possible explanation for how failure can be beneficial for later success is metacognition. Metacognition, or one's ability to regulate one's own thinking and strategies, may play a critical role in making failure productive. Failure also makes the need for metacognition explicit, because it provides an external cue that highlights an inadequacy or incorrect judgment made on the part of the learner that needs adjustment. As a result, experiencing failure provides opportunities for developing greater insight, both about the task at hand, and about one's own understanding and capacity. This chance to engage in a metacognitive experience, especially when confronted with a shortcoming, is an opportunity for the student to monitor their own understanding and capacity, and select or develop alternative strategies and approaches towards the problem space. In the process of monitoring and controlling one's own cognitive processes, students also make implicit realizations about the task at hand, and the knowledge required to complete them. But does this understanding happen so long as you experience failure, or are there specific strategies one can take that help facilitate understanding, both in the moment of the failure, as well as later on, during formal learning? In other words, how do metacognitive monitoring and control behaviors relate to deep conceptual understanding later on?

Another concern relating to the utility of failure is whether students are motivated to engage in productive behaviors when presented the opportunity. Failure in school-like tasks are often high-stakes, and carry with them loaded implications for intrinsic and extrinsic motivation. Furthermore, many school tasks, such as tests and projects, allow little opportunity to engage in metacognitive behaviors in response to shortcomings. Thus, it's difficult to investigate the utility of particular behaviors and strategies in response to failure in traditional school tasks because students are not compelled or afforded the opportunity to engage in such behaviors and strategies. Instead, games are proposed as an alternative that is well-suited for investigating what responses to failure are productive for later learning.

Games are an excellent space for studying responses to failure for several reasons. First and perhaps most importantly, failure is a ubiquitous part of gameplay. Players expect to fail, and therefore are resilient and proactive in the face of failure. In fact, it is the frustration that is experienced during failure, and the resolution of that frustration when the player finally succeeds, is what makes games so enjoyable and gratifying. Secondly, game spaces provide an open-ended but constrained environment that allows players to take a variety of actions within an environment specifically designed to highlight a particular mechanic or system parameter. This allows players to employ a number of approaches and strategies when attempting to solve a game level, all the while observing, evaluating, and testing the parameters of the underlying game system. Thirdly, game spaces serve as a suitable PFL activity, especially when the game allows for goal-directed exploration of the content to be learned later on. Games are especially powerful for PFL when it situates students into the learned content, allowing for manipulation, testing, and direct observations of the system that can serve as grounding for later learning.

Together, this theoretical framework allows us to pose the following questions: What kinds of behaviors in response to failure in game environments are most related to learning later on? Are these behaviors driven by metacognition? If so, how can we induce students to engage in metacognition during failure in game spaces, such that they produce implicit understanding that is beneficial for later learning?

Study 1 Design

Research Questions

Given these threads of research on games, preparation for future learning, productive failure, and metacognition, the aim of Study 1 was to investigate whether the ability to respond to failure in a physics game better prepares students for future learning. Thus, I asked the following questions:

$RQ_{1:}$ Does the affordance for responding to failure within an educational game elicit deeper conceptual understanding and transfer?

RQ_2: Are there particular responses to failure that are better for learning and transfer from a game?

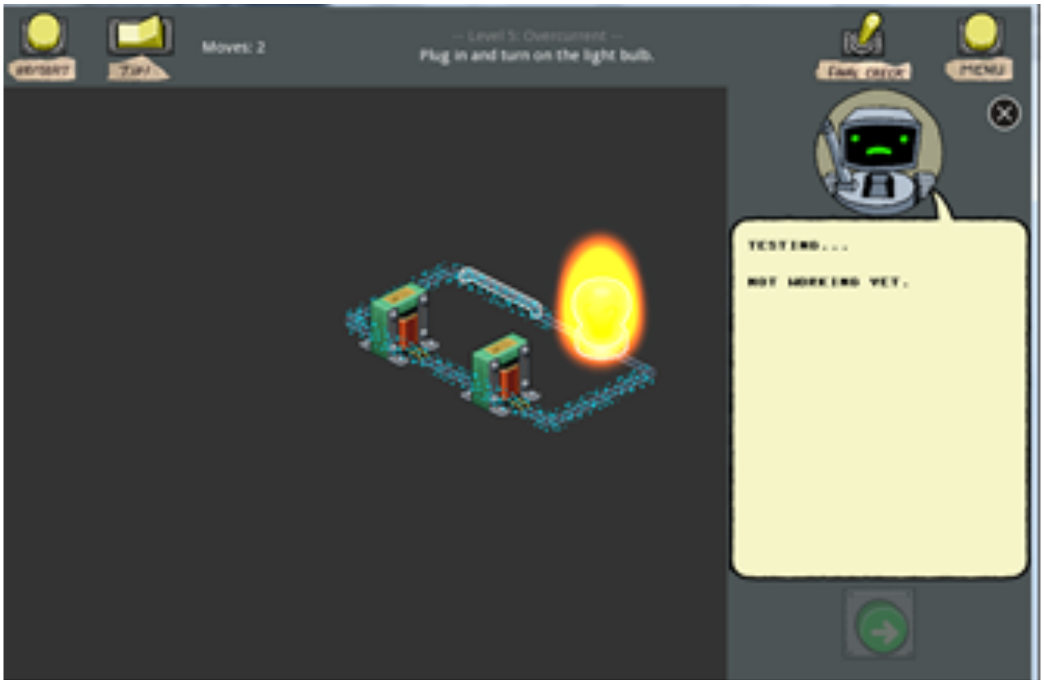
I hypothesized that (H_1) students who have the opportunity to response to their failure within an educational game prior to instruction (Failure Response, or FR participants) will perform better on measures of learning, compared to those who do not have the opportunity to respond to failure in the game (No Failure Response, or NFR participants) and those who play the game after instruction (Tell and Practice, or TnP participants). Furthermore, I hypothesized that (H_2) log data analyses would reveal particular game behaviors enacted in response to failure that were related to deeper learning.

The Game: Electropocalypse

A game was selected using the following criteria: 1) it must be a problem-solving, puzzle-based game; 2) it must include electricity and magnetism (E&M) concepts; and 3) it must allow for a variety of solutions and actions that players can take in response to failure. Puzzle games that allow players to directly simulate and manipulate a system is necessary because while agent-centered games can allow players to explore a space or interact socially with others through a surrogate (Fadjo et al., 2009), grounded interactions, like being able to directly control parameters of a system or manipulate and build parts to problem-solve allow players to directly interact with a concept (Black et al., 2012). This type of simulation interaction can be powerful for allowing students to directly observe, control, and develop insights about a system in a meaningful and constrained way (Honey et al., 2011). Secondly, while physics concepts that are grounded in real-world interactions (like conservation of energy or gravity) are easier to understand because learners have direct experience with them in their everyday lives, abstract systems (like E&M) that are invisible, happening at micro- or macro-level scales, or involve multiple complex features can be more difficult for students to deeply understand. Anecdotal evidence suggests that even for high-performing students, electricity and magnetism is a difficult topic. While over 53,000 students registered for the AP Physics C Mechanics exam in 2016, only 27,000 students –about half of the Mechanics registrants – elected to register for the AP Physics C Electricity and Magnetism exam. Electropocalypse is a commercially available mobile/PC game developed by Stratolab, and features a narrative that takes players through several scenarios (such as a power outage messing up the power grid in a city, or disabling a bomb) involving electrical engineering to solve problems, each involving another physics principle (for example, shorting a circuit, resistance, circuits with resistors in parallel, etc.). In Electropocalypse, players reconfigure electrical circuit puzzles by adding or removing wire, changing the position of switches, resistors and batteries, and measuring voltage and resistance to meet level goals.

Game Versions. The standard version of the game allows players the normal allowances for responding to incorrect attempts, such as looking for hints, fixing their solution, restarting the level, or exiting the level (see

Figure 1). In other words, when a player fails (i.e. submits an incorrect solution) in the standard game, they are free to do as the please in response to that failure. The No Failure Response (NFR) version of the game constrains players, where if they fail (submitted an incorrect solution), they were not permitted to view or fix their solution, but instead received an explanation and screenshot of a correct solution (see
Figure 2). The feedback shows the ideal solution, along with an explanation of the concept in the level. They cannot see or fix their prior circuit solution, nor can they restart the level. This NFR game version was created so that I could isolate the effects of being able to respond to failure without manipulating the naturally-occurring amount of failure participants experienced when playing each level for the first time. Furthermore, this game version has face validity because it mimics the structure of many common classroom activities, where students receive a grade or marks on their assignment indicating whether their solution was correct or not, but do not have the opportunity to fix or respond effortfully to those solutions. In some cases, teachers may review the correct answers to homework or quiz solutions with the class, which mirrors the explanation and screenshot provided to participants in this game version.



Moves: 2

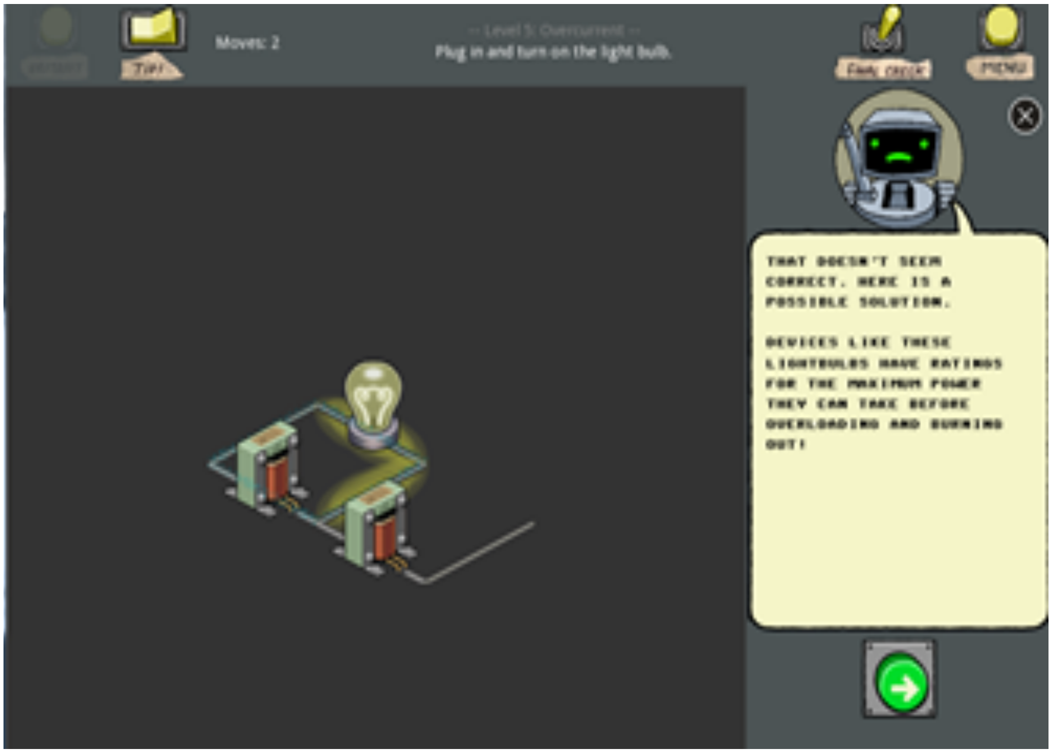
-- Level 5: Overcurrent --

Plug in and turn on the light bulb.

TESTING...

NOT WORKING YET.

Figure 1: Standard Feedback Screen.



Moves: 2

-- Level 5: Overcurrent --

Plug in and turn on the light bulb.

THAT DOESN'T SEEM CORRECT. HERE IS A POSSIBLE SOLUTION.

DEVICES LIKE THESE LIGHTBULBS HAVE RATINGS FOR THE MAXIMUM POWER THEY CAN TAKE BEFORE OVERLOADING AND BURNING OUT!

Figure 2: NFR Feedback Screen.

Study Design

36 adult participants were recruited from a non-random convenience sample to participate in a 3-hour long experimental study with random assignment to one of three conditions. 83% were pursuing a graduate degree, and all of them held at least a bachelor's degree. All participants reported low prior knowledge of the concepts covered in the study, although they also reported having completed at least a high-school level physics course. Participants learned about basic principles of direct current circuits by playing Electropocalypse for 45 minutes and by watching Khan Academy videos about direct current circuits for 45 minutes. Participants played a subset of levels (1-13) covering content like closed/open loops, short-circuiting, and resistors in series and parallel. All participants played the game, watched four Khan Academy videos, responded to surveys, and completed three Open-Ended Worksheets (OEs) and a Post-Test. Participants were randomly assigned to one of the three study conditions: Tell-and-Practice (TnP), Failure Response PFL (FR), and No Failure Response PFL (NFR). The TnP served as the control condition where participants first received instruction, followed by the standard game (See *Figure 3*).

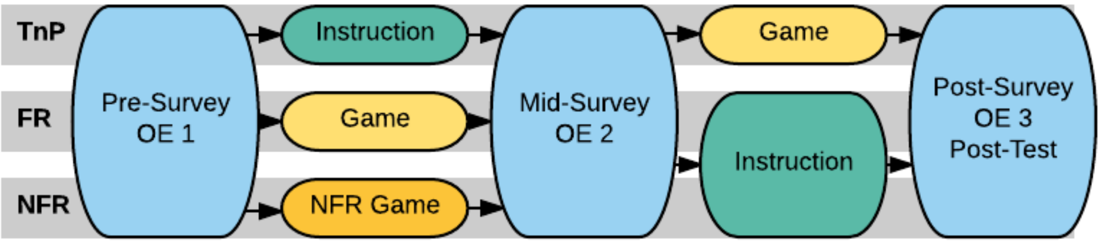


Figure 3: Study 1 Design

The FR and NFR participants played the game first, followed by the instruction. FR participants played the standard version of the game. The NFR participants played a the NFR version of the game.

Instruction. The four Khan Academy videos covered principles surrounding electrical circuits. The first video, "Introduction to circuits and Ohm's Law", covered basic electrical concepts such as voltage, current, resistance, and Ohm's Law. The second video, "Resistors in series", explored circuits with resistors in series, and showed the relationship between current, resistance, and voltage through Ohm's Law. The third video, 'Resistors in parallel", showed how current flows differently when resistors are in parallel, and how that subsequently affects the amount of voltage each resistor gets. This is a critical part of the harder levels of gameplay (and where the most failure occurred), and is most likely where the "A-ha!" moment of understanding occurs if participants bridged their game experiences to the instruction. The fourth and final video, "Voltmeters and Ammeters" covers novel but related concepts that were not explored in the game. This unit explores in detail how voltmeters and ammeters measure voltage and current, and why they need to be put in series or parallel. For participants who played the game prior to instruction, they should be able to access their experiences exploring circuits in series and parallel to make the connection to how and why voltmeters and ammeters should be configured in parallel and in series, respectively.

Learning Measures. Learning measures were assessed through three Open-Ended (OE) Worksheets and a Post-Test. Learning measures were coded on two dimensions: 1) participant's understanding of physics principles (referred to as "correctness") and 2) complexity of his/her conceptual model (only possible on free-response, open-ended items). Correctness and complexity are distinguishable because while the former can be gained by superficial regurgitation of the definition of various basic components of the system, complexity refers to "the number of related dimensions or sources of variation" within the system (Halford, Wilson, & Phillips, 1998). Correctness was operationalized as the number of basic features of the electric circuit system they include in their explanations (i.e. do they include basic components such as a source of voltage, a resistor, and a current in their conceptual model) that was explicitly covered in the video lecture provided. Complexity manifested itself in various ways: through more in-depth explanations of interrelations between components in the system (Halford et al., 1998), such as talking about how current is shared in circuits with resistors in series, therefore reducing the voltage that travels to each resistor; through mentioning multiple levels of components in their causal explanations (i.e. that electrons flowing in the micro-level is what causes current to flow, or that because current is shared between each resistor in circuits in series, the overall resistance of the system is larger) (Jacobson, 2001); and through more complex justifications of their answer, such as using Ohm's Law or real-world examples to illustrate the differences between circuits in series and circuits in parallel. Correctness and complexity scores were blind-coded by two researchers trained on the same coding manual.

The Open-Ended Worksheet (OEs) were included as a form of free recall, to capture how much of the essential systemic structure and features students were able to internalize. The OEs contained the following free-response prompt: "Draw and explain a parallel circuit. Be sure to label all relevant parts of the circuit system, explain what a parallel circuit is, and how it differs from a serial circuit." The open-ended nature of the OEs offers participants the opportunity to provide both as correct and as rich of an account of what they know about the system both visually and verbally without specific prompts – as such, the elements and the relations they describe in both their diagrams and in their verbal explanations will provide measures of correctness and complexity. The correctness score was calculated by the number of correct basic elements of both the diagram and the verbal explanation provided. The complexity score was calculated by the number of additional components or connections participants provided in their explanation, and the richness of their explanation (for example, describing not only how parallel and serial circuits differ in their resistance, but how that impacts the brightness of their overall circuit). The OEs were given as a pre-measure (OE_1), after first activity (OE_2), and as apost-measure (OE_3).

The Post-Test was included as a more traditional form of assessment, using standard items that evaluate student understanding such as mathematical computations and reasoning about the system. The post-test was comprised of four sections: 1) three multiple choice questions about Ohm's Law, 2) three questions on reasoning about a circuit diagram, 3) two analogous reasoning questions, and 4) two PFL questions on voltmeters and ammeters.

The first section, three multiple choice questions about Ohm's Law, asked participants to reason about and calculate voltage and resistance in a circuit. The second section, which involved three questions based on interpreting a circuit diagram, asked participants to compare and contrast brightness of and currents running through light bulbs when the bulbs are in series, compared to when they're in parallel. The third section included two analogous reasoning questions that asked participants to reason about water pipe systems (a commonly employed analogy to explain circuits), and relate components of the water pipe system to circuit systems. These measures are based off of traditional transfer measures, which posit that those who have gained a sufficient understanding of the learned system (in this case, the circuit system) should be able to transfer their understanding the deep structures of that system to reason and map onto another similarly structured system (in this case, the water pipe system) (Gick & Holyoak, 1980). The fourth section involved two PFL questions require participants to reason about why voltmeters and ammeters are configured in parallel or series, which are based on the premise that deeper conceptual understanding of circuit configuration would lead to being better prepared to predict or reason about more complicated concepts. Students did not play levels that related to voltmeters and ammeters, but watched a video on how they are used during their instructional period (see section on Instruction, for further elaboration). All four Post Test sections were coded for correctness.

				SCORES	
Measure	Type*	# Qs	Topic	Correctness	Complexity
OE	FR	1	Circuits in Parallel & Series	Correctness	Complexity
Post-Test	MC	3	Ohm’s Law	Content	
	FR	3	Diagram Reasoning	Content	
	FR	2	Analogous Reasoning	Transfer	Complexity
	FR	2	PFL	Transfer	Complexity

*FR = Free Response
MC = Multiple Choice

Table 1: Study 1 Learning Measures

The first two sections’ (Ohm’s Law & Diagram Reasoning) correctness scores were summed to create the “Content” correctness sub-score. The latter two sections’ (Analogous Reasoning & PFL) correctness scores were summed to create the “Transfer” correctness sub-score. The Analogous Reasoning and PFL questions were also coded for complexity, calculated by the breadth and depth to which answers explicitly made connections between the water pipe and electrical circuit system, and deeply discussed the relationship between why specific forms of measurement (voltage or current) need to be configured in series or parallel.

Behavioral measures. Behavioral measures to capture each participant’s response to failure and problem solving processes were assessed through log data from the game. The log data generated a list of participants’ actions over the duration of the gameplay, which include timestamps of each action (dragging and dropping wire components, submission of answers, button presses such as hints or menus, etc.) and system-triggered events (such as a level start, a circuit explosion, or a feedback panel opening or closing). Exploratory analysis of the log data identified action sequences in response to failure, duration of time spent on a particular action or level, and how participants navigate through the game.

Study 1 Results

Learning outcomes. On the open-ended question given across three time points, a repeated measures ANOVA on OE correctness scores, with time as a 3-level within-subjects factor and condition as a 3 level between subjects factor, showed that there was a significant interaction between condition and time (*F*(2,64)= 3.002, *p* = .025), where participants in the TnP condition demonstrated higher content and complexity scores at OE_2 compared to the other two groups. This was expected; given that participants in the TnP condition received their instructional videos first (between OE_1 and OE_2), it stands to reason that they should perform better during OE_2. However, a post-hoc analyses showed that there were no significant differences in OE correctness scores by OE_3 (*p*=.238), which suggests that all participants demonstrated an equal amount of conceptual understanding by the end of the study. However, while learning of the physics concepts occurred across all of the conditions, participants in the FR Condition produced more complex explanations of direct current circuits in parallel and series (See Figure 4).

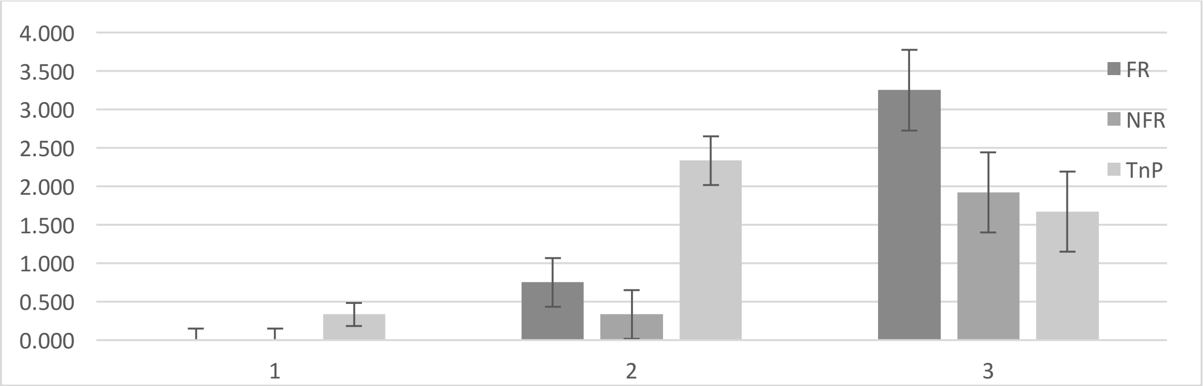


Figure 4: Study 1 OE Complexity Comparison

A RM ANOVA on OE complexity scores revealed that there was a significant interaction between condition and time, *F*(4,64)=6.213, *p*<.001, where the TnP condition produced more complex explanations at OE_2 as expected, but the FR condition provided marginally more robust explanations of electrical circuits at OE_3 than the other two groups, *F*(35) = 2.661, *p* = .085. This suggests that students who had the opportunity to respond to their failure prior to formal instruction can demonstrate a richer understanding of the system, and provides support for H_1. However, an ANCOVA on post-test scores controlling for OE_1 Correctness scores revealed that there were no significant condition differences in the Post-Test overall (.917), or in the content (*p*=.504), transfer(*p*=.612), or complexity (*p*=.560) sub-scores (Table 2).

		Correctness Score (Content + Transfer) (out of 21)		Content Sub-Score (out of 10)		Transfer Sub-Score (out of 11)		Complexity Score (out of 11)	
Game	n	M	SD	M	SD	M	SD	M	SD
FR	12	8.83	3.49	3.66	2.06	5.17	2.04	7.75	3.55
NFR	12	8.42	2.39	2.75	1.42	5.66	1.67	8.58	3.26
TnP	12	9.58	2.75	4.00	1.86	5.58	1.31	8.08	2.71

Table 2: Study 1 Post-Test Score Comparisons

This is contradictory to the earlier OE_3 complexity findings, as we would have expected that the increase in complexity of participants' understanding would also lead to higher scores on the transfer and complexity sub-scores. Furthermore, we would have expected that the FR (and perhaps the NFR) condition would perform better than the TnP condition on the transfer and complexity measures to affirm prior research on PFL.

Behavioral outcomes. When looking at frequency of failure between all three conditions, the NFR condition had an overall higher proportion of attempts that were successful (M_{NFR} = .585, SD_{NFR1} = .13) compared to the FR (M_{FR} = .38, SD_{FR} = .18) and TnP (M_{TnP} = .327, SD_{TnP} = .18) conditions. This result was expected, as the NFR condition directly received the correct answers, and likely replicated these answers in the next time they encountered that level. However, we had expected that the TnP condition would also have a higher rate of success, given that they already received the instruction prior to the gameplay. This suggests that for the TnP condition, the learning that occurred during instruction did not transfer to better conceptual understanding and performance in the applied context – the game.

Game log analysis captured 5 actions that participants in the FR and TnP conditions could take in response to failure: participants could fix their solution, restart the level, resubmit their answer without changing their solution (quick resubmit), move to another level (a previous level, the next level, or a more difficult level), or they could look for information by clicking on a hint or component description. From these actions, log data analyses identified seven distinct responses to failure: 1. Fixing current solution, 2. Quick Resubmit (did not change solution), 3. Restarted Level, 4. Skipped backwards, 5. Skipped to next level, 6. Info-seeking, restarting the level, and 7. Info-seeking, fixing current solution. Participants overwhelmingly favored using the fix and restart responses to failure.

	N	Mean	Std. Deviation
Fix	17	17.94	14.88
Quick Resubmit	13	5.08	7.10
Restart	29	15.07	13.57
Skip Back	22	1.91	1.31
Skip Next	25	4.72	4.03
Info, Restart	21	2.24	1.51
Info, Fix	13	2.77	2.95

Table 3: Study 1 Responses to Failure

We expected that some of the failure responses may be negatively related to failure, since every instance that these behaviors are produced are inherent markers of “lack of understanding” (otherwise they would have gotten it correct). Therefore, we’re particularly interested in when these behaviors are *positively* correlated to learning – that, despite happening in a moment of “lack of understanding”, the experience related to a positive impact on *later* understanding. Of the seven responses to failure identified, “Info-seeking, then fixing your answer” was the only response significantly correlated to Post-Test complexity, using Spearman’s Rho, r(13)=.687, *p* = .01, when adjusting for more conservative Type I error via a Benjamini & Hotchberg correction (1995). To identify whether the effect of info-seeking and fix behavior on learning was accounted for by prior knowledge, OE_1 Correctness score and Info-Seeking, Fix was regressed on Post-Test Complexity score. Results indicated that info-seeking and fixing was significantly predictive of learning complexity, even after controlling for OE1 Correctness, $\beta = .558$, $t(12) = 2.575$, $p = .028$. None of the other failure responses were positively related to any of the post-test or OE_3 measures of learning and complexity. The FR and TnP conditions did not differ by the frequency of failure responses, suggesting that even participants who received the instruction prior to gameplay responded to failure in the same way as those in the FR condition.

Discussion

These findings suggest that engaging with failure in game spaces before formal learning can elicit more nuanced mental representations of a complicated science system. While there were no condition differences in Post-Test and OE_3 conceptual understanding measures, students in the FR condition produced more complex and robust explanations of parallel and serial circuits in the last open-ended worksheet. This suggests that while the various methods of using games for learning across these three conditions can produce benefits to conceptual understanding, the affordance of experiencing failure can better prepare you to learn from formal content later on, thereby producing more nuanced and rich conceptual models. However, the expected finding that the FR condition would also perform better on the Post-Test analogous reasoning transfer and PFL measures was not confirmed, nor were the complexity of their responses on the post-test any better or worse than the other two conditions. One possible explanation for this is that the study did not provide enough of a treatment to have a significant effect on these other dimensions – after all, the game session only lasted 45 minutes and occurred only once. This is supported by Wouters et al.'s (2013) meta-analysis, which suggests that games are better for learning when there are multiple sessions of gameplay. Another possibility is that the measures of transfer used – the analogous reasoning and PFL questions – were not sensitive enough to detect significant differences between groups, or that they were not the right kind of transfer assessment to use for this kind of learning. A final possibility to consider is that these participants (university graduates) were able to learn sufficiently from the videos or already had prior experience and instruction with the content (all of them reported having taken at least high-school level physics), and did not require the use of a PFL activity to provide prior experiences.

These results suggest that even though failure can benefit the complexity of one's conceptual model (as supported by higher performance by FR participants on OE_3 complexity score), simply experiencing failure unto itself is insufficient for improving general conceptual understanding compared to just playing the game after learning (as suggested by the null differences between NFR and TnP in OE_3 scores). In other words, failing, by itself is not enough to prepare students to learn the material. Furthermore, the rate of failure was the same, regardless of whether you learned the material beforehand or not, which suggests that TnP participants' understanding from the lecture did not transfer into better performance in the game. However, the results did demonstrate that there are effortful behaviors in response to failure that are related to better learning. I found that of all the responses that one can take in response to failure in our game environment, the response of "info-seeking, then fixing one's answer" was significantly positively related to learning. When considered within the framework of metacognition, this is beneficial because it required participants to appraise and become aware of knowledge gaps, resolve identified gaps through info-seeking, and then apply the newly acquired information to adjust prior misconceptions. All three of these components - the appraisal (or the awareness of knowledge gaps), the resolution (or "filling-in"), and the application - are equally critical to learning. This is in contrast to Loibl and Rummel's (2014) conclusion that the awareness of knowledge gaps alone account for the benefits of productive failure for learning - we see that a general awareness is not as important here (as evidenced by our nonsignificant differences in Post-Test conceptual questions) so much as the specified appraisals of what one does not know *in the moment of failure*. Furthermore, we see the importance of the "application" component when contrasting "info-seeking, fixing" with "info-seeking, restarting" - if resolving knowledge gaps alone (through info-seeking) were sufficient for productive failure, then we should have seen that both of these strategies (info-seeking, restarting as well as info-seeking, fixing) would be statistically significantly related to learning. However, fixing one's solution after info-seeking as a key component suggests that the metacognitive actions taken after metacognitive judgments made are just as important as the judgments themselves – in particular, knowing that this newly acquired information must be used to address previous failures is a vital part of the learning process. The act of info-seeking and fixing may lead to deeper "appreciation for the deep structures" Schwartz et al. (2011) referred to in their own PFL activity - that is, info-seeking and fixing one's solution may have led to noticing and developing greater intuitions about the underlying concepts and structures of electrical circuits, that then led to more complex conceptual models. However, this exploratory study did not address the question of whether these enacted behaviors were indicative of a deliberate metacognitive strategy, or just so happened to be a pattern of behaviors that produced the most robust understanding. Although this behavioral response of info-seeking and fixing was most related to learning, and that info-seeking has classically been treated as a behavior of metacognition, it would be difficult to causally attribute metacognitive intent as driving these behaviors without a deeper investigation into what participants were thinking as they were enacting them. An alternative possibility is that other behaviors may also be significantly related to learning, but that our sample size for these correlations are too small and therefore not detectable. It is also possible that participants who already had an understanding of electrical circuits may also enact better metacognitive behaviors, because they have the cognitive resources readily at their disposal that otherwise may have been devoted to trying to understand the problem space in the first place (i.e. there's less cognitive load required to process the content for participants with prior knowledge, so they have more capacity to reflect on their strategy and performance). However, the "info-seeking, fix" behavior is not correlated to OE_1 scores (p=.813), which suggests that even those who demonstrated some prior knowledge did not employ this strategy any more than those who did not.

To conclude, using failure spaces in games for learning can be an effective way of improving the complexity of students' conceptual model, but have tenuous impact on more general conceptual learning and transfer. Furthermore, the most effective behavioral response to failure in our game was to fill in knowledge gaps through information seeking, then applying this new information towards fixing one's prior incorrect solution. However, we are limited in our conclusions about the degree to which games in general can be an effective PFL activity, because our population may have already possessed the appropriate prior intuitions about the system, or because the transfer measures were insufficient or inappropriate for this audience, or the treatment duration was insufficient. Furthermore, while the behavioral response to failure appeared to indicate deeper metacognitive strategies, this data does not provide the validation required to make the argument that these behaviors are driven by metacognitive responses and strategies employed in response to failure. To address these limitations, Study 2 examined whether this game can be used as an effective PFL activity for more novice participants with little or no exposure to the content, using alternative transfer measures such as delayed assessments, giving students a particularly challenging level to see what kinds of behaviors students spontaneously employ in response to complex, difficult problems, and using multiple game sessions for the treatment. Above all, Study 2 compared whether the addition of metacognitive prompts after failure that provoke students to appraise, info-seek, and apply information to fix solutions in the game can produce better learning outcomes in both conceptual learning and conceptual complexity.

Study 2 Design

The results of Study 1 suggest that there is an optimal way of responding to failure, but did not explicitly connect these behaviors to metacognitive intent. However, the most plausible explanation of why “info-seeking, then fixing your answer” was the behavior most related to more complex understanding later on is that the act of information-seeking suggests that the participant must have made a metacognitive judgment about what went wrong or what they did not know, and decided to info-seek to remedy that gap. Furthermore, the finding that “info-seeking and fixing your answer” is more valuable than “info-seeking and restarting the level” suggests that applying this newly acquired information to reconcile a prior mistake is a critical feature of why this behavior is significantly related to later understanding. Within this set of actions are several metacognitive monitoring and control behaviors: first, the participant evaluated his own performance and determined that he was lacking in understanding necessary to succeed (metacognitive monitoring); second, he decided to look for and attend to information to fill that knowledge gap (metacognitive control); and finally, he took this newly acquired knowledge and used it to attempt to address his incorrect answer (metacognitive control).

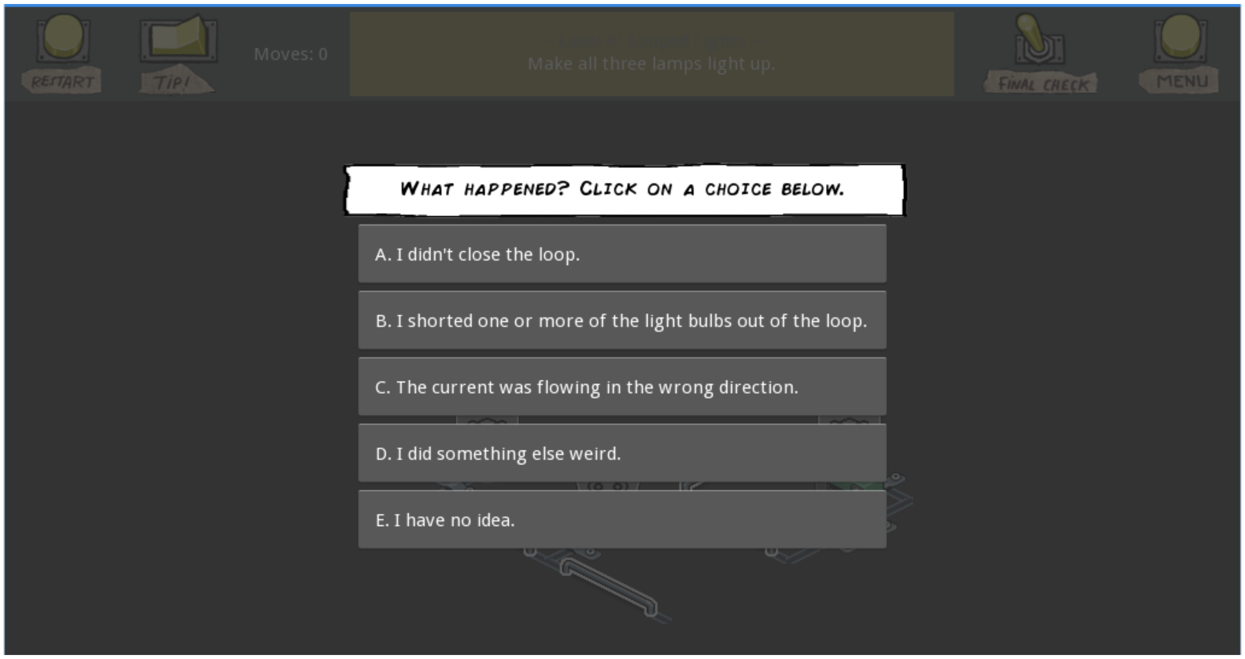
The participants in Study 1 were university-educated adults who presumably have a high repertoire of metacognitive strategies and behaviors at their disposal in a problem-solving environment. Yet, not all of these participants employed the most-effective strategy of “info-seeking, fix”, which suggests that the benefit of these behaviors aren’t obvious even to sophisticated learners. Given these findings, we now turn to the question of whether inducing these behaviors will then produce better learning. In other words, if we prompt participants to take these same steps of appraising the source of error, info-seeking, and fixing their solution, will this lead to deeper conceptual understanding later on?

Overview

In general, my theory is that employing effortful, metacognition-based responses to failure in games will produce more robust and complex understanding than playing a game without employing these responses to failure. However, these kinds of effortful behaviors don't always come naturally to adults, let alone high-school students. As such, I expect that the use of the metacognitive prompt, outlined below, that mirrors the most effective behavior found in Study 1 ("info seeking, fix") will help students develop a richer implicit understanding and intuitions from their game experience that will better ground their understanding of the formal instruction that follows. The goal of this dissertation study is to investigate whether the use of metacognitive prompts that walk students through effective responses to failure will lead to both the formation of a richer grounding experience for learning about electrical circuits, as well as generate more effective approaches to failure in general, even when the prompt is not given.

The metacognitive prompts.

Metacognitive Failure Response (MFR) prompts. In the "Metacognitive Failure Response" (MFR) condition, participants played a version of the game where after they get a certain number of incorrect attempts, they received the 3-part metacognitive prompt that explicitly walked them through the "info-seeking, fix" behavior. This prompt is designed to be embedded naturally into the game's narrative, using character dialogues and personalities to take participants through the response to failure. Panel 0 (not pictured here) features a brief exchange between the narrator (player) and a fellow lab engineer, Shelly, who provokes you to wonder "where did I go wrong here?" The first panel asks participants to make a judgment about what went wrong via a checklist of both possible causes, as well as the options "something else" and "I have no idea" (Figure 5).



Make all three lamps light up.

WHAT HAPPENED? CLICK ON A CHOICE BELOW.

A. I didn't close the loop.

B. I shorted one or more of the light bulbs out of the loop.

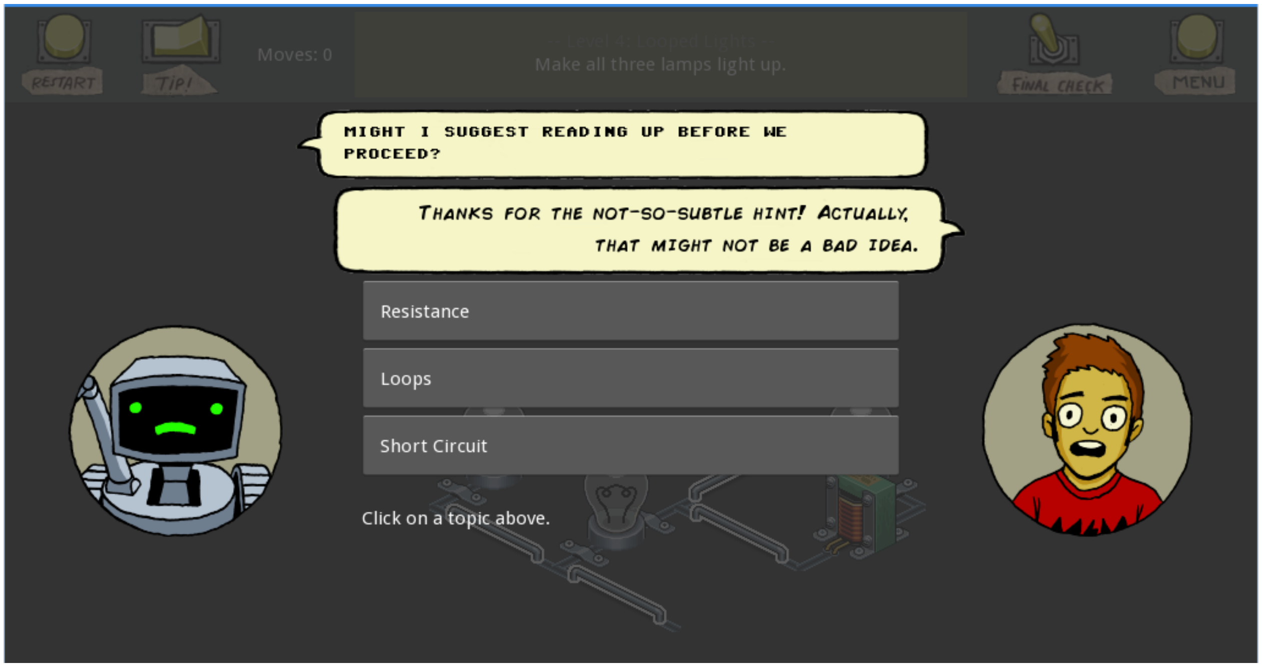
C. The current was flowing in the wrong direction.

D. I did something else weird.

E. I have no idea.

Figure 5: Panel 1 - Error Specification

The second panel shows participants a clickable list of concepts that are covered in the level, along with an expandable description of the concepts (Figure 6).



Make all three lamps light up.

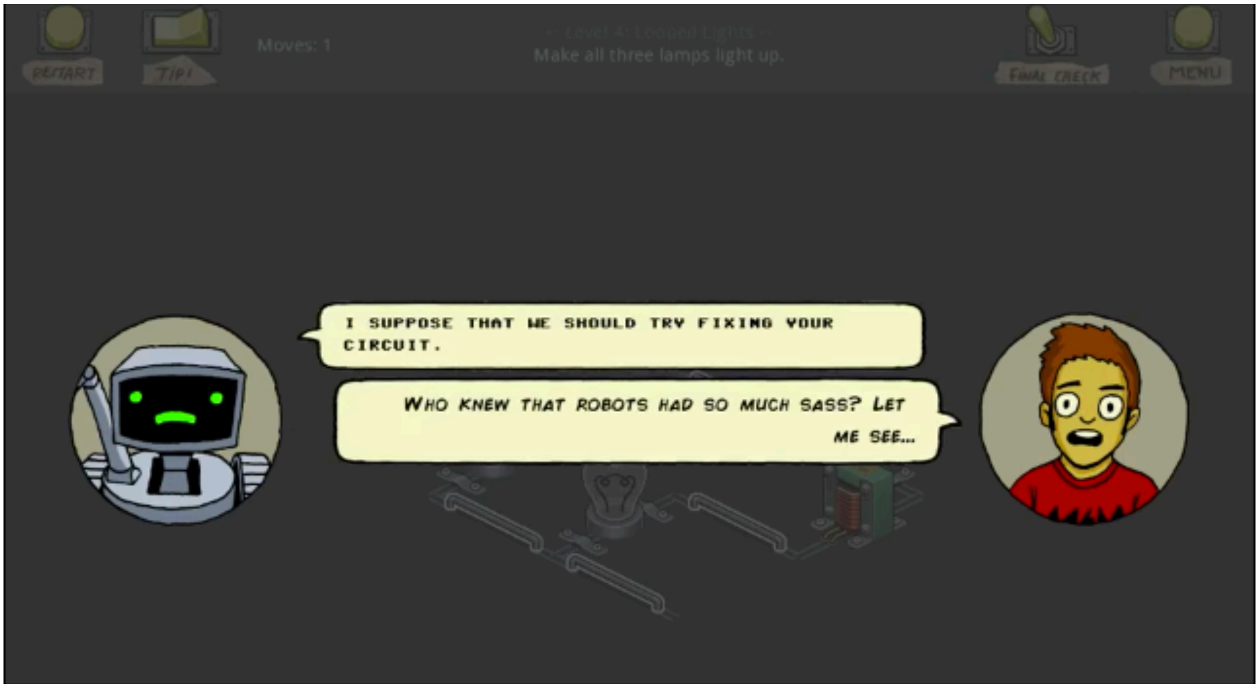
MIGHT I SUGGEST READING UP BEFORE WE PROCEED?

THANKS FOR THE NOT-SO-SUBTLE HINT! ACTUALLY, THAT MIGHT NOT BE A BAD IDEA.

Click on a topic above.

Figure 6: Panel 2 – Info-Seeking

The third and last panel features a conversation between the main character and a non-playable character (NPC), Throckmorton, who suggests that the player fix their incorrect solution (Figure 7). This third panel only triggers if the circuit has not been “blown” (or shorted) – if the circuit is blown, players must restart the level, as the battery, wire, and lights are no longer operational. The metacognitive prompt appeared only after the 3rd incorrect attempt (and every third attempt after) on non-tutorial or scaffolded problems. The selection of which attempt # the prompts should be triggered on was done by looking at the average number of attempts needed to solve the problem by Study 1 participants, and also by considering how many prompts to present in such a way that does not disrupt the flow of the game or is presented on when the player needs it most, without some intelligent algorithm running to predict participants’ knowledge states.



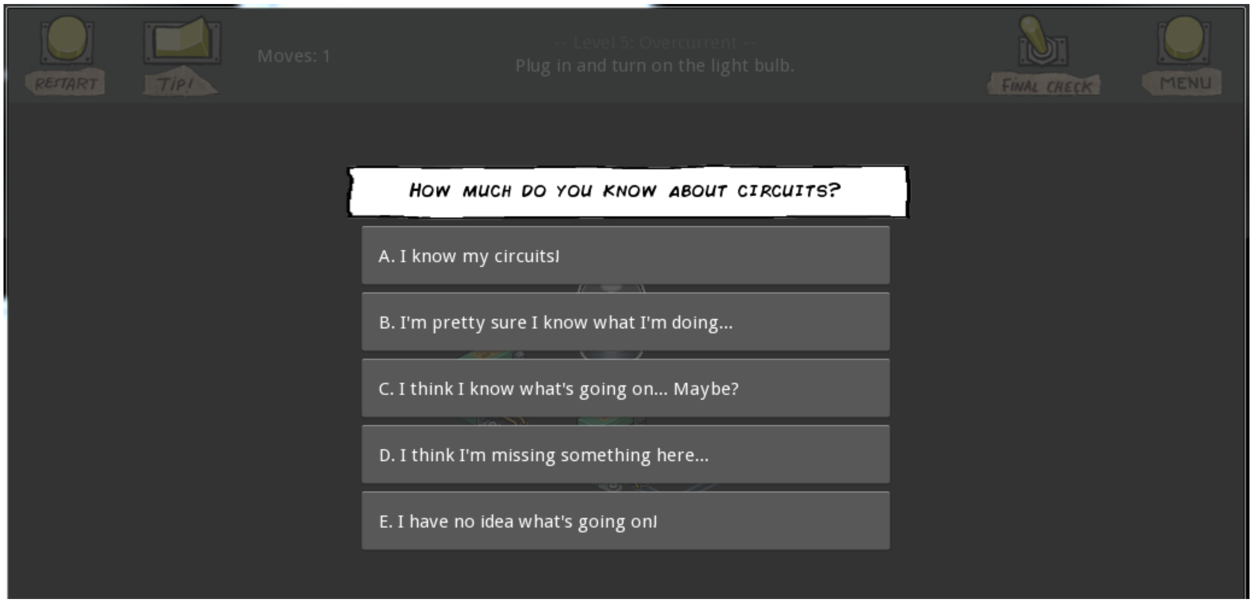
-- Level 4: Looped Lights --
Make all three lamps light up.

I SUPPOSE THAT WE SHOULD TRY FIXING YOUR CIRCUIT.

WHO KNEW THAT ROBOTS HAD SO MUCH SASS? LET ME SEE...

Figure 7: Panel 3 – Fix Suggestion

Global Awareness Prompts: In the “Global Awareness” (GA) condition, participants will play a version of the game where after they get a certain number of incorrect attempts, they will receive metacognitive prompt that asks them to make a global judgment of their understanding of electrical circuits. The prompt, designed after Loibl & Rummel's (2014) assertion that PFL experiences are beneficial predominantly for their elicitation of global knowledge gaps, will allow me to isolate the effects of the "info-seeking, fix" strategy versus just a general awareness of knowledge gaps without specification. This allows me to compare students who are provoked to make a general appraisal of their understanding (GA prompts) to those who are provoked to take specific metacognitive judgments and strategies (MFR prompts) in response to failure to identify which type of metacognitive judgment is more predictive of learning. Panel 0 (not pictured here) features a brief exchange between the narrator (player) and a fellow lab engineer, Shelly, who then provokes you to make a global judgment of knowing (Figure 8). The GA prompt triggers at the same attempt numbers, on the same problems as the MFR prompts.



-- Level 5: Overcurrent --
Plug in and turn on the light bulb.

HOW MUCH DO YOU KNOW ABOUT CIRCUITS?

A. I know my circuits!

B. I'm pretty sure I know what I'm doing...

C. I think I know what's going on... Maybe?

D. I think I'm missing something here...

E. I have no idea what's going on!

Figure 8: Global Awareness Prompt

Research Questions

The research questions driving this dissertation are as follows:

RQ$_1$: Does provoking students to reflect, info-seek, and fix their solutions through a metacognitive prompt in a game space ("MFR participants") prior to formal instruction lead to more complex and robust understanding later on, therefore outperforming students who do not receive the metacognitive prompt during gameplay ("control participants") or prompted only to make a global judgment of knowing ("GA participants")?

RQ$_2$: Do MFR participants use the prompts in different ways, where spending more time on error specification (Panel 1), info-seeking (Panel 2), or both results in better performance later on?

RQ$_3$: On a challenge level that does not contain metacognitive prompts, do MFR participants spontaneously use the "info-seeking, fix" strategy more than the control and GA participants?

RQ$_4$: Do MFR participants develop richer implicit understanding of the electrical circuit mechanics in the game because of the metacognitive prompts, and therefore perform better on the challenge level than the control and GA participants?

RQ$_5$: Do MFR participants behave differently than control and GA participants in their approach towards solving the game levels, such as using hints, needing more attempts, or responding differently to failure on the attempts that don't have metacognitive prompts?

RQ$_1$ poses the most pertinent investigation of this dissertation: whether inducing metacognitive judgments and strategies will lead to deeper learning from the instruction that follows. Both prior work on PFL and Study 1 suggest that all participants, regardless of the condition, learn some content from the formal instruction. However, given both the results of Study 1 and the literature on metacognitive monitoring and control, preparation for future learning, and productive failure, we should expect that students who employ metacognitive strategies in response to failure in a game space will develop deeper intuitions about the problem space, and therefore better prepare them to learn from formal instruction. Thus, I distinguish between two forms of knowledge that could be acquired: a rote, basic conceptual understanding of the concepts that is encoded simply by paying attention to the formal content covered in the lecture; and a deeper understanding that arises from integrating external experiences and prior intuitions (such as from the gameplay) into the formal content in the lecture that results in more robust and complex mental models. Furthermore, the comparison between the MFR and GA conditions allows me to determine whether specifying knowledge gaps and employing strategies to remedy and apply that gap to one's solution is more beneficial than global metacognitive awareness alone. To answer this question, I present my first hypotheses:

H_{1A}: The MFR, the control, and the GA conditions will perform equally on measures of rote conceptual understanding.

H_{1B}: The use of metacognitive prompts during gameplay (MFR condition) will lead to better performance on measures of complexity and retention, compared to the use of global awareness prompts during gameplay (GA condition) or the game alone (control condition).

RQ_2 acknowledges the possibility that participants in the MFR condition might treat the prompts differently, depending on whether they have effective metacognitive skills, or whether they find it useful for their gameplay process or not. For example, a student who has low prior metacognitive skills might not monitor their own understanding or strategies effectively, and so may choose to skip over the error specification (Panel 1) or not to info-seek (Panel 2). Or, a student who has high prior metacognitive skills might decide that they have accurately specified the source of error (Panel 1), and so do not need to info-seek. As a result, time spent on error specification, or info-seeking, may or may not positively impact later learning. I will conduct an exploratory analysis on the time spent on these actions to see whether they relate to more robust, complex understanding.

RQ_3 poses it is possible that the use of a prompt will facilitate more use of this "info-seeking, fixing" strategy in general, even when the prompt does not appear. Participants who receive the prompt may find it to be a useful strategy to employ in general, particularly on a level that was especially difficult. Thus, to answer RQ_2, I present my second hypothesis:

H_2: The MFR participants will learn to use the "info-seeking, fix" strategy from the metacognitive prompts, and therefore will employ the "info-seeking, fix" strategy more on a challenge level, compared to participants who do not receive the "MFR" metacognitive prompt during gameplay (control and GA condition).

RQ_4 asks whether using MFR metacognitive prompts in earlier levels will lead to higher rates or faster success on a particularly challenging game level. It should follow that reflecting on, pinpointing causes of incorrect solutions, and information-seeking would lead to deeper intuitions about the game mechanics (and therefore about electrical circuits in general). This should, in turn, translate into more success on a particularly challenging level. However, I have no prior evidence that those who employ these strategies will necessarily perform better in harder levels of the game itself - after all, those who employed those strategies were only using said strategies because they failed in the first place. Furthermore, productive failure research indicates that success of the problem-solving task is not so important so much as the experience of failing productively during the problem-solving task (Kapur, 2008). Thus, I will conduct an exploratory analysis without specifying a hypothesis.

RQ_5 asks whether participants who receive metacognitive prompts during gameplay (MFR and GA condition) behave differently compared to the control participants in their approach towards solving the game levels, such as using hints, needing more attempts, or responding differently to failure on the attempts that don't have metacognitive prompts? It is possible that receiving the MFR prompts might alter their gameplay approach in other, unanticipated ways. Receiving the prompts might make them more effective at passing levels, or it might make them more confused; it might also make them more inclined to search for hints or information, even when not prompted to do so because they notice the utility of such an approach, or it may make them less inclined to info-seek, because they find the process bothersome and unnecessary. Furthermore, the use of a "global awareness" prompt might also elicit different behaviors in those participants, such as info-seeking. I will similarly conduct an exploratory analysis without specifying a hypothesis about differences in gameplay patterns between conditions.

Study Design

Procedure. This study employed a fully randomized design, with randomized assignment of students within classes to one of three conditions: the control condition, the Metacognitive Failure Response (MFR) condition, and the Global Awareness (GA) condition. The control condition used the same game and study design as the FR condition in Study 1, where participants played the game with the normal allowances for incorrect attempts. In the MFR condition, participants played a version of the game where after they get a certain number of incorrect attempts, they will receive the 3-part metacognitive prompt. In the GA condition, participants played a version of the game where after they get a certain number of incorrect attempts, they received the global awareness metacognitive prompt.

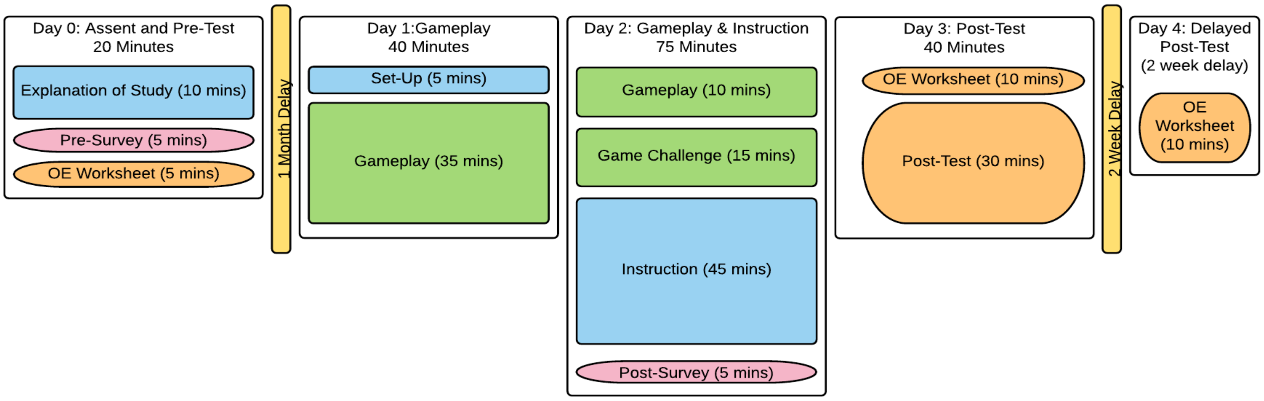


Figure 9: Study 2 Overview

Figure 9 shows the timeline of the study activities. Prior to the study activities, participants completed a pre-survey (Appendix B), and completed the first OE worksheet as a pre-test. A month later, participants began by playing a subset of the Electropocalypse levels (1-13) covering content like closed/open loops, short-circuiting, and resistors in series and parallel, for a full class period (40 minutes). On Day 2, after 10 minutes of continued normal gameplay, participants were given 10 minutes to try and solve a challenge level (further elaborated below). Note that gameplay has been split into two sessions. This is because meta-analyses on educational games indicate that multiple sessions of play are more effective for learning (Wouters et al., 2013), and the prior study suggested that one 45-minute session might not have been a sufficient treatment. After the challenge level, participants received instruction on circuits through a lecture video. They also took a brief post-survey, asking isomorphic questions about their confidence in explaining electrical circuit concepts, and their perceptions of games’ utility for learning, as well as the impact of their game experience on their learning (Appendix B). On Day 3, participants will take their second OE worksheet, followed by the Post-Test. On Day 4, two weeks after their post-test date, participants will take their third and last OE worksheet. Note also that the temporal placement of the OE Worksheets have been moved. This will be discussed further in the measures section.

Procedure

175 participants were recruited from their regularly scheduled physics classes in January 2017, with permission from their teachers and parents. During these classes, I explained the nature of the study activities, and offered them an alternative activity (a reading and writing assignment on a "Scientific American" article) for those who did not wish to participate as part of their class. After permissions were obtained, participants were randomly assigned within class to condition. After the explanation and obtainment of assent, participants were asked to fill out the pre-survey and the first OE worksheet that served as a pre-test. Immediately after this day, I analyzed both the survey questions and the OE responses to ensure that the randomly assigned groups were equivalent, as well as other potentially significant differences that may impact the treatment, such as gender, prior STEM experience, and gaming experience. A month later, I returned to the school during their class time over the course of three days to conduct the gameplay, lecture, and post-test sessions, assisted by the classroom teachers. Two weeks after those sessions, I returned a final time to distribute the last OE worksheet.

Participants. 175 participants (52% female) were recruited from a convenience sample at a high-performing, suburban New Jersey high school. No participants elected to opt out at the start of the study, but 9 participants were dropped due to absences over the course of the study, with a total of N = 166 participants with data. The school population is 69% white, 18.5% Asian, 9% Hispanic, and 3% black; and consistently performs at the 90th percentile or above on the New Jersey High School Proficiency Assessment (HSPA). Study activities were conducted as part of their conceptual and academic physics class (40-minute class periods); higher performing classes at the

Honors and AP level were excluded due to concerns over introducing an additional source of variance (student academic level), and the higher likelihood of prior knowledge. Academic physics is the standard-track physics course offered to students who previously took academic classes but did not do well enough to take or opted out of Honors or AP level courses, while conceptual physics is the remedial-track physics course offered to students who either did not do well in prior academic science classes or lacked the prerequisite math course. 8 course sections were taught by two teachers, who each taught two standard-track and two remedial-track physics classes. The participants were predominantly 16- and 17-years-old, as these courses are typically taken in the junior year of high school.

Materials

Game. The game used in in this study, Electropocalypse is the same as the one used in Study 1, with the prompt modifications by condition described earlier. Game software was installed at the computer lab at the school, where participants were directed to during gameplay sessions. If students completed all 13 levels before the game session was over, they were instructed to replay earlier levels; students often replayed levels on which they did not receive a 3-star rating (achieved by solving the circuit puzzle in the least number of moves), shown on the menu screen. If students asked for help, they were instructed to look to game hints or earlier levels to look for more information. Like Study 1, log data from the game was used to capture participants' problem-solving strategies and behaviors, including # and duration of attempts needed to solve each level, # and duration of hints used, and counts of response-to-failure (R2F) behaviors. Additionally, the log data from the metacognitive prompts (time spent on Panel 1: error specification and Panel 2:info-seeking) in the MFR condition, as well as from the general awareness prompts in the GA condition was analyzed.

Instruction. The instructional videos (three Khan Academy videos, plus an additional video that connects the game puzzles with the learned concepts) covered principles surrounding electrical circuits. The content from Khan Academy, a publicly available online learning platform, covers basic principles about electrical circuits, Ohm's Law, circuits with resistors in series and parallel, and the PFL content of Voltmeters and Ammeters. The gameplay "bridging" video, which takes several puzzle examples in the game to show how puzzle solutions use resistors in series and parallel to meet different goal states, was added to help participants make the connection between their game experiences and the instruction explicit.

Pre Survey. Participants completed an 18-item pre survey prior to the study activities. This pre-survey was given to collect demographic information, as well as other factors that may impact their game behaviors and learning performance. These 18 items ranged from asking about their prior STEM and game experiences, whether they identify as gamers or not, to their confidence in ability to explain electrical circuits to someone (self-efficacy). The pre-survey also included a battery of items on goal orientation, derived from Midgley et al.'s (1998) scale for assessing students' achievement goal orientations. This battery was included because prior literature suggests that students with varying goal orientations might have different self-regulatory (metacognitive), risk-taking, and strategic behavior (Pintrich, 2000), and may be used as a moderator in later analyses.

Open-Ended Worksheets (OE). The Open-Ended Worksheet (OEs) contain the same prompt as used in Study 1(Appendix 1). However, the OEs will be given at three different time points than previously: as pre-test (OE_1), as part of the post-test (OE_2), and as two-week delayed post-test (OE_3). The second and third OE worksheet times were moved because the original OE_2 from Study 1, immediately after gameplay, did not yield any particularly insightful information about participants' conceptual understanding besides that playing the game appeared to only slightly increase participants understanding of the formal circuit system immediately after play. Klahr & Chen (2011) consider temporal interval as a dimension of knowledge transfer; thus, we can also consider this two-week delayed OE worksheet (OE3) another measure of "near transfer" ("near" being somewhat arbitrary, since there is no formalized way of measuring transfer distance); or, at the very least, a measure of robustness of learning, in the form of retention. Correctness and complexity scores (see description in Chapter 2, and shortened coding manual in APPENDIX) were blind coded by 3 researchers, with an intra-class correlation coefficient of the final summed scores at .925 using an absolute agreement definition, with a 95% confidence interval from .866 to .960 ($F(35,70)= 14.893$, $p < .001$).

Post Test. The Post Test used for this study is the same described in Study 1 (see Table 1). Thus, we will still have the same two correctness sub-scores of content (the first two sections that cover the explicit content about DC circuits) and transfer (the latter two sections about analogous reasoning and PFL), and one overall complexity sub-score. Post-Tests were blind-coded by 3 researchers, with an intra-class correlation coefficient of the final summed scores at .935 using absolute agreement, with a 95% confidence interval from .888 to .965 ($F(32,96)= 16.572$, $p<.001$).

Post Survey. The Post Survey included one isomorphic question from the Pre Survey that asked about their confidence in being able to explain electrical circuit concepts to someone, and three questions about how their game experience impacted their learning.

Study 2 Results

I will begin by presenting the variables and predictors I found in the pre-survey and pre-test stage that will be included as interaction or covariate variables in later analyses. I will also present some overall descriptive statistics of learning and game behaviors, and discuss how they differ from the Study 1 population of graduate students. I will then present the results of each condition comparison test and relationship to learning outcomes that relates to my five research questions: a) learning measures (OE worksheets, Post-Test); b) metacognitive prompt use; c) Challenge Level response-to-failure strategy and d) performance; and e) game performance and behaviors.

The learning analyses were conducted with all 166 participants, while game data analysis was conducted with 165 participants, due to one students' game data lost due to technical issues. Due to the large number of tests that were run at every stage of analyses, Benjamini and Hotchberg's adjustments for false discovery rates (FDR) (1995) were applied to any analyses besides hypothesis tests 1 and 2. Tests that were significant with FDR adjustment are discussed as primary results, while tests that were significant at a = .05 but not after adjustment are discussed as promising exploratory results that offer implications for future investigation.

Pre-Survey and Pre-Test (OE1)

Initial one-way ANOVA analyses of the pre-survey demographic data revealed that three variables significantly influenced pre-test (OE_1) scores: gender, prior STEM experience, and identifying as a gamer. Boys and girls performed equally low on OE_1 Correctness (*p* = .118), but did significantly differ on OE_1 Complexity, *F*(1,164) = 8.755, *p* = .004, with boys (*M* = .206, *SD* = .452) outperforming girls (*M* = .046, *SD* = .211). Similarly, prior STEM experience (*N* = 19) did not significantly predict OE_1 Correctness (p = .103) but did significantly predict OE_1 complexity, *F*(1,164) = 4.877, *p* = .029, with those having prior STEM experience (*M* =.290, *SD*

= .625) outperforming those with no prior STEM experience (*M* = .100, *SD* = .300). Students who identified as a gamer significantly predicted both OE_1 Correctness, *F*(1,164) = 8.252, *p* = .005, and Complexity *F*(1,164) = 10.182, *p* = .002, with those identifying as gamers ($M_{Correctness}$ = .912, *SD* = 1.10; $M_{Complexity}$ = .245, *SD* = .496) outperforming those who did not ($M_{Correctness}$ = .475, *SD* = .818; $M_{Complexity}$ = .065, *SD* = .243). Thus, gender, gamer identification and prior STEM experience will be used as both covariates and as comparison variables in later analyses, in addition to pre-test (OE_1 total scores). There were no condition differences in gender (*p* = .299), gamer identification(*p* = .278), and prior STEM experience composition(*p* = .855). Confidence in being able to explain electrical circuits to someone, *r*(166) = .508, *p* < .001, and high reported interest in physics, *r*(166) = .239, *p* = .002, also positively correlated with pre-test scores. ANOVAs showed that there were no condition differences in OE_1 Correctness (*p* = .966) or OE_1 Complexity (*p* = .239).

A set of goal orientation items was also administered as part of the pre-survey, derived from a subset from Midgley et al.'s (1998) goal orientation scale. Three distinct goal orientations emerged from a confirmatory factor analysis: mastery approach, performance approach, and performance avoidance. These goal orientation factors will also be explored to determine whether they impact response-to-failure strategies or condition effects on student outcomes.

Overall Performance and Comparisons to Study 1 Participants. There were significant differences in game behaviors and learning between the two track types that participated in the study, standard-track (academic) and remedial-track (conceptual) physics. As expected, standard-track students outperformed remedial-track students on every learning measure (*p* < .001) besides the baseline measure (OE_1) and post-test content (Table 4). Standard-track students also solved more levels in the game (*M* = 11.38, *SD* = 1.11) than remedial-track students (*M* = 10.83, *SD* = 1.19), *F*(1,163) = 9.351, *p* = .003. Academic Track did not interact with condition on any of the analyses discussed below, and is discussed here only as context for the population. There were no differences between the 8 classes recruited on learning outcomes or game performance, when controlling for track type.

	Academic (M, SD)	Conceptual (M, SD)
OE_1 Correctness	0.73 (1.04)	0.50 (0.80)
OE_1 Complexity	0.16 (0.41)	0.08 (0.27)
OE_2 Correctness*	4.85 (1.48)	3.83 (1.37)
OE_2 Complexity*	2.54 (1.69)	1.48 (1.47)
OE_3 Correctness*	4.5 (1.35)	3.19 (1.21)
OE_3 Complexity*	2.03 (1.42)	0.95 (1.03)
Post Content	3.08 (1.35)	2.83 (1.24)
Post Transfer*	3.64 (1.83)	2.59 (1.67)
Post Complex*	0.88 (0.99)	0.32 (0.68)

*sig. at a < .001

Table 4: Study 2 Standard Track vs. Remedial Track Student Performance

Generally, Study 2 participants (*N* = 166) exhibited lower performance on learning measures and more varied game behaviors compared to the graduate students in Study 1 (*N* = 36). Although both populations exhibited similarly low baseline scores on the pre-test (OE_1), high school participants comparatively lower on the learning measure after the video lecture (OE_2 in Study 2, OE_3 in Study 1, and post-test scores), especially on measures of transfer and complexity (see below for descriptive statistics). This is likely due to a trifold effect of differences in learner ability (with graduate students being much more experienced and adept learners), prior knowledge, and the fact that the instruction was delivered through a long lecture video, which is likely something much more familiar to graduate students compared to high school students. Study 1 participants also had a higher proportion of successful attempts on the game, but made less number of attempts overall (i.e. was less iterative), and had a smaller variety of response-to-failure behaviors compared to Study 2 participants, suggesting that there are game dispositions that differ by population. This is likely due to differences in dispositions: while graduate students are more likely to approach a learning task, even one that's situated in a game, more seriously and cautiously, high school students may concentrate less on the learning goal or performing well, and therefore be more willing exhibit a greater range of playful behaviors. Yet, the original driving question still remains: will we see that the behaviors that were effective for learning for graduate students are also effective for more novice learners?

	Study 1 Mean (SD)	Study 2 Mean (SD)
OE_1 Correctness Score (out of 9)	0.64 (1.38)	0.62 (0.94)
OE_1 Complexity Score (out of 24)	0.11 (0.52)	0.12 (0.36)
OE Correctness Score (after instruction)	6.08 (2.79)	4.34 (1.51)
OE Complexity Score (after instruction)	2.29 (1.89)	2.01 (1.670)
Post Test Content Score (out of 10)	3.47 (1.83)	2.96 (1.30)
Post Test Transfer Score (out of 9)	5.47 (1.67)	3.11 (1.83)
Post Test Complexity Score (out of 10)	6.39 (2.40)	0.60 (0.90)
Prop Attempts Successful	0.43 (0.20)	0.158 (0.09)
Total # of Attempts	51.5 (20.61)	102.448 (38.68)

Table 5: Study 1 vs. Study 2

Learning Outcomes.

I expected that MFR participants will perform the same as control and GA participants on correctness and complexity scores at OE_1 to establish equivalent baseline knowledge. To support H_{1A}, which asserted that all groups would demonstrate the same amount of content learning from the instruction, I also expected that MFR participants will perform the same as control and GA participants on correctness scores at OE_2 and on post-test content scores, since both groups should demonstrate the same amount of content learning from the instruction. However, to support H_{1B}, I expected that MFR participants will outperform control and GA participants on measures of complexity on both OE_2 and post-test, as well as on the post-test transfer score. These findings should indicate that although both groups learned about electrical circuits, participants that receive the MFR prompt will have a more complex understanding of the concepts. Finally, I expected to find that the MFR participants will outperform control and GA participants on both correctness and complexity scores on OE_3. This should indicate that the MFR participants not only have a more complex understanding of the concepts, but that their understanding is also more robust.

An initial correlation of all the learning measures indicated that prior knowledge was not significantly related to learning. Of all the correlations of OE_1 Correctness and Complexity to the various learning measures (OE_2 and OE_3 Correctness and Complexity, and Post-Test sub-scores), only OE_1 Complexity was significantly correlated to Post-Transfer, $r(166) = .163$, $p = .036$, and Post-Complexity, $r(166) = .19$, $p = .014$; however, these correlations were no longer significant with the FDR adjustment. This suggests that students' prior knowledge of electrical circuits was only at best modestly related to learning outcomes within this context, likely because most students had little to no prior knowledge of the system. All of the other learning measures were all strongly correlated to one another.

To test these hypotheses using the OE scores, I calculated two repeated measures ANOVAs on correctness and complexity scores, with time as a 3-level within-subjects factor, condition as a 3-level between-subjects factor, and track, gamer identification, gender, and prior STEM experience as covariates. All demographic differences in learning will be discussed in detail in a later section, labelled "Other Group Comparisons", and are mentioned here only for reference. Students performed fairly poorly on the OE free recall measures, getting less than half of the possible 9 points on their correctness scores. The RMANCOVA on correctness scores revealed a significant main effect of time on learning, $F(2,158) = 146.26$, $p < .001$, and a significant interaction between gender and time ($p = .008$), with girls outperforming boys on OE_2 correctness, and a significant interaction between track and time ($p < .001$), with standard track students outperforming remedial track students (see Table 4 for mean comparisons), but no significant interaction between condition and time (p = .557). Gamer identification (p = .773) and STEM experience (p = .791) were not significantly predictive of OE correctness scores. This suggests that learning occurred equally across all groups, validating H_{1A}, and that learning was equally robust between conditions even after a delay, rejecting H_{1B}. (Figure 10)

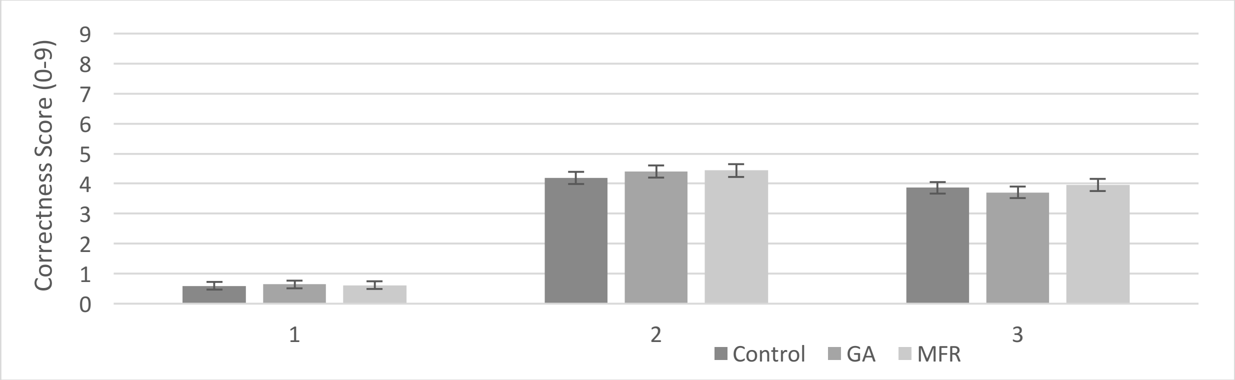


Figure 10: Study 2 Condition Comparisons on OE Correctness

Coding for complexity required criteria that would encapsulate all the possible ways students could demonstrate knowledge complexity, from using mathematical formulas to coding verbal explanations that featured more complex connections between system features. Despite this, Study 2 participants scored particularly low on measures of OE complexity, often getting no more than 2 points out of the 24 possible. The results of the RMANCOVA on complexity scores showed a similar significance of time on knowledge complexity, $F(2,163) = 149.90$, $p < .001$, controlling for the demographic covariates listed, but not between time and condition. The interaction between track and time was significant as expected (p < .001), but gamer identification (p = .437), STEM experience (p = .525), and gender (p = .091) was not.

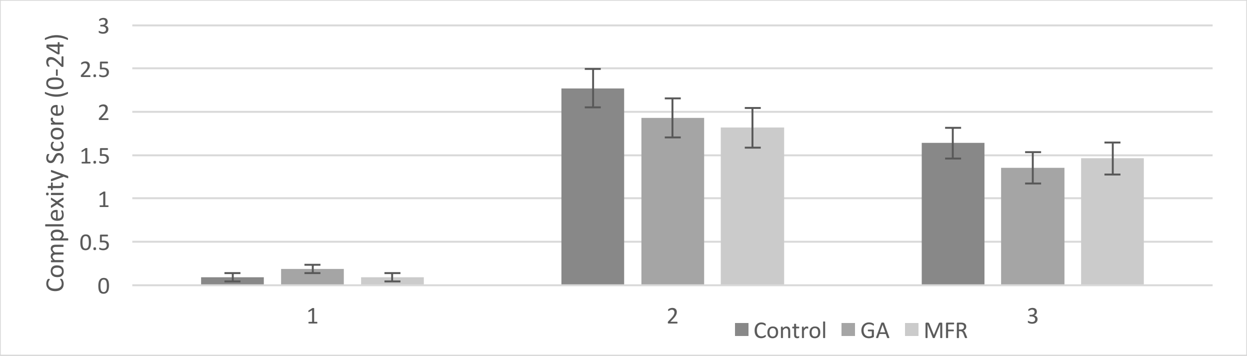


Figure 11: Study 2 Condition Comparisons on OE Complexity

To test the learning hypotheses using the Post-Test scores, I calculated a MANCOVA on Post-Test Content and Post-Test Transfer scores, with condition as a 3-level between-subjects factor, and OE_1, track, gamer identification, gender, and prior STEM experience as covariates. The results indicated that groups performed equally on the measures of rote learning (p = .098) and knowledge transfer (p = .401). Standard-track students outperformed remedial track students on Post-Test Transfer scores (p = .001), but did not differ on Post-Test Content scores. Gender trended towards predicting Post-Test Transfer (p = .055), and gamer identification trended towards predicting both Post-Test Transfer (p = .068) and Post-Test Content (p = .051). An ANCOVA on Post-Test Complexity, with condition as a 3-level between-subjects factor, and OE_1, track, gamer identification, gender, and prior STEM experience as covariates, revealed a similar nonsignificant relationship (p = .320). Track also significantly predicted differences on Post-Test Complexity scores, with standard-track students outperforming remedial track students (p < .001), but no other demographic variables were predictive of Post-Test Complexity performance.

Together, these learning outcomes analyses suggest that presenting the MFR prompts in the game did not yield the global learning benefits on conceptual complexity and retention predicted.

Metacognitive Prompt Use

Although gross comparisons between conditions did not yield differences in learning, there is still the possibility that using the metacognitive prompts more carefully might benefit learning. On average, MFR participants, received the prompt 5 times, used Panel 1 (error specification) an average of 33.40 seconds total, and Panel 2 (info-seeking) an average of 55.47 seconds total. To test whether there is a relationship between panel use and learning, I correlated the time MFR participants spent on Panel 1 (error specification) and Panel 2 (info-seeking) to the 7 learning measures. Correlations of time spent on the MFR prompts with Post-Test measures were not significant, but many with the open-ended (OE) measures were; this suggests that while metacognitive prompts might not have yielded better performance on a more standard form of assessment (i.e. through multiple choice questions, and prompted reasoning), using metacognitive strategies like error-specification and info-seeking does impact learners' abilities to freely recall system structures and features.

Spearman's Rho	OE_2 Correctness	OE_2 Complexity	OE_2 Total	OE_3 Correctness	OE_3 Complexity	OE_3 Total
TimeOnPa1	0.255	0.218	0.279*	0.433**	0.268	0.425**
TimeOnPa2	0.247	0.439**	0.396**	0.417**	0.371**	0.495**

* Sig. at a = .05
** Sig. at a = .01

Table 6: Study 2 Correlations of MFR Panel Use to Learning Measures

Time spent on Panel 1 was significantly correlated with OE_2 total score (Correctness and Complexity summed), and with OE_3 Correctness and total score (see Table 6). This suggests that more time spent on error specification is related to learning both immediately after instruction, and with long term retention. Time spent on Panel 2 had even more significant positive relationships to both OE_2 measures and to all OE_3 measures, suggesting that info-seeking after failure as a strategy is a powerful way to facilitate preparation for learning from the formal instruction, thereby yielding more complex and robust understanding. These were all statistically significant, even with FDR adjustments. This provides promising support for the original hypothesis that time spent information-seeking after error-specification as a response to failure is an effective way of preparing students for future learning. To identify whether eliciting a general awareness of knowledge gaps (GA prompts) similarly correlated to learning, the same analysis was conducted between time on GA panels and the learning measures. However, none of the correlations were significant or even trending, suggesting that provoking students to appraise their general understanding of the content did not yield learning benefits. However, it could be that simply failing in itself was enough to provoke this general awareness, and that the prompts did not elicit any deeper or more accurate appraisals of global knowledge gaps in GA participants than control participants.

Did students with different demographic backgrounds use the MFR prompts differently? ANOVAs on Time on Panel 1 and Time on Panel 2 suggest that students do not differ on Time on Panel 1 (error specification) or Time on Panel 2 by academic track, gamer ID, gender, or STEM experience, after adjusting for FDR. This suggests that MFR prompt use was not impacted by these demographic factors, such as STEM experience or academic ability, but might still be impacted by other individual differences not captured in the demographic factors.

Although students did not differ in their prompt use by demographic, the question remains of whether student prior ability, prior knowledge, or other demographic factors accounted for the benefits of MFR prompt use on learning. To test this relationship, I ran regression analyses on all 7 of the learning measures, with total time spent on the MFR prompts, OE_1, academic track, gamer identification, gender, and STEM prior experience as regressors. Results indicated that total time spent on the prompts predicted OE_2 Complexity, $\beta = .004$, $t(52) = 3.83$, $p< .001$, even after controlling for prior knowledge and experience and FDR. Total time spent on the MFR prompts also predicted OE_3 Correctness, $\beta = .006$, $t(52) = 2.08$, $p = .043$, and OE_3 Complexity, $\beta = .013$, $t(52) = 3.09$, $p = .015$; however, these were no longer significant after adjusting for FDR. Despite the vulnerability to Type 1 error, together these results suggest that there is a benefit to spending time on the prompts on knowledge complexity (OE_2) and promising evidence for positive effects on knowledge retention (OE_3), even when accounting for academic track, STEM experience, and prior knowledge differences.

Although these analyses provide evidence that the relationship between MFR prompt use and learning is not accounted for by demographic factors measured in the study (such as gender, prior knowledge, and academic ability), there is still a possibility that there are other, undocumented latent factors that impacted students' willingness to use the MFR prompts. I now aimed to explore how MFR high-users and low-users navigated the game environment differently, and whether they differed in behavioral and learning outcomes. This allows me to compare students who used the intervention with fidelity (MFR high-users) with those who did not (MFR low-users), and to the other control and GA participants.

MFR High vs. Low Users

In this section, I will conduct a set of exploratory analyses comparing MFR high-users and MFR low-users to determine whether they differed on other dimensions, such as game behaviors or motivation; as such, both statistically significant and trending results will be discussed. To create the high vs. low users category, MFR participants were split into high/low use categories using the mean time on Panel 2 of 42.38 seconds. Mean time on Panel 2 was used in lieu of Panel 1 or total time because Time on Panel 2 had stronger correlations overall to the OE learning measures, and because, while error specification is a vital part of this strategy, info-seeking (and willingness to devote effort doing so) may have a more direct impact on producing insights about the problem space that lead to greater learning outcomes. Using the Panel 2 mean split, 27 MFR participants were categorized as "low-users" and 25 were categorized as "high-users". This categorization will allow for exploratory comparisons between MFR high and low users, and with the other participants (control and GA), to identify whether MFR high-users demonstrated higher learning outcomes even when controlling for demographic factors.

To determine whether these MFR Panel 2 high-users and low-users differed demographically, I conducted a X^2 test on track by high/low panel use, $X^2(2, N = 52) = 1.322$ ($p = .250$), which indicated that high-users were not significantly more representative of standard-track students than remedial-track students. X^2 analyses indicated that MFR high- and low-users also did not significantly differ in STEM experience ($p = .704$) or gamer identification ($p = .309$), but girls trended towards being MFR high-users (58% of girls) more than boys (33%), $X^2(2, N = 52) = 3.067$, $p = .080$. ANOVAs indicated that MFR high- and low-users also did not differ in OE_1 correctness ($p = .526$), complexity ($p = .838$), mastery goals ($p = .370$), or performance-avoidance goals ($p = .091$), but MFR high-users trended towards being less performance-approach goal oriented ($M = 13.04$, $SD = 3.40$) compared to MFR low-users ($M = 15.10$, $SD = 2.22$), $F(1,50) = 6.914$, $p = .063$. This provides evidence that while high-users did not differ from low-users in academic track, prior knowledge, STEM, or game experience, they may be less driven to demonstrate their competency and therefore be more willing to read the information presented in the prompts instead of rushing back to solve the level successfully. ANOVAs conducted between on game behaviors revealed no other differences between MFR high and low-users, which suggests that MFR high and low-users did not differ in their engagement, game success, persistence, or problem-solving. If there were other latent factors that accounted for MFR high-users' higher performance on learning measures, those potential latent factors were not evidenced by differences in their demographic factors, game success or problem-solving behaviors. Motivation, in the form of performance approach goal orientation, appears to be the only measured difference between the two groups; however, goal orientation scores were not predictive of learning outcomes and thus could not explain the higher learning outcomes exhibited by MFR high-users.

MFR High/Low Use, Control, and GA comparisons

In this section, I compare MFR high-users and MFR low-users to control and GA to explore whether MFR high-users differed from the other groups on learning outcomes. To explore whether MFR high-users demonstrated higher outcomes, I re-calculated two repeated measures ANOVAs, with time as a 3-level within-subjects factor, condition (control, GA, MFR low-users, and MFR high-users) as a 4-level between-subjects factor, and academic track as a covariate, on correctness and complexity scores. The RMANOVA on OE correctness scores revealed a trending interaction between time and condition on learning, Roy's Largest Root *F*(3,160) = 2.062, *p* = .107. Post-hoc ANCOVA analysis on OE_3 Correctness, covarying track, revealed a significant difference between conditions, *F*(3, 160) = 2.687, p = .048, with MFR high-users (*M* = 4.54, *SD* = 1.406) outperforming MFR low-users (*M* = 3.49, *SD* = 1.52), control (*M* = 3.86, *SD* = 1.35), and GA participants (*M* = 3.71, *SD* = 1.44). Thus, we have evidence that MFR high-users better retained their understanding compared to the other three groups, even when controlling for prior academic ability.

A similar RMANOVA using OE complexity scores was also significant, Roy's Largest Root *F*(3,160) = 2.79, *p* = .042. Post-hoc ANCOVA analysis on OE_2 Complexity scores, controlling for track, were trending, *F*(3,161) = 2.585, *p* = .055, with MFR low-users (*M* = 1.31, *SD* = 1.19) performing lower than MFR high-users (*M* = 2.40, *SD* = 1.91), control (*M* = 2.27, *SD* = 1.71), and GA participants (*M* = 1.93, *SD* = 1.65). These condition differences in complexity were no longer significant by OE_3 (p = .291). This suggests that using the MFR prompts less led to less complex understanding immediately after instruction (OE_2 Complexity Scores) compared to the other three groups, even when controlling for track effects.

To compare performance on Post-Test scores between MFR high-users and the other participants, I calculated a MANCOVA on Post-Test Content and Post-Test Transfer, with condition as a 4-level between-subjects factor and controlling for Track and OE_1. The results indicated that groups performed equally on the measures of rote learning (Post-Test Content, *p* = .318) and Transfer, (*p* = .630). An ANCOVA on Post-Test Complexity scores revealed that all groups performed equally on Post-Test Complexity (*p* = .144).

Although these results provide promising evidence that using this MFR strategy more effortfully can lead to higher learning outcomes, there may be other latent factors that may explain the relationship between MFR panel use and learning, such as conscientiousness or engagement. Thus, while I cannot conclude that this metacognitive response to failure led to deeper learning, these results suggest that students who spent more time specifying the source of their error and info-seeking produced higher learning outcomes, even when controlling for prior knowledge, prior ability, or other demographic differences that traditionally impact science performance.

Challenge Behaviors and Performance.

The challenge level integrates all previously covered circuit principles into one particularly large and complex circuit system. This served as a space to see if the two conditions will differ in success in completing a particularly hard level, as well as whether they enact different responses to failure (RQ_3, RQ_4). I expected that MFR participants will develop an appreciation for the "info-seeking, fix” strategy through the metacognitive prompts, such that they will employ this response-to-failure more in the challenge level (H_2). However, ANOVAs on counts of response-to-failure behaviors by condition indicated that the three conditions did not differ in the types of strategies they employed in response to failure on this challenge level. In fact, all students generally ignored the “info-seeking, fix” strategy (only 11% used it at all), instead opting predominantly for fixing (*M* = 20.447, *SD* = 14.342) and restarting the level (*M* = 7.397, *SD* = 3.931) as their default response to failure (Table 7). This suggests that MFR participants did not transfer the “info-seeking, fix” strategy they experienced through the prompts to a challenging problem, thus rejecting H_2.

Challenge Level R2F	N	Mean	Std. Deviation
Fix	152	20.447	14.34
Info-Seeking, Fix	19	1.053	.23
Info-Seeking, Restart	9	1.000	.00
Info-Seeking, Resubmit	11	1.273	.47
Quick Resubmit	107	5.009	4.67
Restart	156	7.397	3.93
Skipped Back	7	1.000	.00

Table 6: Study 2 Challenge Level Responses to Failure

This hypothesis, admittedly, was overly ambitious; after all, if even more sophisticated learners (like the graduate students in Study 1) hesitated to use this strategy, it was less likely that high school students would have noticed the utility of such a strategy through merely two sessions of gameplay prompting them to do so. As earlier analyses made clear, many MFR participants opted to pay insufficient attention to the prompts in the first place; as such, it is clear that the prompts themselves, while effective for those who take them seriously, is not an effective mode of teaching this strategy. To invoke the transfer of this R2F strategy to either more challenging problems or to other contexts would require much more explicit and involved training of the skill itself.

All participants also performed equally poorly on the challenge level, with less than 17% of students succeeding on the challenge level. MFR participants had a slightly higher proportion of students who succeeded (22%) compared to control participants (14%) and GA participants (12%), but not significantly so, $X^2(2, \mathrm{N} = 166) = 2.373$ ($p = .305$). A logistic regression on challenge level success, with condition, track, and gamer identification as predictors, revealed a non-significant relationship between any of the variables and challenge level success ($p = .155$) Thus, we conclude that using the MFR prompts did not lead to more success on a challenging level within the game.

Gameplay Behaviors and Response-to-Failure

General Game Behaviors: The study also aimed to investigate what kinds of general game behaviors were related to learning, and whether comparison groups differed in their behaviors and performance. Analyses of game data identified several in-game constructs: 1) game performance, measured by proportion of attempts that were successful, number of successful attempts total, and number of levels solved (out of 13 levels total); 2) failure/iteration, measured by number of failed attempts and total number of attempts made; 3) general info-seeking, measured by total time spent on in-game information; 4) solution reflection, measured by time spent between attempt submission and next action; and 5) conceptual failure, measured by the number of “exploded” failed attempts. Explosions only happen if the solution “shorts” the circuit, or in the case of one level, triggers a switch that sets off a bomb (circumventing the trigger requires “shorting” around the switch). In other words, explosions only occur if you don’t understand the basic structure of a short circuit – that currents flow through the path of least resistance, but paths without resistance will result in the entire circuit getting fried. Explosions on earlier levels are expected, as they directly involve players experimenting with shorts to power light bulbs, as a way to teach the concept; however, explosions that occur later in the game, on levels that get increasingly more complicated and using other concepts (such as resistors in series vs. parallel), are a sign that the player never fully understood the basic structure and features of the circuit.

First, I correlated each of these game behaviors to the learning measures to identify what kinds of behaviors are positively or negatively related to learning, using Spearman's Rho with a Benjamini & Hotchberg correction (1995) to account for vulnerability to false discovery; as previously noted, statistically significant tests will be discussed first, followed by explorations of tests that were significant at a = .05 but no longer significant after adjusting for FDR.

		OE2	OE2	OE3	OE3	Post	Post	Post
	Spearman‘s Rho	Correctness	Complexity	Correctness	Complexity	Content	Transfer	Complex
Iteration	Total # of Failed Attempts	-0.037	0.031	-0.065	0.106	0.144	0.004	-0.077
	Total # of Attempts	-0.049	0.037	-0.061	0.108	0.168*	0.02	-0.044
Performance	Total # of Successful Attempts	0	0.139	0.148	0.087	0.288**	0.108	0.198**
	Prop Attempts Successful	0.037	0.009	0.116	-0.103	0.02	0.044	0.16*
	# Levels Solved	0.246**	0.306**	0.371**	0.359**	0.174**	0.173*	0.206**
Conceptual Failure	Explosions	-0.176*	-0.088	-0.17*	-0.045	0.129	-0.048	-0.085
General Info-Seeking	Time on Info	0.187*	0.133	0.146	-0.024	0.2*	0.021	0.147
Solution Reflection	Solution Reflection	-0.016	-0.071	-0.037	0.025	0.143	0.009	-0.119

*Sig. at a = .05
** Sig. wi B&H adj.

Table 7: Study 2 Correlations between Game Behaviors and Learning

In-game performance, in the form of number of levels solved, was the only significant relationship to all measures of learning and complexity except for Post-Test transfer, while total number of successful attempts was also significantly related to post-test outcomes. Number of successful attempts is a related but distinct construct from number of levels solved; students could and often did play a level more than once in pursuit of a "3 Star" rating on the menu screen, which required them to optimize their circuit solution using the fewest moves. As such, number of successful attempts could be a measure of "performance optimization" – the participant's desire to produce the best solution to a problem. Overall, these results suggest that doing well in the game is related to deeper learning.

Additionally, there were other constructs that were significantly related to learning at the a = .05 level, but were no longer significantly related after the conservative adjustment; still, these relationships are worth interpreting, as they lend some insight into what other possible game behaviors might lead to deeper learning. For example, one measure of iteration, total number of attempts made, was positively correlated with post-content scores, which aligns with prior literature on the benefits of trying many solutions as part of productive failure. General info-seeking in the game was also positively correlated with measures of rote learning immediately after instruction, which suggests that looking for information in the game helped students better understand some basic concepts better than those who did not. Conceptual failure, in the form of explosions, was negatively related to basic understanding both immediately after instruction (OE_2 Correctness) and retention (OE_3 Correctness). This suggests that this specific kind of failure, one that both marks a fundamental misunderstanding about a concept, and inhibits your ability to tackle more complex concepts in the game, can be debilitating for using failure spaces as PFL. This is a revalidating finding: a basic lack of comprehension or misconception that is not addressed (either by the learner or by an external agent) could be detrimental to learning (Chi, 2005), and specific types of game failures, in the form of persistent explosions in the game space, may be used to indicate such a fundamental gaps in understanding or misconceptions has not been effectively resolved. In short, perhaps not all failure is good for learning, and some failures can be a red flag for a fundamental knowledge gap that cannot be effectively resolved by the learner. Furthermore, this kind of failure, a kind of impasse, should have given rise to a global awareness of knowledge gaps, and therefore led to deeper learning (Loibl & Rummel, 2014; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). Yet, we see a negative relationship between this form of impasse and later performance. Another possibility is that the explosions made failure more salient in the game, which may have demotivated or frustrated players, which in turn led to less engagement in fruitful metacognitive reflection and strategies, such as info-seeking or error specification.

To more rigorously test the relationship between these game behaviors and the learning outcomes, I ran a linear regression analysis on the one learning measure most related to game behaviors, Post-Content, with Total # of Attempts, Total # of Success, # of Levels Solved, Time on Info, and the demographic variables (track, gamer identification, prior STEM experience, and gender) as regressors. The results indicated that iteration, in the form of total # of attempts, was the most significant predictor of learning, $\beta = .006$, $t(163) = 2.18$, $p = .03$, after controlling for in-game success, time on general info-seeking, and demographic variables. In fact, no other regressors were significant, after controlling for iteration. This suggests that iteration is the most powerful game behavior that predicts learning, even when controlling for demographic differences and in-game measures of success.

Given that there is a strong relationship between in-game performance, iteration, and learning, I now sought to identify whether groups differed in these game behaviors. ANOVAs on all game behaviors between conditions revealed that all groups enacted all these behaviors equally in their gameplay, with no significant differences. However, it is worth nothing that one measure, total number of attempts made, was just barely non-significant at p = .051, with control participants (*M* = 111.509, *SD* = 42.411) making more attempts overall compared to the GA participants (*M* = 101.482, *SD* = 33.676) and MFR participants (*M* = 93.558, *SD* = 37.970). This suggests that control participants were more iterative and had the opportunity to submit more solutions compared to the groups that received prompts. One possible explanation for this is that the prompts the GA and MFR participants received simply displaced the time that could have instead gone towards making more attempts. This makes sense, given that any time spent on prompts could have instead been spent problem-solving in the game space, and the differences in means of number of attempts between groups decrease by about 11 attempts per number of prompt parts presented (i.e. GA and Control participants differ by approximately 11 attempts, which may be accounted for by the appearance of the one-step prompt, while MFR and GA participants differ by approximately 12 attempts, which may be accounted for by the extra step MFR participants received). Given that iteration is a critical part of productive failure, both from prior research (Kapur, 2008; Schwartz et al., 2011) and from the present analyses, this tradeoff between iteration and metacognitive strategy use will be further discussed in the discussion section below.

Was MFR prompt use related to these game behaviors? To test these relationships, correlations between Time on Panel 1 and Time on Panel 2 for MFR participants were correlated to the game behaviors listed. Time on Panel 1 was positively correlated to Total # of Failed Attempts and Total # of Attempts, likely because the panels triggered on failed attempts. However, Time on Panel 1 and Time on Panel 2 also positively correlated (although this correlation was no longer significant after adjusting for FDR) with # of Levels Solved, which suggests that more time spent specifying errors and info-seeking was related to higher game completion. Time spent in Panel 1 (error specification) was similarly tenuously positively correlated with Solution Reflection, which suggests that students who were more willing to spend time specifying the source of their errors in the prompts were also more likely to exhibit similar behavior (appraising their own solution, presumably to identify where their solution broke down) even when the prompt was not present. (Table 8)

	Total # of Failed Attempts	Total # of Attempts	Total # of Success	Prop Attempts Successful	# Levels Solved	Explosi ons	Total Info Time	Solution Reflecti on
TimeOnPa1	.377**	.373**	0.12	-0.27	.320*	0.159	0.03	.317*
TimeOnPa2	0.272	.285*	0.227	-0.093	.331*	0.081	0.071	0.252

*. Sig. at a = .05.
** Sig with BH alpha

Table 8: Study 2 Correlations between Prompt Use & Game Behaviors

Response-to-Failure (R2F) Behaviors. There were 10 different responses to failure that students enacted: 1) fixing the solution; 2) info-seeking and fixing, 3) info-seeking and restarting; 4) info-seeking, quick resubmit; 5) info-seeking, then skipping back; 6) info-seeking, then skipping forward; 7) quick resubmit; 8) restarting the level (resetting the puzzle); 9) skipping backwards; and 10) skipping forward. Fixing and restarting were overwhelmingly the predominant responses to failure (Table 9).

	N	Mean	Std. Deviation
Fix	159	53.912	32.2460
Info-Seeking, Fix	115	2.122	1.1932
Info-Seeking, Restart	65	1.385	.6776
Restart	165	26.636	13.4654
Skipped Back	24	1.167	.3807
Skipped Forward	74	1.203	.5963
Info-Seeking, Resubmit	23	1.174	.3876
Info-Seeking, Skipped Forward	2	1.000	.0000
Info-Seeking, Skipped Back	1	1.000	
Quick Resubmit	136	6.838	6.2352

Table 9: Study 2 Responses to Failure

Next, Spearman correlations were run between these response-to-failure behaviors and learning outcomes. Of the 10 response-to-failure behaviors, “fixing the solution” was the only one significantly correlated with OE3 Complexity, *r*(159) = .160, *p* =.044, and Post-Content, *r*(159) = .179, *p* = .024, while restarting the level was negatively correlated to OE2 Correctness, *r*(165) = -.157, *p* = .045; however, these correlations were no longer significant after adjusting for FDR. This provides tenuous evidence that corroborates with our initial findings from Study 1: fixing may be a conduit for making failure more productive for future understanding because it provides an opportunity to reflect on one's solution and specify the source of error; restarting the level removes that opportunity. The fact that other R2F behaviors were not related to the learning measures could be because Study 2 participants simply did not enact these other response-to-failure behaviors enough for there to be a relationship, positive or otherwise. This is expected, given that even Study 1 participants, graduate students who presumably have a more sophisticated repertoire of metacognitive strategies, very rarely employed the response-to-failure behaviors most beneficial for learning.

There were, however, group differences in how often participants employed these R2F behaviors. Although this comparison was no longer significant after adjusting for FDR, control participants appeared to use the quick resubmit behavior a higher proportion (*M* = .089, *SD* = .053) than the GA (*M* = 058, *SD* = .050) and MFR participants (*M* = .062, *SD* = 046), F(2, 133) = 4.497 (*p* = .013). Quick resubmits may have been used by participants to revisit what happens when the current is turned on (the current only flows through when the player hits the "submit" button), or simply because they were confused. Control participants may have used more quick resubmits because they were more uninhibited by failure, while GA and MFR participants were warier of submitting incorrect answers, either because failure was made more salient by the prompts, or simply because they didn't want to trigger another prompt.

Other Group Comparisons

Gamer Identification. 52 (32.5%) of the 166 participants identified as gamers. X^2 analyses showed that boys were more likely to identify as gamers (54%) than girls (13%), X^2(1, N = 166) = 32.94 (*p* < .001), but did not differ in prior STEM experience (*p* = .127) or track (*p* = .914). To investigate the relationship between gamer identification and learning, I compared students who identified as gamers to those who did not on game behaviors, performance, and learning outcomes. ANOVAs revealed that gamers iterated more in the game, and used "fix" behaviors more often than non-gamers, and provided evidence (not significant with FDR adjustment) that they were also more successful and used more "quick resubmit" behaviors (Table 8). This suggests that gamers do exhibit a more playful, iterative disposition in the game space – they attempted more in general, pursued more successes (which does not necessarily translate into more levels solved – they could have played one level several times to get an ideal star count on the menu), attempted more "fix" behaviors rather than restarting or exiting the level, and, by extension of the earlier posited explanation for quick resubmits for control participants, seemed more resilient and undeterred by failure in the game. These playful behaviors – which highlight the resiliency, effort, and motivational benefits that both gamers employ and game spaces can encourage– were related to learning in earlier analyses. As such, it's unsurprising that gamers also performed higher on the Post-Test content and total scores. ANOVAs showed that gamers within the MFR condition did not significantly differ from non-gamers on time on error specification (*p* = .401) or info-seeking (*p* = .198) in the MFR prompts.

			Gamers	**Non**
Comparison	**Variable**	**P-value**	**M (SD)**	**M (SD)**
Learning	Post Content	p = .028	3.278 (1.535)	2.804 (1.153)
	Post Total	p = .038	7.352 (3.385)	6.344 (2.653)
R2F	Fix	p = .001*	66.412 (33.365)	48.009 (30.092)
	Quick Resubmit	p = .043	8.348 (6.711)	6.067 (5.867)
Game Behavior	Total # of Attempts	p = .002*	115.815 (42.610)	95.946 (35.020)
	Total # of Success	p = .023	15.704 (8.261)	13.541 (3.910)

*sig. with B&H adj.

Table 8: Study 2 Gamers vs. Non-Gamers

The question then, is whether the differences in game behaviors mediated gamers' higher performance. Iteration, in the form of number of attempts made, is a particularly noteworthy construct to investigate because it has classically been cited as a crucial part of PFL activities

(Kapur, 2008; Schwartz et al., 2011), and because games are cited as a particularly motivating context to encourage persistence and solution iteration in the face of failure; to find that gamers are more prone to iteration in a game environment would make a powerful case for encouraging game-like behaviors and activities in the classroom as a PFL activity. A regression analysis was used to investigate whether iteration mediated the effect of identifying as a gamer on post-test content scores. Results indicated that iteration, in the form of number of attempts made, was a significant predictor of post-test content performance, β = .007, $t(164) = 2.73$, $p < .01$. However, gamer identification, β = .347, $t(164) = 1.591$, $p = .113$, was no longer a significant predictor of post-test content scores, when controlling for the number of attempts made, β = .006, $t(164) = 2.277$, $p = .02$, suggesting that iteration fully mediated the relationship between identifying as a gamer and learning. These results show the promise of both using games in the classroom as a PFL intervention, because it provides the opportunity and motivation for iteration and metacognitive responses to failure, and for encouraging playful, game-like behaviors in the classroom that can help motivate students to be more resilient and effective in the face of failure.

Gender Comparisons. Given the extensive history of gender differences in both STEM fields and in video game play, I also sought to see whether boys and girls differed in their game behaviors and learning outcomes. ANOVAs on game behaviors and learning outcomes revealed several significant differences between gender that were no longer significant after adjusting for FDR; still, given the historical evidence and interest in gender differences in games and STEM, results will be discussed. Girls performed higher on immediate measures of rote learning (OE_2), but were less iterative in the game compared to boys (Table 10). Girls also employed less Fix and Quick Resubmit behaviors in response to failure. This is in alignment with our earlier findings on gamer identification and gameplay behaviors, given that more boys identified as gamers overall, and given that boys historically play more video games than girls do, which may have led to boys exhibiting more gamer-like behaviors in general. However, these more game-like behaviors did not lead to higher learning for boys; instead, girls appeared to performed better on the immediate measure of free recall. One possible explanation for this is that although boys demonstrated game behaviors that should have led to better preparation for future learning, girls may have attended more closely to the lecture that followed and therefore recalled more of the essential structures and features of electrical circuits; during the study, it was anecdotally observed that boys appeared to be more disruptive and off-task during the video lecture portion of the study compared to girls. However, this is merely speculative, as there were no quantitative or systematically codified measures of attentiveness to the video lecture. ANOVAs showed that boys and girls within the MFR condition did not significantly differ on time on error specification (*p* = .397) or info-seeking (*p* = .209) in the MFR prompts.

Comparison	**Variable**	**P-value**	**Girls (M,SD)**	**Boys (M, SD)**
Learning	OE2 Correctness	p = .01	4.63(1.46)	4.03(1.52)
R2F	Fix	p = .02	48.25(32.65)	60.09(30.84)
	Quick Resubmit	p = .02	5.58(5.03)	8.06(7.04)
Game Behavior	Total # of Attempts	p = .01	95.06(37.58)	110.49(38.50)

Table 10: Study 2 Gender Differences in Game Behaviors and Learning

Prior STEM Experience. Another comparison worth noting is whether students who reported having prior experiences in STEM, such as camps, clubs, or extracurricular classes, would behave differently or perform better in learning measures compared to those who did not. 19 (11.4%) of the 166 participants reported having prior STEM experiences. STEM experiences did not differ by gender, but X^2 analysis showed that standard-track students were more likely to report having prior STEM experience (18%) than remedial-track students (4.8%), X^2(1, N = 166) = 7.495 (*p* = .006). ANOVAs on game behaviors showed that students with prior STEM experience did not behave differently in their gameplay from those without prior STEM experience. ANCOVAs on learning measures by STEM experience showed that STEM experience was still significantly predictive of Post-Test Transfer scores, *F*(1, 163) = 5.259, *p* = .023, even after controlling for track. This suggests that while having prior STEM experience is related to academic ability, having those prior experiences within the field significantly predicted students' ability to transfer their understanding to novel content, even when accounting for their track level. ANOVAs showed that those with prior STEM experience within the MFR condition did not significantly differ from those without prior STEM experience on time on error specification (*p* = .189) or info-seeking (*p* = .730) in the MFR prompts.

Goal Orientation. Given that students with different goal orientations may have different responses to experiencing failure, I sought to investigate whether different goal orientations – mastery approach, performance approach, or performance avoidance - would correlate with different gameplay behaviors and outcomes. Correlations of the three goal orientation scores with game behaviors revealed that there were no significant relationships between goal orientations and gameplay. However, correlations with learning measures indicated that mastery approach orientation scores correlated with higher post-test transfer scores, *r*(166) = .216, *p* = .01, suggesting that students who emphasized mastery of learned content were able to better transfer their learning to novel contexts. Correlations between goal orientation scores and time spent on the MFR prompts showed that students who had a performance-approach goal orientation spent less time on info-seeking, *r*(52) = -.308, *p* = .027. This provides further evidence that students who are more motivated to perform well may have devoted less effort to carefully attending to the information presented, instead opting for returning to solving the game levels as quickly as possible.

Study 2 Discussion

In this section, I will review the results of Study 2, highlighting the important findings surrounding the use of a metacognitive strategy in response to failure, and its relationship to gameplay and learning. I will also discuss the limitations of the study, as well as implications for metacognitive strategy instruction, game-based learning, and industry.

Summary of Findings

Convention highlights the importance of failures for eventual success, and motivation research emphasizes theories that elucidate what aids in persistence in the face of failure (Duckworth et al., 2007; Dweck, 2006), implying that overcoming failure is an unavoidable, perhaps even useful pathway to better outcomes (Hong & Lin-Siegler, 2011). Prior literature on productive failure as preparation for future learning (Kapur, 2008) and metacognition as mediator of productive failure (Loibl & Rummel, 2014) showed that there is a benefit to the experience of failure for preparing students to better understand the instruction that follows, thus yielding deeper and richer understanding. However, this body of literature does not address what kinds of affordances are presented in the moment of failure that lead to deeper understanding.

Failure – and persistence through it – is an essential part of gameplay; in fact, games that fail to sufficiently challenge players are less intrinsically motivating, because they fail to induce the flow states that compel players to commit their best efforts to seek ambitious but achievable success (Csikszentmihalyi, 2000). As such, a possible insight to what kinds of opportunities for deeper understanding are elicited by failure can be found in the way gamers respond productively to failure. Skilled gamers appear to be quite effective at making use of their failures; they evaluate and select strategies to change outcomes, test parameters of the game space, and seek help from in-game resources or from game communities when they reach an impasse. Games encourage these effective responses to failure because they provide both the motivation to encourage persistence through failure, and a variety of affordances and scaffolds that empower players to employ a myriad of failure responses. Thus, an essential question this theoretical framework posits is: what is an optimal way to respond to failure in a game space, that may then lead to deeper learning from instruction that follows?

Study 1 helped answer this question by demonstrating the utility of one particular strategy, info-seeking and fixing, that sophisticated learners used when playing a game as preparation for future learning. Within this strategy, several metacognitive actions are taken: the player is appraising his/her incorrect attempt in an effort to pinpoint what went wrong (error specification); the player is looking for information to help fill that gap in understanding or to better understand the mechanics of that error (info-seeking); and the player is using this newly acquired information to then resolve their prior error. Thus, I characterized this suite of actions as a metacognitive failure response (MFR). However, not all the participants used this MFR strategy in their gameplay, suggesting that it perhaps is only needed when learners reach an impasse, that is not an intuitive strategy to employ, or that it is simply preferable (to the player's interest) to ignore such a strategy in favor of continuing gameplay. Regardless, these results lead to the next critical question: will inducing this metacognitive failure response lead to deeper learning for less sophisticated learners, such as high school physics students?

This dissertation sought to determine whether prompting students to use error specification, info-seeking, and fixing in response to failure (MFR participants) would lead to better preparation for future learning, compared to students who were not prompted to use this strategy (control participants) or those who were prompted to make a global metacognitive judgment instead (GA participants). Participants were randomly assigned to three game conditions: MFR (n = 53) and GA (n = 56) groups were given their respective prompts after every 3rd failure they encountered in the game, while the control participants (n = 57) were not given any prompts. All three conditions took pre-surveys, took the pre-assessment (OE_1), played the game and challenge level for over an hour, received the video lecture, and took several learning assessments, including an immediate open-ended worksheet (OE_2), a four-part post-test, and a two-week delayed open-ended worksheet (OE_3).

Although the results did not yield gross condition differences between the three conditions on learning, further analyses on MFR prompt use indicated more time spent on error specification and info-seeking led to higher performance on learning outcomes surrounding complexity and robustness of knowledge, even when controlling for prior knowledge, ability, and other demographic factors. This finding provides promising evidence that using metacognitive responses to failure, in the form of error specification, information-seeking, and fixing one's errors, is an effective way to make failure productive for future learning. Yet, prompt presentation in the game led to tradeoffs from other productive gameplay behaviors, such as iteration. Analyses indicated that iteration (in the form of more attempts) is a crucial mediator between game experience and learning, yet students who received any prompt, MFR or GA, attempted less tries, presumably because that time was spent looking at the prompts instead. Furthermore, prompt use also did not yield better performance or transfer of that strategy to a more difficult game level. Altogether, this suggests that while the prompts managed to provoke some students to use the error specification and info-seeking in response to failure, this approach was not the optimal way of teaching this strategy.

MFR Prompt Use and Learning

Condition Comparisons. Analyses of learning differences between the three conditions, in the form of repeated measures ANOVA on the OE free-recall worksheet, and multivariate analyses of covariance with the post-test measures, indicated that there were no group differences in conceptual complexity and robustness. This provided evidence that either this strategy is ineffective for less sophisticated learners, our Study 2 population, or that the prompts didn't consistently provoke MFR strategy use.

MFR Prompt Use. Further analyses of the MFR prompt use, in the form of total time spent on error specification (Panel 1) and info-seeking (Panel 2), showed that there was a great deal of variance in the way MFR participants used the prompts, particularly the info-seeking panel. Correlations between time spent on error specification, info-seeking, and learning measures indicated a positive relationship, suggesting that more time and effort dedicated to error-specification and info-seeking resulted in more robust and complex understanding. Regression analyses controlling for demographic covariates (such as prior knowledge, academic track, gender, or other STEM and game experiences) showed that time spent on the MFR prompt positively predicted knowledge complexity and retention. Furthermore, X^2 analyses indicated that students who used the info-seeking panel more than average ("MFR high users") were equally representative of all demographic groups, and ANOVA analyses indicated that high/low users also did not differ in their game behaviors and successes. Overall, these analyses suggest metacognitive response use was not explained by demographic factors, and that there is in fact a relationship between using such metacognitive responses to failure and deeper understanding.

Although demographic factors did not explain the relationship between MFR prompt use and learning, there may be a latent variable, conscientiousness, that may explain the differences between MFR high- and low-users, distinctive from prior ability or prior knowledge. Students who were more conscientious in using the MFR prompts seriously were able to demonstrate more robust knowledge retention compared to those who were less conscientious, or who simply did not have the opportunity to use the MFR prompts at all. In contrast, students who were less conscientious in using the MFR prompts exhibited poorer knowledge complexity compared to their more conscientious peers and to those who played the game without this strategy prompt. This is an interesting, if rather unsurprising, explanation; after all, using a metacognitive approach to failure requires one to be careful, reflective, and willing to vigilantly take advantage of resources provided. Less conscientious students may have neglected both to use the MFR prompts, and may have been generally less vigilant and reflective in their gameplay, resulting in lower knowledge complexity. Yet, if conscientiousness was the primary factor that accounted for the relationship between prompt use and learning, we should have also seen that GA prompt use also positively correlated with learning outcomes. Since GA prompt use was not related to learning, this suggests that while there may be individual differences that account for whether students chose to take the MFR prompts seriously or not, those individual differences alone did not wholly account for the benefits of the MFR prompt on learning. Still, the prompts, while effective for conscientious students who recognized the utility of a metacognitive response to failure, was not successful in provoking this strategy use for all MFR participants. As such, to seriously evoke such strategy use will require direct instruction of the strategy to increase adoption, even for those less-conscientious students. Nevertheless, these results provide evidence that using this MFR approach in response to failure can improve long-term learning outcomes.

The Global Awareness (GA) condition was included in this study because it provided the opportunity to contrast the benefits of a general, non-specific metacognitive awareness of one's own understanding with a more specific and strategy response to failure, error-specification and info-seeking. Our results suggest that eliciting a global awareness of one's knowledge gaps was not as effective for improving students' preparation for future learning as eliciting a specific metacognitive failure response. GA participants performed equally with control participants on all learning measures, which suggests that deliberate provocation of a global awareness of a knowledge gap was no better than simply playing the game without provocation. However, it's possible that the control participants also made these global judgments of knowing, even without the aid of a GA prompt. Thus, I have tenuous evidence that the benefits of failure for future learning are not in a generic, non-specific awareness per se, but in the specification of one's knowledge gaps and immediate strategies one can employ to investigate the source of an error and lack of understanding.

Game Metrics and Learning

General Game Behaviors and Learning. Other game behaviors and performance were also related to learning. In-game success, in the form of both number of levels completed and in performance optimization (number of successes overall), were strongly correlated with learning measures. While the relationship between game performance and learning outcomes should not come as a surprise to games researchers and developers (after all, educational games are typically developed to directly deliver practice or instruction on the learned content), this is in contrast with other kinds of PFL activities, such as productive failure (Kapur, 2008) or inventing with contrasting cases (Schwartz et al., 2011), where in-task success did not translate into deeper learning. However, productive failure and inventing tasks were purposely designed to have a low likelihood of success in mind, in order to encourage student production of intuitions and experiences that would then better prepare them for future learning. In contrast, games are designed to have the “just-right” balance of difficulty and success, allowing for players to struggle and be challenged enough for the game to be compelling, but not so difficult that success is near impossible – after all, few people would play a game where success (or at least progress) is not an option. Furthermore, games provide a variety of affordances for players to take advantage of in response to their failure, such as restarting a level, referencing earlier levels, looking for hints or information, and tinkering. This allows players to engage with failure in reflective ways that hopefully point them towards eventual success or understanding. In this regard, this actually makes a *stronger* case for using games as PFL – it provides the space and encouragement for students to persist and react productively to failure, without feeling the negative emotional effects from never experiencing success.

Iteration and general info-seeking in the game was also related to learning, corroborating with earlier research that showed the importance of persistent effort and iteration in PFL activities (Kapur, 2008; Schwartz et al., 2011), and in the utility of info-seeking when help is needed to benefit understanding (Aleven et al., 2006). In fact, regression analyses of the battery of game behaviors on learning outcomes revealed that iteration was the most significant game behavior related to learning, holding other game behaviors (such as game performance and information-seeking) constant. Iteration and failure are inextricably tied, as iteration enables failure experiences, and failure experiences provide opportunities for improvement on subsequent iterations. As such, the finding that iteration in the game was related to later learning highlights this relationship, especially when related to game performance; while iteration itself historically is an important part of PFL activities, iteration that then led to game success in this particular task was related to deeper learning later on, suggesting that those who both iterated and were able to glean something from their failed iterations (i.e. got enough information out of that attempt to adjust their answer or approach, eventually leading to solving the level) benefited the most from that experience.

Not all failure was good for learning, however; one particular type of failure documented in the game system, explosions, was negatively related to learning. This failure was a sign that the player was unable to master (or at least grapple meaningfully with) more complex topics in later levels because they lacked a fundamental understanding of the basic structure of a circuit. This type of recurring failure is unlikely to be productive because, despite being a highly salient signal of a knowledge gap (the entire circuit goes up in flames and the player is forced to restart with unburnt materials), players who frequently encountered this kind of failure failed to address this lack of understanding, which then undermined their ability to fully benefit from their game experience as preparation for future learning. This failure was not effectively addressed even with the MFR prompts, which suggests that students who experienced this kind of failure could not even specify the source of their error, much less actively seek out information to resolve it. This highlights the delicate balance between student ability, meaningful failure and struggle, and instructional intervention; while there are all kinds of failures that might be productive for students to engage in (particularly those that students can effectively address, especially through metacognitive approaches), the kind of helpless, fundamental failure signaled by explosions in the game demonstrated that certain kinds of impasses cannot be addressed by students' metacognitive strategies alone. Instructional intervention, either by a teacher capable of detecting such fundamental gaps in understanding, or by a technological system attuned to detect such impasses, should guide these students through more carefully scaffolded activities. After this fundamental knowledge gap is successfully addressed, students are then better prepared to tackle more complex ideas and problems productively, even if they do experience failure.

Gamer Identification, Iteration, and Learning. A critical question surrounding the utility of games for learning is whether gamer dispositions and familiarity might yield different behaviors and learning outcomes when using games in the classroom. Since games are quite unlike most standard academic tasks in many ways – goal expectations, stakes of outcomes, engagement, task features, and many more – it would be expected that the characteristics required of the player to do well in the game would differ from the characteristics required of the student to do well on academic tasks such as a test. Yet, given that the PFL intervention is grounded in the rich interactions students encounter in the game space, it's possible that students who have more familiarity with game environments and exhibit more playful behaviors may produce deeper intuitions for the game space that translate into better preparation for future learning.

The relationship between the utility of iteration and learning is most effectively embodied in the analyses involving gamer identification. Analyses revealed that students who identified as gamers were more iterative in the game, sought more successful attempts (performance optimization), and learned more from the material that follows, compared to students who did not identify as a gamer. Further mediational analyses showed that iteration mediated the relationship between gamer identification and learning outcomes, suggesting that the benefits of being a gamer when using games as PFL came from their iterative behaviors. Performance optimization, in the form of pursuing the most elegant or more than one successful solution to a level, is also something frequently encouraged by games, and the process of developing multiple or optimal solutions may also provide the opportunity to produce intuitions about the game's content or mechanics. Thus, we can conclude that students who have prior game experience enacted more playful behaviors, such as attempting frequently and optimizing one's answer, that then translated into deeper learning. This provides further support for using games in the classroom; by introducing more games into the classroom, more students may adopt more playful dispositions and attitudes in response to failure, that in turn translates into better learning outcomes.

Condition Differences in General Game Behaviors. Although using the metacognitive prompts effortfully was related to deeper learning, the presence of prompts appeared to be a tradeoff on time that could have been spent making more attempts, a measure of iteration. Each additional prompt section presented to players cost them an average of 12 attempts they could have made during that time instead. This is a significant concern, as iteration, as previously discussed, is not only a vital part of game play, but also of PFL activities. Furthermore, the introduction of prompts after the occasional incorrect attempt may interrupt players' flow and desire to continue problem-solving, which detracts from the allure of playing games as part of a classroom activity in the first place. A term coined in the educational games sphere, "chocolate-covered broccoli", adequately captures the feeling students express when they're told they're playing a game, but actually experience an academic task loosely couched in a gamified elements; the introduction of a prompt may add suspicions of "broccoli" hidden in the folds of the game task, which would undermine the motivational benefits of games in the classroom. Malkiewich and Chase (in press) found that in an engineering intervention which asked participants to tinker with Lego structures to learn about center of mass, participants were much more interested in the playful tinkering component of the task, and neglected to attend to the information regarding the concepts that would have helped them successfully build the structure. It's possible that the MFR low-users took a similar approach to their gameplay, where they dismissed the prompts in an effort to return to the much more interesting gameplay. Furthermore, participants who identified as gamers both iterated more in the game and performed higher on learning measures; mediational analyses confirmed that iteration mediated the effect of being a gamer on learning. Thus, we could posit that the benefits of games are that they promote iteration in the face of challenges, and that players who iterate more experience higher benefits from learning that follows than those who iterate less. Yet, responding to failure effectively is also an important conduit to deeper learning, but in-game prompts to employ this strategy resulted in a tradeoff with iteration. The question that remains, then, is how we can induce students to use this response-to-failure strategy, without a cost to gamers' natural predisposition for iteration.

Challenge Level Behaviors and Performance. There were no significant condition differences in the way students approached failure in the challenge level, nor in their success rate in solving the challenge level. Participants predominantly used the fix and restart behaviors in response to failure on the challenge level, rather than opting to use more varied strategies, such as info-seeking and fix. This is likely due to two factors: first, participants did not often enact other strategies besides fixing, restarting, and quick-resubmits to begin with, because other strategies (like info-seeking and fixing) are not as automatic and require more involved judgments of one's performance; second, because this activity was posed as a challenge with a time limit (so that enough time in the class period would be left for the video lecture), participants may have viewed it as a competition amongst peers on who could solve the challenge the fastest, and therefore was less likely to use more methodical, careful strategies in favor of returning to problem-solving. This is corroborated by our findings that performance-approach scores were negatively correlated with Panel 2 use; students who were more performance-approach goal oriented used the info-seeking panel less, perhaps because they would rather continue solving the level rather than trying to understand the mechanisms underlying the problem space.

Limitations and Future Directions

There were several limitations to this study design and outcome. Firstly and most importantly, this study only looked at whether *provoking* students to use error-specification and info-seeking before fixing as a response-to-failure strategy would benefit their PFL learning outcomes, and did not actually teach students to use this strategy on their own. This led to variance in prompt use, which was contingent on whether participants chose to take the prompts seriously or not. Although we found that higher prompt used led to higher learning outcomes even when controlling for demographic variables, students with higher academic ability did opt to use the prompt more, suggesting that prompts were less effective in provoking lower-ability students to use the strategy seriously. Furthermore, a possible latent variable, conscientiousness, may have accounted for whether students took the prompts seriously, even when accounting for prior ability and knowledge. Future studies should look at how students can be directly taught or motivated to use this strategy when they encounter failure, compared to students who were not taught to use the strategy, to mitigate this confound between individual differences and metacognitive response to failure. Instruction of the strategy may also then lead to transfer to other contexts, perhaps to other game levels (such as our challenge level), other games, other physics learning tasks such as labs and engineering projects, or even to other academic subjects, such as math problems or essay writing.

Another limitation of this study is in the design and delivery of the prompts themselves: perhaps prompts could be effective in provoking strategy use, but only when it is needed (i.e. when students have reached a point in their problem-solving that actually requires error specification and info-seeking, as opposed to using a set amount of failures). This study looked at failure as a general phenomenon, rather than a set of possible indicators of different states of the learner; yet, it is clear now that failure- and it's varied instantiations – can indicate a broad set of player states and understanding. For example, early failures could be a result of deliberate exploration and testing the parameters of the system that could lead to greater insight (and therefore success on later problems), or it could be the start of a spiral towards wheel-spinning, confusion and frustration. Future studies should look at how prompts could be delivered in conjunction with data mining techniques that detect the affective and cognitive state of the student, such as detectors of impasses or engagement, to encourage metacognitive strategy use when it is most helpful (i.e. when students are stuck or clearly do not understand the content. Furthermore, prompts could be delivered in a way that ensures students are paying attention to them (in the case of those who don't spend enough time reading that information) or are not spending too much time on info-seeking in lieu of problem solving (in the case of those who spend more than the optimal amount of time on the info-seeking panel).

Another limitation of this study is that it only identified one metacognitive approach to failure – there may be other strategies that may be equally or more effective, depending on the level of understanding and problem-solving ability of the player. For example, there may be a more appropriate strategy for when students are at an impasse that is pointing to a fundamental lack of understanding, such as the explosions in this study. Future studies should seek to identify what other kinds of response-to-failure strategies are enacted and when, to establish a set of behaviors that can make failure more productive. Learning algorithms can also detect when students are at an impasse that prevents their experiences from being productive, such as the explosions in the current game. Systems that could identify when students are lacking a fundamental understanding could provide more guidance or deliberate strategies to ensure that students have the opportunity produce better intuitions about the learned content, such that they're better prepared to grapple with tougher concepts that come in later levels or problems.

Finally, this study only looked at the use of this strategy within a particular physics game; as such, the implications for broader types of problem-based learning contexts (like other games, PFL activities, and generic educational activities like labs and projects) as well as for other academic subjects (such as history, math, and literature) are limited. The implications of the benefits of this response-to-failure strategy are also limited to learning outcomes; it is possible that the strategy, while effective for improving understanding and knowledge complexity, may not be as effective when applied to designing and improving a product or solution. This study also only sampled students from a high-performing high school, and may not be as relevant for other populations. Future studies should look at whether direct instruction of this strategy for different populations (i.e. lower-performing high school students, middle school students, or even undergraduates) impact their ability to learn from problem-solving environments.

Implications

The finding that a particular strategy in response to failure can lead to deeper understanding is an important one, with implications for both education and industry. While the results of this study indicate that this MFR strategy of error-specification, info-seeking, and fixing was strongly related to learning outcomes, the delivery of this strategy through an in-game prompt was insufficient for encouraging effective strategy use across all students. This provides strong evidence that in order for students to use this, and other reflective, metacognitive-based strategies in learning and problem-solving, these skills must be directly taught – a system-delivered mechanic is simply not enough to encourage strategy use. This finding is in alignment with the recent push by education foundations, think-tanks, and blogs, who increasingly call for direct instruction of the so-called “21st century” or “non-cognitive” (a completely erroneous but nonetheless well-adopted term) skills, such as critical thinking, communication, learning and study skills, and self-regulation. The fact that these somewhat ambiguously-defined skills can impact the way students learn from and succeed in both traditional and non-traditional learning tasks (like an educational game) call attention to the need for teachers to provide “21st century” skill development in the classroom, but teachers are often ill-equipped and (rightly) confused as to how to actually teach this wide and ill-defined spectrum of skills, dispositions, and strategies. This study has demonstrated that one particular skill in this category – the ability to respond effectively and strategically to failure – can facilitate deeper learning; as such, it provides direct evidence and concrete instructional implications for how this can be taught in the classroom, discussed below. Furthermore, while some might argue that the applicability and relevance of pre-calculus and chemistry for the typical citizen is rather low, the real-world relevancy of such skills as critical thinking, reflection and self-appraisal, and effective problem-solving cannot be denied. This study adds to the many other voices in the field calling out for a shift in the emphases and requirements of modern education – to emphasize not only domain-specific learning outcomes such as math and science, but to highlight the need and utility of skills that can be more broadly applicable to work, life, and citizenship.

For the PFL body of literature, the study suggests that there’s a way that teachers or the task itself can guide students’ problem solving in the PFL exploration phase in a way that optimizes their interaction with the novel content. Helping students be reflective about their errors, seek out new information, and resolve mistakes could produce even more powerful intuitions about the learned system that leads to more complex and robust understanding. Future PFL interventions developed for classroom use should encourage students not only to iterate on their solutions, but also take a metacognitive approach to their failures within the PFL space to produce richer intuitions about the learned concepts.

Traditional STEM instruction and learning can also benefit from using such an approach when introducing problem-solving tasks. In addition to direct instruction of this metacognitive approach, teachers could encourage students to use these strategies by simply giving students the opportunity and support to specify the source of their errors, resolve knowledge gaps, and work out novel and successful solutions. For example, this could be carried out very easily with a shift in the way teacher approach grading (i.e. allowing students to earn points for completing this process after getting several problems wrong, or grading a lab or project after several cycles of feedback that allow students to engage in error specification and info-seeking).

In addition to the push for 21st century skill development in the classroom, the influx of interest in design-thinking, applied subjects such as coding and Makerspaces, and hands-on inquiry-based learning also shows the myriad of applications that using a metacognition-based failure responses would be useful in. These fields traditionally emphasize iteration, rapid prototyping, and team-based design and learning; these processes should incorporate the error specification and info-seeking phase after each cycle of design, to optimize the next iteration students produce.

For educational technologists and designers, these findings presents a unique opportunity to integrate metacognition-based strategies into their learning systems and games in ways that deeply engage students and provoke meaningful problem-solving. As discussed previously in the future directions section, data mining techniques can be leveraged to optimize the delivery of prompts or other strategies in problem-solving contexts, detect when students encounter impasses that require more guided play or instruction, and constrain students when they are abusing or ignoring system mechanics, such as hints and prompts.

There are also wide implications for adoption beyond education, ranging from design and engineering to management consulting. Design-based industries, from fashion, to product innovation, to engineering and public planning, could benefit from research surrounding the strategies that could make failure productive, and could explicitly incorporate error-specification and information-seeking as part of their design process. Management consultancy companies such as Deloitte and KPMG, who rely on flexible and talented recruits to quickly learn about an industry and company to provide guidance to clients about improving processes, corporate structures, and products, could directly improve their services by training new recruits on how to accurately specify weaknesses and errors within a system (or their own understanding), seek out information to contextualize such weaknesses, and find solutions that alleviate those problems. It is likely that many senior designers, engineers, and consultants already adopt these kinds of processes in their system implicitly, but providing explicit training to entry-level staff could improve their skillsets more quickly and directly.

Finally, this study provides a concrete example of how games can enhance behaviors that are productive for learning, and how one such game can be incorporated into high school curriculum to bolster learning outcomes. Educational games can provide a motivating and effective problem-solving environment that allows students to meaningfully and concretely grapple with abstract concepts, that in turn can enhance their understanding of the formal instruction that follows. Furthermore, gameplay behaviors that translated into better learning outcomes, such as iterative dispositions and productive failure responses, could potentially translate into other academic tasks if teachers are willing to encourage transfer of such behaviors.

Conclusion

Failure is a contentious, seemingly paradoxical phenomenon: success stories often include a long string of failures that inevitably lead to triumph, yet many who encounter repeated failure see it as a sign to throw in the towel. People are often encouraged to be resilient and optimistic in the face of failure, yet persistence through failure without a productive response only leads to wheel-spinning and inevitable decay of confidence and helplessness. Failing can produce greater insight about the system the failure occurred in, but only if one possesses the skill to identify what caused the failure and how. Worse, while some failures signal an opportunity for potential growth and change that can be addressed by oneself, others signal a genuine lack of pre-requisite skill or understanding that require intervention or help from others. So, how can failure lead to success when there's seemingly such a wide variety of failures that can happen, and so much required to make that failure productive?

To productively diagnose, address, and respond to failure requires one to be resilient, accurate in self-appraisal, and well-quipped with skills to fix what went wrong. Yet, there's surprisingly little empirical literature on the process of making failure productive. Games are an environment where players appear to be especially resilient and skillful in response to failure, making it an optimal space to study the mechanisms through which failure can be made productive. Using a preparation for future learning (PFL) framework, this research sought to investigate what kinds of responses to failure that occur in an educational game environment are most effective for preparing students to learn from instruction that followed. Prior PFL research provided theoretical and empirical evidence that experiencing failure and grappling with concepts produces insights and intuitions that then better prepare students to learn from formal instruction that follows, but did not explicitly examine the role of how failure responses benefit understanding.

Study 1 indicated that having the opportunity to respond to one's failure in a physics game prior to instruction lead to deeper and more complex understanding, and that one particular strategy sophisticated learners used, info-seeking and fixing one's solution, was related to higher learning outcomes. This strategy encompassed several metacognitive steps: first, students had to come to a realization that there were elements of the system they did not understand; second, they had to specify what the source of their misunderstanding was that led to that error; third, they had to deliberately close that knowledge gap through info-seeking, and finally, they had to return to their incorrect solution in order to address their prior error, equipped with this new information. Study 2 investigated whether the induction of this strategy in response to failure would lead to higher learning outcomes, compared to those who were not provoked to use any strategy or those who were instead asked to make a global judgment of their understanding. Results indicated that effective use of this strategy was related to higher learning outcomes, but not all students who were prompted to use the strategy spent the necessary time on error-specification and info-seeking. Iteration was also a key feature of making the game experience productive for future learning, highlighting the benefits of experiencing failure frequently for producing richer intuitions about system mechanics. Yet, not all kinds of failure were effective for producing higher learning outcomes; one kind of failure signaled a fundamental lack of understanding of the underlying system, which undermined the benefits of grappling with the content prior to instruction.

Together, these studies add to the burgeoning body of research on the relationship between failure and learning, the complexities of when failures can be appropriately addressed by a learner versus when they’re a signal for external intervention, and the strategies one should use to make failure productive. These results imply that metacognition-based strategies in response to failure are an effective method to improve student learning outcomes, and offer promising implications for educational interventions, 21st century skill instruction, and other industrial applications, such as design, engineering, and consulting.

References

Adey, P., & Shayer, M. (1993). An Exploration of Long-Term Far-Transfer Effects Following an Extended Intervention Program in the High School Science Curriculum. *Cognition and Instruction*, *11*(1), 1–29.

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education*, *16*(2), 101–128.

Arena, D. (2012). Commercial video games as preparation for future learning, (Doctoral dissertation, Stanford University).

Baker, R. S. J. d, D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human Computer Studies*, *68*(4), 223–241.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, 128(4), 612–637.

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and Transfer: The Role of Mastery-Approach Goals in Preparation for Future Learning. *Journal of the Learning Sciences*, 2*1*(3), 399–432.

Belmont, J. M., Butterfield, E. C., & Ferretti, R. P. (1982). To secure transfer of training instruct self-management skills. In: Detterman D K, Sternberg R J (eds.) *How and How Much Can Intelligence Be Increased*, (pp. 147-154), Norwood, NJ: Ablex.

Benjamini, Y., Hochberg, Y., & Series, B. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological), 57*, 289-300.

Black, J. B., Khan, S. A., & Huang, S. C. D. (2014). Video games as grounding experiences for learning. In F.C. Blumberg (Ed.) *Learning by playing: Frontiers of videogaming in education.* New York, NY: Oxford University Press.

Black, J. B., Segal, A., Vitale, J., & Fadjo, C. (2012). Embodied cognition and learning environment design. In D. Jonassen and S. Lamb (Eds) *Theoretical Foundations of*

Learning Environments, New York, NY: Routledge.

Blumberg, F. C., Rosenthal, S. F., & Randall, J. D. (2008). Impasse-driven learning in the context of video games. *Computers in Human Behavior*, *24*, 1530–1541.

Borkowski, J., & Muthukrishna, N. (1992). Moving metacognition into the classroom: “Working models” and effective strategy teaching. In M. Pressley, K R Harris, & J. T. Guthrie (Eds.), *Promoting academic competence and literacy in school* (pp. 477-501). San Diego, CA: Academic Press.

Bransford, J. D., & Schwartz, D. L. (1999). Chapter 3: Rethinking transfer: A simple proposal with multiple implications. *Review of research in education, 24*(1), 61-100.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*, 32–42.

Chi, M. T. H. (2005). Commonsense Conceptions of Emergent Processes: Why Some Misconceptions are Robust. *The Journal of the Learning Sciences*, *14*(2), 161–199.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence.* New York, NY: Oxford University Press.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and Useradapted Interaction*, *4*, 253–278.

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety.* San Francisco, CA: Jossey-Bass.

Darnon, C., Butera, F., Mugny, G., Quiamzade, A., & Hulleman, C. (2009). Too complex for me!” Why do performance-approach and performance-avoidance goals predict exam performance?. European Journal of Psychology of Education, *24*, 423–434.

Davidson, J. E., Deuser, R., & Sternberg, R. J. (1994). The role of metacognition in problem solving. In J. Metcalfe and A. Shimarmura (Eds.) *Metacognition: Knowing about knowing*, 207-226. Cambridge, MA: Bradford

Detterman, D. K., & Sternberg, R. J. (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Norwood, NJ: Ablex Publishing.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087.

Dweck, C. S. (2006). *Mindset: The new psychology of success.* New York, NY: Random House.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review*, *95*(2), 256.

Fadjo, C. L., Hallman Jr, G., Harris, R., & Black, J. (2009). Surrogate embodiment, mathematics instruction and video game programming. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2009* (pp. 2787–2792).

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist 34,* 906.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation and learning, *Simulation & Gaming*, *33*, 441–467.

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-Learning and Digital Media*, *2*, 5–16.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–831.

Halverson, R., Shaffer, D., Squire, K., & Steinkuehler, C. (2006). June). Theorizing games in/and education. *In Proceedings of the 7th International Conference on Learning Sciences, International Society of the Learning Sciences*, *2006 SRC*, 1048–1052.

Hammer, J., & Black, J. (2009). Games and (preparation for future) learning. *Educational Technology Magazine: The Magazine for Managers of Change in Education, 49*, 29–34.

Honey, M., & Hilton, M. (2011). *Learning science through simulations and games.* Washington, DC: National Academies.

Hong, H.-Y., & Lin-Siegler, X. (2011). How learning about scientists' struggles influences students' interest and learning in physics. *Journal of Educational Psychology*, *104*(2), 469–484.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity*, *6*(3), 41–49.

Juul, J. (2013). *The art of failure: An essay on the pain of playing video games*. Cambridge, MA: Mit Press.

Kapur, M. (2006). Productive failure. *ICLS 2006 - International Conference of the Learning Sciences, Proceedings*, *1*(November 2011), 307–313.

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*, 379–424.

Klahr, D., & Chen, Z. (2011). Finding one's place in transfer space. *Child Development Perspectives*, *5*, 196–204.

Kuhn, D., & Pease, M. (2010). The dual components of developing strategy use. In H. S. Waters & W. Schneider (Eds) *Metacognition Strategy Use & Instruction*, (pp. 135-159). New York, NY: Guilford.

Loibl, K., & Rummel, N. (2014). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*, 74–85.

Malone, T. (1981) Toward a theory of intrinsically motivating instruction, *Cognitive Science, 5*, 333-369.

Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., & Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, *23*, 113–131.

Perkins, D. N., & Salomon, G. (1992). Transfer of learning. *International Encyclopedia of Education*, *2*, 6452-6457.

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology* 92(3), 544.

Reese, S. D. (2007). The Framing Project: Model for Media Research Revisited. *Journal of Communication*, *57*, 148–154.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, *21*, 267–280.

Rowe, E., Asbell-Clarke, J., & Baker, R. S. (2015). Serious games analytics to measure implicit science learning. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious Games Analytics* (pp. 343-360). Switzerland: Springer International Publishing.

Rowe, E., Baker, R. S. J. D., & Asbell-Clarke, J. (2015). Strategic game moves mediate implicit science learning. *Proceedings of the 8th International Conference on Educational Data Mining*, 432–435.

Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–52). Greenwich, CT: Information Age.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology 103(4)*, 759-7753.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in newton's playground. *The Journal of Educational Research*, *106*, 423–430.

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science*, *3*, 1–5.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Son, L. K., & Schwartz, B. L. (2002). The relation between metacognitive monitoring and control. in Perfect, T. J., & Schwartz, B.L. (Eds.), *Applied Metacognition* (pp. 15-38). Cambridge, UK: Cambridge University Press.

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions: Functions involving attention, observation and discrimination. *Psychological Review*, *8*, 553–564.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why Do Only Some Events Cause Learning During Human Tutoring? *Cognition and Instruction*, *21*(3), 209–249.

VanLehn, K., & Springer, U. S. (1988). Toward a theory of impasse-driven learning. In: Mandl H., Lesgold A. (eds), *Learning Issues for Intelligent Tutoring Systems. Cognitive Science.* (pp. 19-41). New York, NY: Springer.

Wouters, P., van Nimwegen, C., van Oostendorp, H., & van der Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*(2), 249–265.

Appendix A: Learning Measures

Open-Ended (OE) Measure

Code:___________________ 1 2 3

Draw, label, and explain a circuit with resistors in parallel.

How does it differ from a circuit with resistors in series?

Table 1: OE Correctness Rubric (9 Points Total)

Variable	0 Point	0.25	0.5	0.75	1 Point
Diagram					
Closed Loop	a line				shows that circuits behave in a loop that's fully enclosed.
Voltage	no voltage source depicted		depict voltage source but does not include positive/negative terminal	voltage source with positive/negative terminal	voltage source with positive/negative terminal AND "V" "Voltage"
Current	doesn't delineate current symbolically or with a label	symbol and/or label without direction	arrow or some other symbol showing direction of current OR label		indicates with both arrow and label "I" "Current"
Resistance	no resistors depicted		object on circuit (light bulb, squiggle, etc) OR label "R"		1 Point: Object WITH labels (symbol "R" or Resistor)
Parallel Resistance	resistors are not in parallel			resistors are parallel to one another	resistors are parallel to one another AND labeled "R1, R2, R3"
Parallel current	doesn't indicate (either with arrow or with "I") that current flows through both		depicts with arrows but not with label		depicts with labels "I1, I2, I3" that show separate currents flowing through resistors
Verbal					
Parallel current	none of the above		Ambiguous answer stating that current has different paths (i.e. "flow goes in more than one direction")	1 of 2	1) currents in parallel circuits split paths 2) currents may differ in their separate paths depending on resistance
Parallel voltage					Mention that voltage remains constant across parallel circuits
Series current			Ambiguous answer stating that current is universal to all components ("i.e. "flows through all elements")	1 of 2	Explicitly states that: 1) one single current/pathway flows through all components of the system AND 2) current is constant across all circuit components

Table 2: OE Complexity Rubric (24 Points Total)

Variable	0 Point	0.5	0.75	1 Point
Diagram				
Electrons				depicts electrons (i.e. electron flow, arrow from negative to positive, dots with label electron) in ANY diagram
Switch				diagram includes switch
Series Resistors				includes diagram with resistors in series
Series Current				indicates that current is the same with just "I" consistently
Game Example				Uses game example in their diagram (i.e. light bulb, looks like the game puzzle)
Formulas				
Ohm's Law				Uses Ohm's Law (V=IR)
Parallel Current				Parallel Current: IT = I1 + 12 + 13....
Parallel Resistance				Parallel Resistance: 1/RT = 1/r1 + 1/r2...
Series Voltage				Series Voltage: VT = V1 + V2 + V3
Series Resistance				Series Resistance: RT = R1 + R2 + R3...
Verbal				
Parallel current	none of the above	Ambiguous answer stating that current has different paths (i.e. "flow goes in more than one direction")	1 of 2	1) currents in parallel circuits split paths 2) currents may differ in their separate paths depending on resistance
Parallel voltage				Mention that voltage remains constant across parallel circuits(i.e. "The amount of the total voltage remains consistent"' "giving each equal voltage")
Series current		Ambiguous answer stating that current is universal to all components ("i.e. "flows through all elements")	1 of 2	Explicitly states that: 1) one single current/pathway flows through all components of the system (i.e. "all of the resistors get the same current", "resistors are all on the same path") 2) current is constant across all circuit components
Voltmeters			1 of 2	1) voltmeters are in parallel AND 2) voltmeters have very high resistance
Ammeters		ammeters go "to the side" or "next to" the element (not explicitly stating series)	1 of 2	1) ammeters are in series AND 2) have very low resistance

Electrons		Electrons travel from negative terminal to positive terminal
Voltage Compare	mentioning resistors (lightbulbs) are weaker in series than in parallel	ONE POINT EACH (TOTAL OF 3 POSSIBLE) 1) resistors in parallel circuits receive full (same) power, (leading to brighter light bulbs) 2) resistors in series circuits receive lower voltage because they share voltage; and 3) resistors in series may receive voltage in proportion to their resistance
Current Compare	current flows slower in series than in parallel	ONE POINT EACH (TOTAL OF 2 POSSIBLE) 1) current in series overall is slower BECAUSE all resistors contribute to slowing down 2) current in parallel may be faster depending on the resistance of the individual resistors
Resistance Compare	resistance is higher in series (without justification) compared to parallel	Compares the overall resistance in parallel and series circuits. I.E.: resistance in series are summed, but resistance in parallel are shared. Therefore, resistance is higher in series circuits compared to parallel
Electron Relate		Any answer that explicitly discusses electron flow in relation to resistance, current, or voltage.
Current Resistance		Discusses the reciprocal relationship between resistance and current i.e. More current flows through the resistors with lower resistance"
Voltage Discussion	"electrical pressure" or "how badly electrons want to get from one place to another"	"potential difference in charge between two points"
Other Examples		Using other, real-world examples to show differences between parallel and serial circuits
Resistor Infer	mention that parallel circuits don't break, but series circuits do	Inferring about what happens in other parts of the system when a resistor fails In parallel circuits, one failed resistor does not break the system because the current can flow through other paths; In series circuits, one failed resistor will break the system because the loop will have been opened, and the current has no other pathway to flow though

Post-Test

Quiz

Physics concepts:

Ohm’s Law: V = IR
Voltage: (V or v - Volts) The electrical potential between two points in a circuit.
Current: (I or i - Amperes) The amount of charge flowing through a part of a circuit.
Power: (W - Watts) Simply P = IV. It is the current times the voltage.
Source: A voltage or current source is the supplier for the circuit.
Resistor: (R measured in Ω - Ohms) A circuit element that "constricts" current flow.

* Required

1. **Code**

2. **Two identical resistors are connected in series. The voltage across both of them is 250 volts. What is the voltage across each one?** *

   Mark only one oval.

   ○ R1 = 125V and R2 = 125V
   ○ R1 = 250V and R2 = 0V
   ○ R1 = 150V and R2 = 100V
   ○ None of the above.

3. **Three resistors with 1Ω, 2Ω and, 3Ω are connected in parallel. What is the total resistance?**

   Mark only one oval.

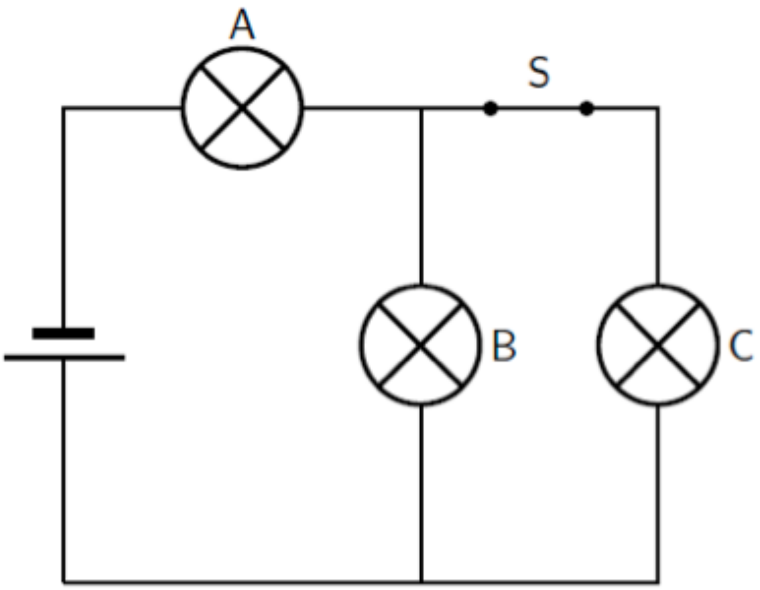
   ○ 6/3Ω
   ○ 3/6Ω
   ○ 11/6Ω
   ○ 6/11Ω

4. **Two resistors are connected in parallel with a voltage source. How do their voltages compare?** *

   Mark only one oval.

   ○ The voltage across both resistors is the same as the source.
   ○ The voltage across both resistors is half the voltage of the source.
   ○ One has full voltage, the other has none.
   ○ None of the above.

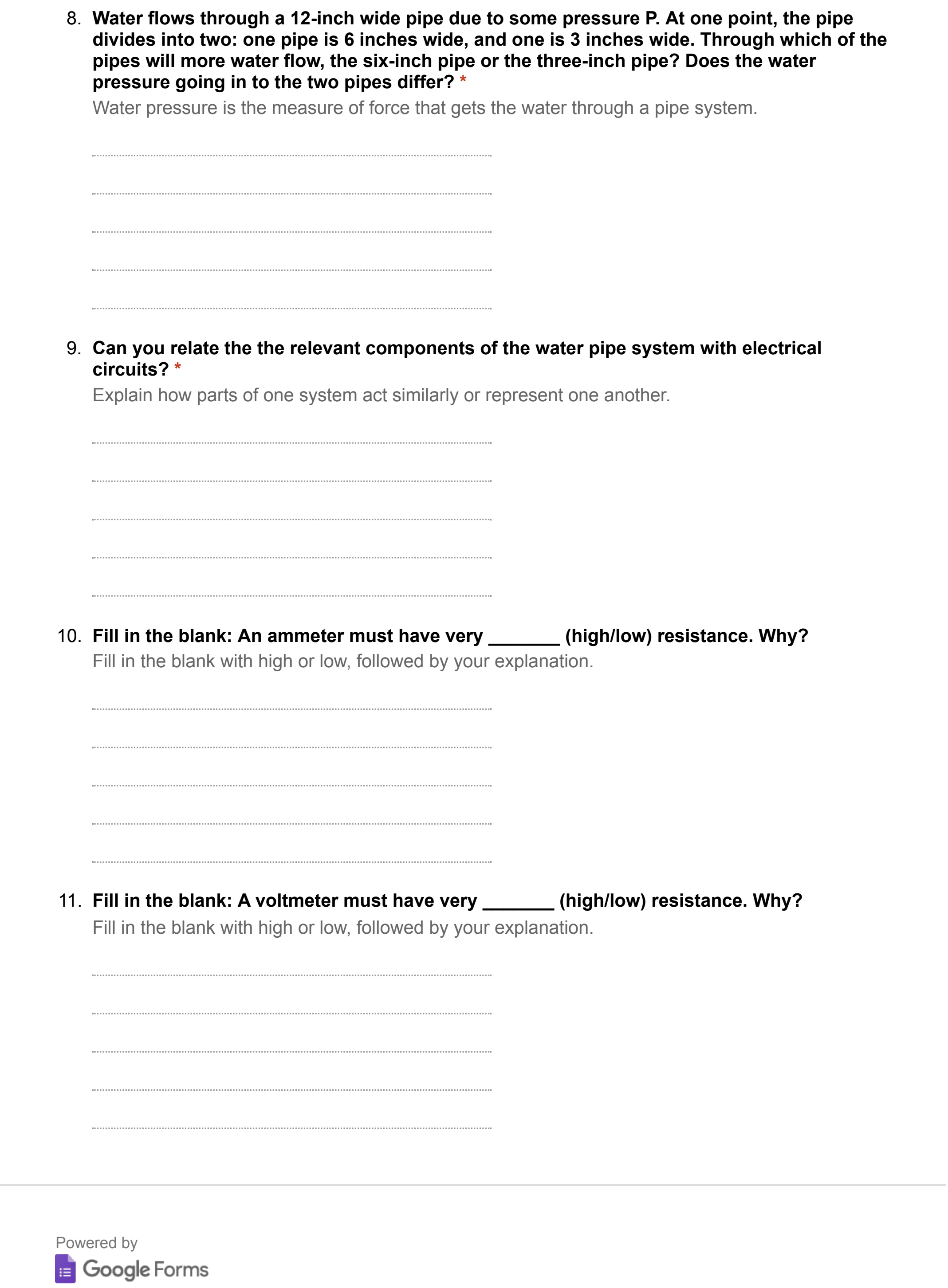
Please use the following circuit diagram for the next three questions.



5. **How bright is bulb A compared to B and C?** *

6. **How bright are the bulbs after switch S has been opened?** *

7. **How do the currents in bulbs A and B change when switch S is opened?** *



8. **Water flows through a 12-inch wide pipe due to some pressure P. At one point, the pipe divides into two: one pipe is 6 inches wide, and one is 3 inches wide. Through which of the pipes will more water flow, the six-inch pipe or the three-inch pipe? Does the water pressure going in to the two pipes differ?** *

   Water pressure is the measure of force that gets the water through a pipe system.

9. **Can you relate the the relevant components of the water pipe system with electrical circuits?** *

   Explain how parts of one system act similarly or represent one another.

10. **Fill in the blank: An ammeter must have very _______ (high/low) resistance. Why?**

    Fill in the blank with high or low, followed by your explanation.

11. **Fill in the blank: A voltmeter must have very _______ (high/low) resistance. Why?**

    Fill in the blank with high or low, followed by your explanation.

Powered by Google Forms

Table 3: Post-Test Transfer/Complexity Rubric

Question	*Transfer (9 Points Total)*	*Complexity (10 Points Total)*
Analogy_Flowrate	1 point: pipe A (six inch).	1 point: Mentioning pipe size is comparable to resistance 1 point: mentioning water will flow through pipe with less "resistance"
Analogy_WaterPressure	1 point: no, they do not differ. pipe A and B have the same pressure.	1 point: Mentioning water pressure is similar to voltage
Analogy_Map	1point for each: - Voltage = water pressure - Battery = pump - Current = water flow - Resistance = smaller pipes or pipes that make it difficult for water to flow through - Pipes = wire	1 point: Mentioning pipes are in parallel - ANYWHERE in analogy answer
PFL_Amm	1 point: "low" 1 point: it must have a low resistance so as to not disturb the current flow as it goes through the ammeter.	1 point: goes in series (.5 for ambiguous answer) 1 point: measures current 1 point: discussing that ideal ammeters don't exist
PFL_Volt	1 point: "high" 1 point: it must have a high resistance so that it does not disturb the flow of electrons through the resistors that the voltmeter is trying to measure	1 point: goes in parallel (.5 for ambiguous answer) 1 point: measures voltage 1 point: discussing that such ideal voltmeters don't exist

Appendix B: Survey Measures

Pre-Survey (Administered on Google Forms; current copy for reference only)

Code: ___________ Age: ____________ Gender: _________

1. Have you ever participated in physics/engineering extracurricular activities? (Please include and elaborate on summer camps, internships, extracurricular clubs, or other instances where you've participated in STEM-related activities. STEM stands for science, technology, engineering, and math. Check all that apply.)
   □ STEM summer Camp
   □ STEM Internship
   □ STEM Club
   □ I have not participated in STEM extracurricular activities.

2. On a scale of 1 to 5, how would you rate your interest in physics or engineering?
   □ 1 – Not at all
   □ 2 – A little
   □ 3 – Somewhat
   □ 4 - Very
   □ 5 – Absolutely

3. How confident are you that you could explain the concepts of an electrical circuit system to someone?
   □ 1 – Not at all
   □ 2 – A little
   □ 3 – Somewhat
   □ 4 - Very
   □ 5 – Absolutely

4. How confident are you that you could define what an electrical current is?
   □ 1 – Not at all
   □ 2 – A little
   □ 3 – Somewhat
   □ 4 - Very
   □ 5 – Absolutely

5. Do you like playing digital games?
   □ 1 – Not at all
   □ 2 – A little
   □ 3 – Somewhat
   □ 4 - Very
   □ 5 – Absolutely

6. Do you consider yourself a gamer?
   □ Yes
   □ No

7. Do you enjoy playing educational games?
   - □ 1 – Not at all
   - □ 2 – A little
   - □ 3 – Somewhat
   - □ 4 - Very
   - □ 5 – Absolutely

8. Do you often play educational games at school?
   - □ 1 – Not at all
   - □ 2 – Rarely (once a semester)
   - □ 3 – Somewhat (several times a semester)
   - □ 4 - Often (once a week or more)
   - □ 5 – All the time (several times a week)

School and Goals: Please tell us how much you agree or disagree with the following items on how you consider your work at school. Your responses will not be shared with anyone, and the researchers will not look at your responses until after the collection period is over.

On the following items, a "1" means "Strongly disagree - This doesn't apply to me at all", a "3" means "This is somewhat true for me", and a "5" means "Strongly agree - This is completely true for me."

9. I do my school work because getting good grades is important to me.
10. I do my school work because I'm interested in it.
11. I like school work best when it really makes me think.
12. It's important to me that I do as well or better than most of the other students in my classes.
13. I like school work that I'll learn from, even if I make a lot of mistakes.
14. It's very important to me that I don't look stupid in my classes.
15. It's important to me that I show my teachers that I'm smarter than the other students in my classes.
16. I like when I don't have to try very hard to do well in a class.
17. When I'm working on something difficult or challenging, I keep working until I've completely mastered it.
18. I like working on schoolwork that challenges me or is very difficult, even if it feels frustrating in the moment.

Program Post-Survey (Administered on Google Forms; current copy for reference only)

Code: ______________

1. Do you feel like your experiences with the game impacted how much you understood or learned through the class?
   - □ 1 – Negatively Impacted
   - □ 2 –
   - □ 3 – Null Effect
   - □ 4 -
   - □ 5 – Positively Impacted

2. How did the game impact how much you understood or learned through the class?

3. How confident are you that you could explain the concepts of an electrical circuit system to someone?
   - □ 1 – Not at all
   - □ 2 – A little
   - □ 3 – Somewhat
   - □ 4 - Very
   - □ 5 – Absolutely