Addressing the Gender Gap in Middle School Math Education through Digital Learning Games

CMU-HCII-24-100
January 2024

Huy Anh Nguyen
Human-Computer Interaction Institute, Carnegie Mellon University
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
hn1@cs.cmu.edu

Thesis Committee:

Bruce M. McLaren (Co-Chair)	Human-Computer Interaction Institute, CMU
John Stamper (Co-Chair)	Human-Computer Interaction Institute, CMU
Jodi Forlizzi	Human-Computer Interaction Institute, CMU
Jessica Hammer	Human-Computer Interaction Institute, CMU
Derek Lomas	Industrial Design Engineering, Delft University of Technology

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

There is an established gender gap in middle school math education in the U.S., where girls report higher anxiety and lower engagement than boys, which negatively impacts their performance and even long-term career choices. While digital learning games, with the ability to promote learning motivation and outcomes, have the potential to address this gap, there have been mixed results regarding gender differences in learning with games. Furthermore, prior research on the gender effects of learning games remains focused on the distinctions between boys and girls, without accounting for the spectrum of variance in gendered behavior, which can develop as early as middle school.

In my work, I have identified *Decimal Point*, a digital learning game that teaches decimal numbers and operations to middle school students, as an excellent platform for studying gender effects in digital learning games. Based on data from five prior *Decimal Point* studies with over 1,000 students, I have observed a consistent gender difference across all studies – girls tended to have lower prior knowledge, but better self-explanation performance and higher learning gains from the game. The difference in self-explanation also explains the relationship between gender and learning outcomes, suggesting that girls' better learning could be attributed to their self-explanation performance. At the same time, there were no differences in how boys and girls enjoyed the game, indicating that digital games can help close the gap in learning while not sacrificing enjoyment for all students.

I conducted two follow-up studies to better understand the reasons for the observed gender differences and the extent to which they generalize. In both studies, I also employed a multidimensional representation of gender, one that captures not only birth-assigned gender and gender identity, but also gender-typed occupational interests, activities and traits. The first study investigated how self-explanation and different learning platforms influenced the relationship between gender and learning outcomes. The second study examined the ways in which gender differences in learning outcomes and enjoyment manifested when changing the game narrative. Results from both studies indicated that, across different learning platforms and game narratives, girls learned more than boys thanks to their better performance in the self-explanation activities. Furthermore, analyses of multiple gender dimensions led to a more nuanced understanding of the gender effects than analyses of binary gender alone. For instance, while boys generally reported higher levels of engagement with the game than girls, analyses of multidimensional gender further revealed that girls with strong masculine-typed behaviors were also more engaged than others.

In summary, this work contributes (1) robust evidence of the benefits of self-explanation in helping girls learn from digital games, (2) insights on the use of multidimensional gender representation to capture nuances in gender differences with respect to learning, enjoyment and game preferences, (3) guidelines for designing effective learning games to bridge the gender gap in math education. In a broader sense, this research will advance knowledge on the multidimensionality of gender in learning game research and inform practical recommendations on aligning game features with individual learners to optimize learning and engagement.

1. Introduction

Digital and computer games are becoming increasingly popular and accessible, especially to young people. For example, Clement (2021) found that more than 80% of adolescents in the U.S play video games. Additionally, children's gaming behaviors have greatly increased due to social distancing and quarantine practices from the COVID-19 pandemic. Among 2,863 children surveyed by (Zhu et al., 2021), 83% admitted to playing video games every day during the school closure period, and 55% reported having played games for longer than intended. The appeal of games expands to other age groups as well: according to a 2019 survey by the NPD Group (NPD, 2019), 73 percent of Americans aged 2 and older play digital games, a 6% increase from the prior year. There are reports of more than 2.6 billion people world-wide being video game players, with an expected rise to over 3 billion people by 2023 (Gilbert, 2021).

The appeal of digital games to young people has led to the conceptualization of learning games (Gee, 2003; Prensky, 2006; Shaffer & Gee, 2006; Squire & Jenkins, 2003), which aim to promote learning through engaging game environments. While early digital learning games yielded rather weak evidence of learning benefits (Honey & Hilton, 2011; Mayer, 2014; Tobias & Fletcher, 2007), in recent years, more mature games have emerged as promising instructional platforms (Clark et al., 2016; Mayer, 2019; B. M. McLaren & Nguyen, 2023). These modern learning games – such as *Crystal Island* (Taub et al., 2020), *Zoombinis* (Rowe et al., 2021), *Physics Playground* (Shute et al., 2019), *TLCTS* (Lewis, 2010) and *Decimal Point* (McLaren et al., 2017a) – were developed based on established learning principles and have led to clear learning improvements for tens of thousands of students in rigorous experimental studies (McLaren & Nguyen, 2023). In turn, they open up further research opportunities on how learning games can benefit different populations of learners and how these benefits can be enhanced with the support of other instructional techniques.

A learning domain to which these questions are especially relevant is middle-school math education, where there exists an established gender gap with long-lasting implications. In particular, while boys and girls have been shown to have similar performance in standardized tests (Hyde et al., 2008; Lindberg et al., 2010), gender differences favoring boys still emerge when focusing on data representing top performers among students or in advanced areas of math (Breda et al., 2018; Ellison & Swanson, 2023). At the same time, girls often hold less positive attitudes towards math (Breda et al., 2018; Mejía-Rodríguez et al., 2021; Rodriguez et al., 2020). In high school, several studies have reported that girls have lower confidence, less excitement about and greater frustration toward math than boys (Arroyo et al., 2013; Else-Quest et al., 2010, 2013). However, this difference isn't present in elementary school (Punaro & Reeve, 2012; Ramirez et al., 2013; Young et al., 2012), suggesting that middle school is when math anxiety emerges among girls and, as such, a crucial time for addressing this issue. This is particularly important given the negative association between math anxiety and math performance – a meta-analysis by Namkung and colleagues (2019) found an overall effect size of $r = -.34$, with a stronger negative correlation on more complex math topics. Furthermore, while math self-efficacy is a predictor of greater interest in math careers for boys, math anxiety is a predictor of lower interest in math careers for girls (X. Huang et al., 2019). Thus, there remains a significant challenge in promoting math interest and achievement among girls while reducing their math anxiety, one which digital learning games may be well suited to address, thanks to their purported motivational benefits (Chapman & Rich, 2018; Hussein et al., 2021).

Unfortunately, digital game designers often work without empirical guidance for how to make learning games more effective, especially in how games differ in their support of girls versus boys. In some cases, this results in uninformed adoption of extrinsic rewards (referred to as “gamification”), such as points, badges, competition and levels, that often do not foster productive learning processes (Nicholson, 2013, 2012; Seaborn & Fels, 2015). In an attempt to appeal to young girls, the game industry too often has employed gender stereotypes without a clear understanding of gender-based preferences or outcomes (Everett et al., 2017; Shaw, 2015). Greater evidence of when and how boys and girls learn from digital learning games -- and especially how they might learn differently from games -- will help inform teachers’ choices about which digital learning games to incorporate into their teaching and how to enhance learning for all students.

Additionally, much of the current research on gender differences in educational technology usage has focused on distinctions between boys and girls, without accounting for the spectrum of variance in gendered behavior (Hyde et al., 2019). For digital games and learning games, in particular, a large body of research has identified gender differences in game preferences, but only through the lens of binary gender categories (Aleksić & Ivanović, 2017; Chou & Tsai, 2007; Greenberg et al., 2010; Hamari & Keronen, 2017; Romrell, 2014). Towards developing a more nuanced understanding of how gender influences the learning and playing experience, modern learning game research would benefit from incorporating additional dimensions of gender, such as gender-typed interest, activities, and traits (Liben & Bigler, 2002). Examining these attributes would clarify which gender dimensions and game features best predict learning outcome and how they interact (Egan & Perry, 2001). Furthermore, they will contribute to the development of more inclusive learning platforms across different age groups.

This thesis work examines how digital learning games can bridge the gender gap in middle-school math education and how their gender effects can be better understood via a multidimensional gender framework. I will explore these topics through research and development of the game *Decimal Point*, which has proven to be an excellent platform for exploring gender differences in learning with games. From the early development stages, *Decimal Point* was carefully designed to be appealing to all students (Forlizzi et al., 2014) while incorporating evidence-based learning principles of self-explanation (Isotani et al., 2010a; Wylie & Chi, 2014) and example-tracing tutor design (V. Aleven, McLaren, et al., 2016). An initial study of *Decimal Point* has demonstrated its advantages over a conventional tutor in promoting learning and enjoyment (McLaren et al., 2017a). Following up on this research, since the start of my Ph.D., members of the McLearn Lab and I have been conducting several studies on extended versions of the game, to investigate a variety of learning game research topics, including the effect of agency (Nguyen et al., 2018), indirect control (Harpstead et al., 2019a), instructional context (McLaren et al., 2022c), balance between learning and enjoyment (Hou et al., 2020a), and different types of prompted self-explanation (McLaren et al., 2022a; 2022b).

While these studies have uncovered important lessons on different aspects of learning game design, one finding remains steady throughout: the game has led to greater learning benefits for girls than for boys. This highly consistent result, in turn, inspires the overarching question of my thesis work:

Why and how do digital learning games lead to gender differences in learning outcomes?

The first part of this thesis reports on the experimental settings and key results of five previous *Decimal Point* studies, conducted over a period of four years, with more than 1000 students in grades 5 and 6. The Fall 2017 study investigated whether giving students control over which mini-games to play and when to stop, i.e., providing them with more agency, would lead to better learning or enjoyment (Nguyen et al., 2018). As a follow-up, the Spring 2018 study examined how game interface elements may inadvertently exert indirect control over students' choices and sense of agency (Harpstead et al., 2019a). The Fall 2019 study's focus was on evaluating the effects of exposing students to the game's underlying models of their learning and enjoyment (Hou et al., 2020a, 2022a). In the Spring 2020 study, a 2x2 experiment was conducted to test the effect of incorporating hints and error messages, as well as the effect of playing the game in the classroom versus at home (McLaren et al., 2022c). Finally, the Spring 2021 study compared the benefits of three forms of prompted self-explanation activities – menu-based, scaffolded and focused (McLaren et al., 2022b). A summary of the study topics and sample sizes is included in Table 1. Across all five studies, the following results have been consistently observed:

- Girls have tended to perform worse than boys at pretest.
- Girls have tended to have higher learning gains than boys in the test problems that are procedurally similar to the in-game exercises.
- Girls have reported similar levels of enjoyment from the game as boys.

Table 1.1: Overview of prior studies. M indicates boys and F indicates girls.

Study topic	**Sample size**	**Age *M* (*SD*)**	**Key outcomes**
Student agency	158 (81 M, 77 F)	11.15 (0.60)	See *Prior Study Results - Fall 2017*
Study agency and indirect control	237 (107 M, 130 F)	11.86 (0.47)	See *Prior Study Results - Spring 2018*
Learning versus enjoyment	159 (82 M, 77 F)	10.93 (0.64)	See *Prior Study Results - Fall 2019*
Hints and error messages, in-person study	153 (79 M, 74 F)	11.06 (0.86)	See *Prior Study Results - Spring 2020*
Hints and error messages, remote study	125 (61 M, 64 F)	11.80 (0.57)	See *Prior Study Results - Spring 2020*
Types of prompted self-explanation	208 (97 M, 111 F)	11.58 (0.58)	See *Prior Study Results - Spring 2021*

To explain why these consistent gender differences manifested, I then compared boys and girls' behaviors in the problem-solving and self-explanation activities in the game (for details about learning activities, refer to the *Decimal Point* game description below). I identified a similarly strong trend, where girls made significantly fewer errors in the self-explanation activities than boys. Furthermore, I found that self-explanation performance could explain the relationship between gender and learning outcomes in several *Decimal Point* studies. In particular, girls made consistently fewer self-explanation errors than boys and therefore achieved better learning outcomes, as measured by posttest and delayed posttest performance. On the other hand, there were no consistent gender differences in the problem-solving portion of the game, suggesting that either the prompted self-explanation activity or its interaction with the game environment of *Decimal Point* had led boys and girls to learn differently from *Decimal Point*.

Building upon these findings, I conducted two additional studies that (1) captured multidimensional gender data, extending upon the binary gender categories that have been employed thus far, and (2) investigated two potential pathways that may explain the observed gender differences. In particular, the gender effects in *Decimal Point* could be induced by the playful features of the game, which reduce the saliency of the math content and the likelihood of invoking math stereotype threats in girls (the *stereotype threat hypothesis*). Such effects may also result from the game's thematic details, which can be more appealing to girls, thereby promoting higher engagement from girls than from boys (the *engagement hypothesis*). These hypotheses were developed in consultation with my collaborators and advisors, and are expected to cover the most likely mechanisms that induced the gender effects in *Decimal Point*. To test the proposed hypotheses, I conducted a classroom study in the Fall of 2022 which manipulated whether students learned from the game *Decimal Point* or a conventional tutor with identical learning materials, and whether students performed self-explanation as part of their learning activities. The second study, conducted in the Spring of 2023, compared the original *Decimal Point* to a new game version that retained the instructional content but pivoted to a different thematic narrative, one more closely aligned with boys' preferences (H. A. Nguyen et al., 2023), with its emphasis on adventures and naval battles. Results from these studies further reinforced the benefits of self-explanation prompts in helping girls achieve better learning outcomes, in addition to demonstrating how students' engagement levels were influenced by the change in narrative. Additionally, I was able to uncover more nuanced variances in the learning outcomes and enjoyment of students via the use of multidimensional gender representation.

From this research project, I expect to make multiple contributions to the areas of gender studies, digital learning game design, and AI in education.

1. This thesis will provide an understanding of how gender – and in particular, multiple dimensions of gender – interacts with game features to produce different learning outcomes in digital learning games. Lessons learned from analyzing *Decimal Point* data will contribute foundational knowledge related to the learning processes of students with diverse gender backgrounds.
2. The project has the potential to transfer *Decimal Point*'s success with promoting girls' decimal learning outcomes to learning platforms in other domains where a similar gender

gap exists, including most STEM domains (Baram-Tsabari & Yarden, 2011; Eddy et al., 2014; Wang & Degol, 2017). By revealing mechanisms that underlie gender-based differences in digital learning games, this research can provide guidance for learning game designers to incorporate empirically validated game elements that promote inclusiveness and learning efficacy.

3. The gender dimensions examined in this work could serve as useful features for constructing individualized student models and enabling real-time adaptivity within AI-based games (B. M. McLaren & Nguyen, 2023), as well as digital learning platforms in general. Given prior evidence that students' interactions with intelligent tutors can vary considerably across different cultures (Ogan et al., 2015), the use of more nuanced demographic features, such as gender dimensions, would constitute a meaningful step towards increasing the effectiveness of AI in education (Holstein & Doroudi, 2022) and combating potential biases in the student modeling process (Baker & Hawn, 2022; Paquette et al., 2020).

It is also important to note that, while my research investigates learning differences through the lens of gender, I do not expect birth-assigned gender or gender identity alone to be predictive of learning outcomes. In fact, recent meta-analyses of standardized test data have suggested that, outside of the top performers, the gender gap in math learning between boys and girls has mostly closed (Lindberg et al., 2010; Meinck & Brese, 2019; Reardon et al., 2019). Instead, the observed gender differences in *Decimal Point* are likely driven by differences in students' interests or preferences, which the additional gender dimensions in my proposed work are designed to capture. Understanding these underlying individual factors will allow for more practical customizations that match students' playing and learning needs, without subjecting them to existing gender and demographic stereotypes.

In the next sections, I will cover related work from several areas of research that intersect with the studies of gender in *Decimal Point*, followed by a detailed description of the game and study measures. Then, I will report the gender effects identified in the five *Decimal Point* studies conducted in the past years[1]. While these studies have manipulated the game in different ways to examine a variety of game topics – such as student agency, indirect control, and types of prompted self-explanation – the analyses presented here will focus on the gender comparisons across several learning, enjoyment and game play measures. The corresponding publications, cited along with the introduction of each study, provide further details into the effects of the game feature manipulation for interested readers. I will then follow up with a meta-discussion of the gender trends observed across all prior studies, in combination with the new results utilizing multidimensional gender representations from the two follow-up studies.

[1] While the past work I will describe was substantially done by myself, much of the research would not have been possible without the significant contributions of others, particularly my advisors and those in the McLearn Lab. When possible, I will include a brief text acknowledging those who contributed.

2. Background

Gender and Math Learning

Prior work has shown that, overall, there are only small differences in boys and girls' math performance (Meinck & Brese, 2019; Reardon et al., 2019); however, more nuanced differences emerge when looking at specific age groups, skill levels and types of math. In particular, boys and girls have mostly similar math achievements in elementary and middle school, but more consistent differences favoring boys start to emerge in high school (Milovanović, 2020). In addition, across all grade levels, boys tend to perform better than girls among higher-performing students (Breda et al., 2018; Cimpian et al., 2016; Keller et al., 2022). On the other hand, girls tend to have higher math grades and do better in statewide standards-based math tests, while boys do better at tests that are less tied to the school curriculum, such as the SAT (Hyde et al., 2008; Lindberg et al., 2010). Finally, boys tend to do better in advanced areas of math, such as those that involve problem-solving (Hyde et al., 1990), but girls have an advantage on basic numerical skills and routine math problems that have set procedures for solving (Vasilyeva et al., 2009).

Larger gender differences, however, do manifest in other math-related outcomes. Compared to boys, girls often hold less positive attitudes toward math (Breda et al., 2018; Hill et al., 2016; Levine & Pantoja, 2021) and have lower confidence in their math skills (Ganley & Lubienski, 2016). This phenomenon may be attributed to the stereotype threat, which posits that being reminded of social group stereotypes impacts the performance of members in that group (Doyle & Voyer, 2016; Picho et al., 2013; Starr & Simpkins, 2021). Although gender-based differences in math achievement have diminished in recent decades (Lindberg et al., 2010; Reardon et al., 2019), stereotypes about men being better at math than women can still emerge early in childhood and persist through adulthood (Cvencek et al., 2011; Doyle & Voyer, 2016; Passolunghi et al., 2014; Starr & Simpkins, 2021). In turn, such perception may influence girls' performance in math and reduce their interest in STEM careers (R. B. Adams & Kirchmaier, 2016; Bian et al., 2017; Goldman & Penner, 2016; Ochsenfeld, 2016). Thus, broadening STEM participation entails fostering positive math affect among girls, particularly during late elementary and middle school, before they make choices about STEM coursework in preparation for college. Digital games such as *Decimal Point* provide a promising pathway towards this goal, given their increasing popularity among young players (Homer et al., 2012; Lobel et al., 2017; NPD, 2019) and inside the classroom (Takeuchi & Vaala, 2014), as well as their ability to both engage students and promote learning (Gee, 2003; Mayer, 2019; McLaren & Nguyen, 2023).

Gender and Digital Learning Games

Digital games are popular among men and women, and a recent meta-analysis found no gender differences in participants' intentions to play games (Hamari & Keronen, 2017). However, there are consistent gender differences in preferences relating to game speed, type, opportunities for social interaction, and avatar characteristics (Aleksić & Ivanović, 2017; Chou & Tsai, 2007;

Greenberg et al., 2010; Romrell, 2014). Specifically, male players tend to prefer faster-paced and more action-style games, while female players tend to prefer more puzzle-style games and games with social interaction (Chou & Tsai, 2007).

Gender differences in game preferences apply to digital learning games as well. Girls tend to rank goal clarity and social interaction as more important in digital learning games than boys, while boys tend to place more importance on challenge, progress feedback, and visual appeal (Dele-Ajayi et al., 2018). These preferences can produce meaningful differences in learning behaviors; for example, one study found that girls reported more positive feelings and increased help-seeking behaviors when a non-player “learning companion” was present, while boys did best without a learning companion (Arroyo et al., 2013). Drawing from the broader literature on digital game preferences, some educational game researchers have proposed adapting digital learning games based on gender to create more inclusive, equitable learning experiences (Connolly et al., 2009; Hou et al., 2020b; Kinzie & Joseph, 2008; Law, 2010; Pezzullo et al., 2017; Steiner et al., 2009). However, recommendations for adapting games based on gender typically rely on the intuitions of game designers or preferences observed through playtesting and focus groups. There remains a need to empirically validate these recommendations across multiple studies and student populations to better understand the interaction between game features and gender.

Among studies examining gender differences in learning from digital learning games, girls have sometimes been shown to have greater learning outcomes (Khan et al., 2017; Klisch et al., 2012; Tsai, 2017), enjoy learning games more (Adamo-Villani et al., 2008; Chung & Chang, 2017), and see greater value in educational games compared to boys (Joiner et al., 2011). At the same time, other research has reported no gender differences in learning outcomes or motivation (Chang et al., 2014; Clark et al., 2011; Dorji et al., 2015; Manero et al., 2016; Papastergiou, 2009). Few studies have taken an empirically rigorous approach to testing learning outcomes of digital learning games (i.e., randomly assigning students to a learning game versus a comparable non-game control, using both pretests and posttests) and fewer have reported investigating gender differences within those games. Among the six rigorous, controlled studies of math digital learning games identified in Mayer (2019)’s review, only two reported analyzing gender differences in learning (McLaren et al. 2017b; Papastergiou, 2009). While Papastergiou (2009) found no gender effect on learning, McLaren et al. (2017b) reported that girls benefited more from the game than boys in *Decimal Point*, which was then replicated across four other *Decimal Point* studies (Nguyen et al., 2022). My analysis presented in this work extends these prior results by performing a more comprehensive comparison between boys and girls in all published studies of *Decimal Point*.

Multidimensional Framework of Gender Representation

While the majority of prior research in social science and psychology that investigates gender differences has focused on the distinction between boys and girls (Cameron & Stinson, 2019), this view was significantly challenged in recent years. From a social perspective, the transgender activist movement (Beemyn et al., 2016; Stryker, 2017) and intersex activist movement (Dreger & Herndon, 2009; Reis, 2007) have raised awareness to the wide spectrum of gender identities, including the distinction between individuals who identify as *cis-gender* (whose birth-assigned gender and gender identity align), *trans-gender* (whose birth-assigned gender and gender identity do not align), and *non-binary* (whose gender identity is neither exclusively male nor exclusively female). From a research perspective, Hyde and colleagues (2019) have synthesized empirical evidence from multiple disciplines – including neuroscience, behavioral neuroendocrinology, psychology and developmental research – that undermines the gender binary framework, proposing instead that gender is complex and dynamic, comprising multiple interrelated but separate dimensions.

Recent research in gender studies has identified several prominent gender dimensions, including *birth-assigned gender* (the most common operationalization of gender, typically following a binary male-female categorization), *gender identity* (the internal sense of one's own gender, such as male, female, non-binary, gender fluid and self-defined), *gender typicality* (one's perceived similarity to both their own and another gender - Egan & Perry, 2001; Martin et al., 2017), *gender-typed interests, activities and traits* (one's masculine- and feminine-stereotyped occupational interests, activities and traits – Liben & Bigler, 2002). These dimensions can be captured via survey questionnaires, which have been shown to provide reliable measures, even for late elementary and middle school youth (Fast & Olson, 2018; Hyde et al., 2019; Liben & Bigler, 2002; Martin et al., 2017). While birth-assigned gender and gender identity are mostly aligned for individuals at this age range (Zhang et al., 2020), the other dimensions have been shown to be interrelated but separable (Hyde et al., 2019). In particular, gender-typed interests, activities and traits were shown to only have minor correlations with each other and with binary gender identity, in samples of late elementary and middle school students (Cook et al., 2019; Perry et al., 2019).

Additionally, while children tend to think of gender from a traditional male-female dichotomy in younger ages, their gender perception becomes more nuanced in late elementary schools (Brinkman et al., 2014). During workshops and narrative interviews, middle-schoolers have demonstrated the ability to treat gender as a multidimensional representation of themselves, and to incorporate expanded vocabularies of gender identity and expression to portray their own gender dimensions (Bragg et al., 2018; Renold et al., 2017). Notably, this nuanced understanding is found not only within trans-gender and non-binary youth, but also those who identify as cis-gender, and even children who do not perceive gender in terms of these distinctions can still reliably report on their gender-typed occupational interests, activities and traits (Cook et al., 2019; Egan & Perry, 2001; Liben & Bigler, 2002; Martin et al., 2017; Perry et al., 2019). Thus, these gender dimensions have the potential to elucidate the relationship between gender and learning outcomes in *Decimal Point* as well.

Prompted Self-explanation in Digital Learning Games

Self-explanation, a cognitive process which involves the student explaininng their approach or solution to deepen their understanding, is an established learning strategy that promotes deep and robust learning (Joo et al., 2020; Lawson & Mayer, 2021; Wylie & Chi, 2014). However, it has seen limited usage in digital learning games, due to concerns of disrupting the game flow (Killingsworth et al., 2015) or inducing extraneous cognitive processing (D. M. Adams & Clark, 2014; O'Neil et al., 2014). Among the game studies that do incorporate self-explanation activities, there have been mixed results regarding their learning benefits. O'Neil and colleagues (2014) reported that having self-explanation prompts aimed at helping learners make connections between math and game terminology was more effective than not having them. On the other hand, Adams & Clark (2014) compared menu-based self-explanation with explanatory feedback and a control condition with neither self-explanation nor explanatory feedback, but found no differences in learning across conditions.

An explanation for these inconsistent results is proposed by Wylie and Chi (2014), who bring attention to the different types of prompted self-explanation used in digital learning environments, noting that they belong to a continuum from highly constrained to unconstrained self-explanations. Highly constrained self-explanation prompts – or menu-based explanation, as defined by Johnson and Mayer (2010) – present the learner with a small set of options to choose from. Scaffolded self-explanations induce higher cognitive load by, for example, prompting learners to fill in multiple blanks in a statement with a given word bank. Focused self-explanation asks for the learner's own explanation, with some guidance on what to explain (e.g., "is 1.0111 bigger or smaller than 1.1? How do you know?" - McLaren et al., 2022). Finally, open-ended self-explanations prompt learners to generate their own explanation without guidance or focus, thus leading to the highest cognitive load. Although typing an answer into a text box is a rather unnatural interaction during game play that may disrupt the game flow (Csikszentmihalyi, 1990; Killingsworth et al., 2015), Wylie and Chi (2014) advocated for less constrained prompts, as they facilitate more active and constructive engagement, which in turn helps learners to activate and connect prior knowledge, fill the gaps in their understanding, and ultimately achieve more robust learning outcomes. With respect to *Decimal Point*, all studies prior to the spring of 2021 have employed menu-based prompted self-explanation and have reported consistent learning benefits; however, these prior studies did not aim to distinguish between the role of the problem-solving activities and the role of the self-explanation prompts in inducing these learning benefits. My analyses reported below will show that self-explanation does indeed play a mediating role in the relationship between gender and learning outcomes in the game.

3. The Learning Game *Decimal Point*

Decimal Point (McLaren et al., 2017a) is a single-player digital learning game designed as an amusement park-like experience and targeted at 5th and 6th grade students learning about decimal numbers. The game runs on the Internet, within a browser, and can be played on both computers and tablets. The game's source code was developed in Flash and later ported to HTML/JavaScript, while its back-end functionalities are supported by the Cognitive Tutor Authoring Tools (CTAT - Aleven et al., 2016). The game and all related materials (tests and questionnaires) have been deployed on the web-based learning management system, TutorShop (Aleven et al., 2009a), which serves as a platform for students to play the game and for researchers to collect data from game play. The anonymized log data are stored in the DataShop repository[2] (Koedinger et al., 2010) for subsequent data analyses.

Game Description

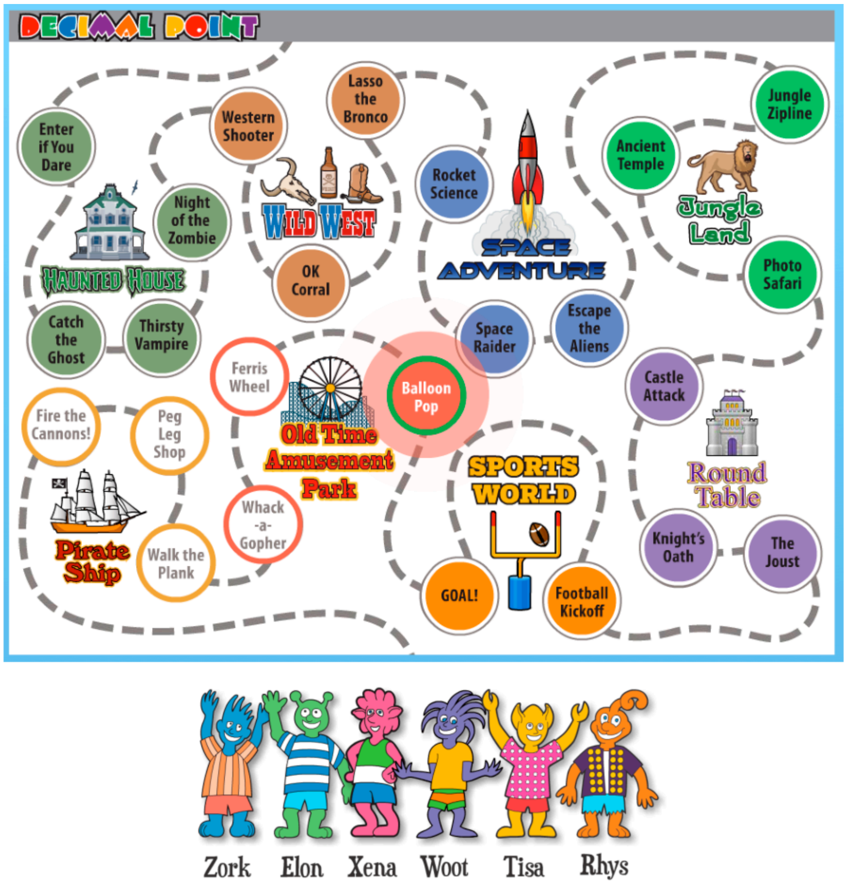


Figure 3.1: The main game map of *Decimal Point* and the alien characters.

The game is composed of a series of mini-games within the larger amusement park map (Figure 1, top). Students travel through different theme areas (e.g., *Haunted House*, *Wild West*), playing

[2] https://pslcdatashop.web.cmu.edu/Project?id=567

a variety of mini-games within each area (e.g., *Western Shooter* and *OK Corral* in the *Wild West* theme area). The mini-games that have been completed become colored in the game map, allowing students to track their progress; there is also a pulsing animation situated at the next mini-game that should be played (e.g., see “Balloon Pop” in Figure 1). Throughout their game play, students are accompanied by six fantasy characters introduced as aliens visiting Earth to learn about decimals (Figure 1, bottom). These characters show up at various points in the mini-games, encouraging students to play, providing feedback on incorrect answers and congratulating them on getting correct answers. Students do not score points or compete with others; rather, they simply make their way through the amusement park and are commended upon completing the journey by the fantasy characters (Figure 2).

DECIMAL POINT
The FANTASTICALLY FABULOUS WORLD of FRACTIONAL FUN!
We hope you learned something, too!

Figure 3.2: The ending scene of *Decimal Point.*

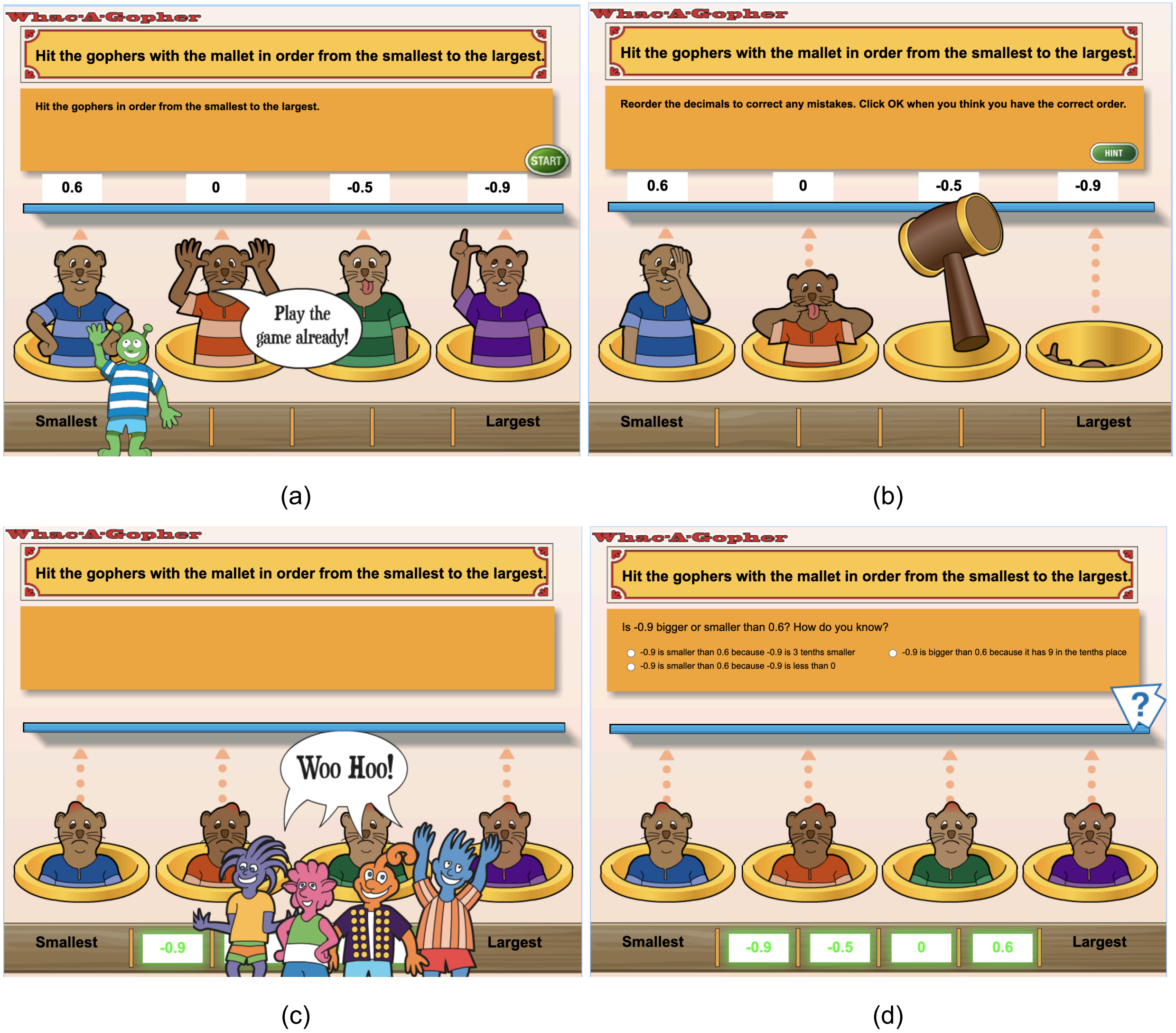
While the amusement park and alien characters contribute to an immersive gaming experience, the primary learning activities take place within the 24 mini-games shown in the game map (Figure 3.1). Each mini-game involves one of five types of decimal problems, as shown in Table 2.1. After solving each problem, students answer a multiple-choice self-explanation prompt to reinforce their learning. As an example, in the mini-game *Whac-A-Gopher*, students have to hit the four gophers in the correct order based on their associated number labels, from smallest to largest (Figure 3.3). These number labels were designed to target the misconception that decimals smaller than 1.0 are negative (Isotani et al., 2010a; Stacey et al., 2001). Students also need to be quick in thinking and acting, as the gophers pop up and retreat at random times. Once the four gophers have been hit, students receive immediate feedback about the correctness of their ordering, and can rearrange the number labels if they are incorrectly sorted. After successfully finishing this activity, students move on to a multiple-choice self-explanation question relevant to the sorting exercise they just completed. Every mini-game has a similar outline of problem-solving embedded in game activities, followed by prompted self-explanation.

Students don’t face any penalty on incorrect responses and can resubmit answers as many times as needed; however, they are not allowed to move forward without successfully finishing both the problem-solving and self-explanation prompt in a mini-game.

Table 2.1: The list of game types and their game activities in *Decimal Point*.

Game type	Activity
Number Line	Locate the position of a given decimal number on the number line
Addition	Add two given decimal numbers by entering the carry digits and the sum
Sequence	Fill in the next two numbers in a sequence of given decimal numbers
Bucket	Compare given decimal numbers to a threshold number and place each number in a “less than” or “greater than” bucket
Sorting	Sort a list of given decimal numbers in ascending or descending order

Decimal Point is built from rigorous research in learning science and game design. From a learning science perspective, the game targets decimal numbers due to the established difficulties that students have faced in this domain (Glasgow et al., 2000; Irwin, 2001), which may persist even into adulthood (Stacey et al., 2001). The in-game exercises were designed to target the most common decimal misconceptions (Isotani et al., 2010b) and leverage the benefits of prompted self-explanation in promoting deep and robust learning (Chi et al., 1989, 1994; Johnson & Mayer, 2010; Mayer & Johnson, 2010; Rittle-Johnson, 2006; Wylie & Chi, 2014). From a game design perspective, development of the game began with a competitive analysis of over 100 educational games for middle-school children, which identified five prominent design patterns: adaptivity, optional help, on-demand support, detailed tutorials, and immediate feedback. These patterns were consolidated into three initial game concepts, which were further refined through playtesting co-design sessions (Burkett, 2012; Walsh, 2009) with 32 middle school students. By consolidating the characteristics that were proposed during these sessions – such as the inclusion of diverse actions and colors, as well as familiar places and events – the research team settled on the amusement park theme. In light of prior research on gender preferences in games and learning games, the amusement park was chosen to be equally appealing to both boys and girls. Subsequent development was carried out over a year, focusing on brainstorming the theme areas and mini-game settings that align with the overall narrative and support student learning. Further details about the design process are reported in Forlizzi et al. (2014).



Hit the gophers with the mallet in order from the smallest to the largest.

Hit the gophers in order from the smallest to the largest.

Reorder the decimals to correct any mistakes. Click OK when you think you have the correct order.

Is -0.9 bigger or smaller than 0.6? How do you know?

○ -0.9 is smaller than 0.6 because -0.9 is 3 tenths smaller
○ -0.9 is smaller than 0.6 because -0.9 is less than 0
○ -0.9 is bigger than 0.6 because it has 9 in the tenths place

(a)

(b)

(c)

(d)

Figure 3.3: An example mini-game, *Whac-A-Gopher*, in the *Sorting* problem type and *Old Time Amusement Park* theme. Students are introduced to the game by an alien character (3a), play through the game (3b), get congratulated on finishing the problem-solving activity (3c), and finally solve the self-explanation prompt (3d).

Building on the above work, at the start of my Ph.D., I converted *Decimal Point*'s code base from Flash to HTML5, CSS3 and JavaScript. This transition helps the game comply with modern browser standards and allows for more rapid development of additional game features. For the past four years, members of the McLearn Lab and I have conducted five classroom studies on the game. The first three studies manipulated how students progressed through the mini-games, while maintaining the original learning content and gameplay mechanics (Harpstead et al., 2019; Hou et al., 2021; Nguyen et al., 2018). The latter two studies retained the linear mini-game progression from the base game, but adjusted the in-game instructional support and self-explanation prompts. Critically, all five studies employed the same pretests and posttests, which allow for a consistent measure of learning across studies. The survey measures did differ from study to study, depending on which enjoyment constructs were relevant to the game elements being manipulated. However, the overall experimental procedure, as outlined below, was consistent across studies.

Experimental Procedure

All *Decimal Point* studies were conducted on local elementary and middle schools in a northeastern U.S. city. Each study took place during students' regular class times and lasted six days; the materials tackled on the first five days included a pretest, a demographic questionnaire, game play, an evaluation questionnaire and posttest; the sixth and final day was reserved for the delayed posttest. Participants completed the pretest and demographic questionnaire on the first day, played the game for up to three class days, proceeding at their own pace, then completed an evaluation survey and posttest immediately after finishing the game, as well as the delayed posttest one week later.

The test items were identical across all four studies. Each test consists of 43 items; most items were worth one point each, while some multi-part questions were worth several points, for a total of 52 points per test. The questions were designed to probe for specific decimal misconceptions and involved either one of the five decimal activities in Table 2.1 or conceptual questions (e.g., "Is a longer decimal number larger than a shorter decimal number?"). Three test forms (A, B and C) that were isomorphic and positionally counterbalanced across conditions were used. In other words, one student may have forms A, B, C for pretest, posttest and delayed posttest, while another student may have forms B, C, A instead. Results from all previous studies indicated no student performance difference among the three test forms at pretest, posttest, or delayed posttest (Harpstead et al., 2019; Hou et al., 2020; McLaren et al., 2022b; 2022c; Nguyen et al., 2018).

Each study of *Decimal Point* also incorporated two surveys: a pre-intervention demographic survey and post-intervention evaluation survey. The demographic survey asked for basic information about the student's age, gender (male/female) and math experience. In the evaluation survey, which was taken immediately after game play, students rated several statements about their enjoyment of the game elements, on a Likert scale from 1 ("strongly disagree") to 5 ("strongly agree"). The survey contents were based on the study topics and therefore differed across studies. A summary of the evaluation survey items in each study is as follows (see Tables 2-5 in the Appendix for the full surveys):

- In the Fall 2017 study, the evaluation survey consisted of 11 items and covered three factors: *enjoyment of content* (4 items – e.g., "I liked doing this lesson"), *enjoyment of interface* (5 items – e.g., "I liked the way the material was presented on the screen"), and *math attitude* (2 items – e.g., "the lesson made me feel that math is fun").
- The Spring 2018 study surveyed students in two enjoyment factors: *enjoyment of content* (4 items – e.g., "I would like to do more activity like this") and *enjoyment of interface* (4 items – e.g., "It was easy to enter my answer into the system").
- In the Fall 2019 study, the evaluation survey covered three factors: *multidimensional engagement* (6 items adapted from Ben-Eliyahu et al. (2018) – e.g., "I felt frustrated or annoyed"), *game engagement* (5 items adapted from Brockmyer et al. (2009) – e.g., "I

lost track of time"), and the enjoyment dimension of *achievement emotions* (6 items adapted from Pekrun (2005) – e.g., "reflecting on my progress in the game made me happy"), for a total of 17 items.

- The Spring 2020 and Spring 2021 studies retained two factors from the Fall 2019 study – *multidimensional engagement* (Ben-Eliyahu et al., 2018) and *achievement emotions* (Pekrun, 2005), while incorporating 9 additional items from the *Player Experience Inventory* (Abeele et al., 2020 – e.g., "playing the game was meaningful to me").

Measures of Game Play, Learning and Enjoyment

To measure gender differences in learning, the 43 test items were partitioned into three groups, based on their level of learning transfer: 20 items were classified in the *Near transfer* group, 8 items in the *Middle transfer* group, and 15 items in the *Far transfer group*. This assignment is based on the taxonomy of transfer by Barnett and Ceci (2002), where near transfer items can be solved with identical procedures from those learned in the game, middle transfer items required modifications of the learned procedures but retain the problem representation, and far transfer items require an understanding of the underlying decimal principles. For example, based on the sorting game in Figure 3.3, a near transfer problem is "Sort the following list of decimals from largest to smallest: 7.681, 7.2, 7.15, 7.9," a middle transfer problem is "Which number is closest to 4.5? 4.555, 4.05, 4.4, or 4.6," while a far transfer problem is "Is a shorter decimal always smaller than a longer decimal number?" (more examples of the test items at each transfer level are included in Table A.1 of the Appendix). Under this classification, I measured the *pretest scores*, *learning gains* (difference between posttest and pretest scores) as well as *delayed learning gains* (difference between delayed posttest and pretest scores) at each transfer level.

To measure gender differences in game play, I considered four metrics: *problem-solving duration*, *problem-solving errors*, *self-explanation duration,* and *self-explanation errors*, where the durations are measured in minutes. The first two metrics reflect how students played through the problem-solving activity in the mini-games (Figure 3.3b), while the latter are based on their attempts at the self-explanation prompt at the end of each round (Figure 3.3d). As the number of mini-game rounds played by each student may differ, each of the four metrics above is summed over the student's entire playthrough and then divided by their number of mini-game rounds, yielding an average-per-round measure.

To measure gender differences in enjoyment, I computed the average Likert ratings of the relevant items for each enjoyment factor in the post-intervention evaluation survey. For example, if the survey contained three items related to the *enjoyment of interface* (e.g., "I liked the way the material was presented on screen," "I liked the way the computer responded to my input"), each student would have a representative enjoyment score for this factor, computed from their average ratings of the three relevant items. While it is not possible to observe enjoyment trends across studies, due to the enjoyment factors differing from study to study, it would still be meaningful to compare boys and girls' enjoyment within each individual study, to see whether any particular game element resonates strongly with students from one gender group.

4. Investigation of the Gender Differences in *Decimal Point*

This section reports on the experimental design and results of the five prior *Decimal Point* studies that I and members of the McLearn Lab conducted, as outlined in Table 1.1. While these studies explored different learning game topics, my data analyses focus on the gender effect via the following research questions:

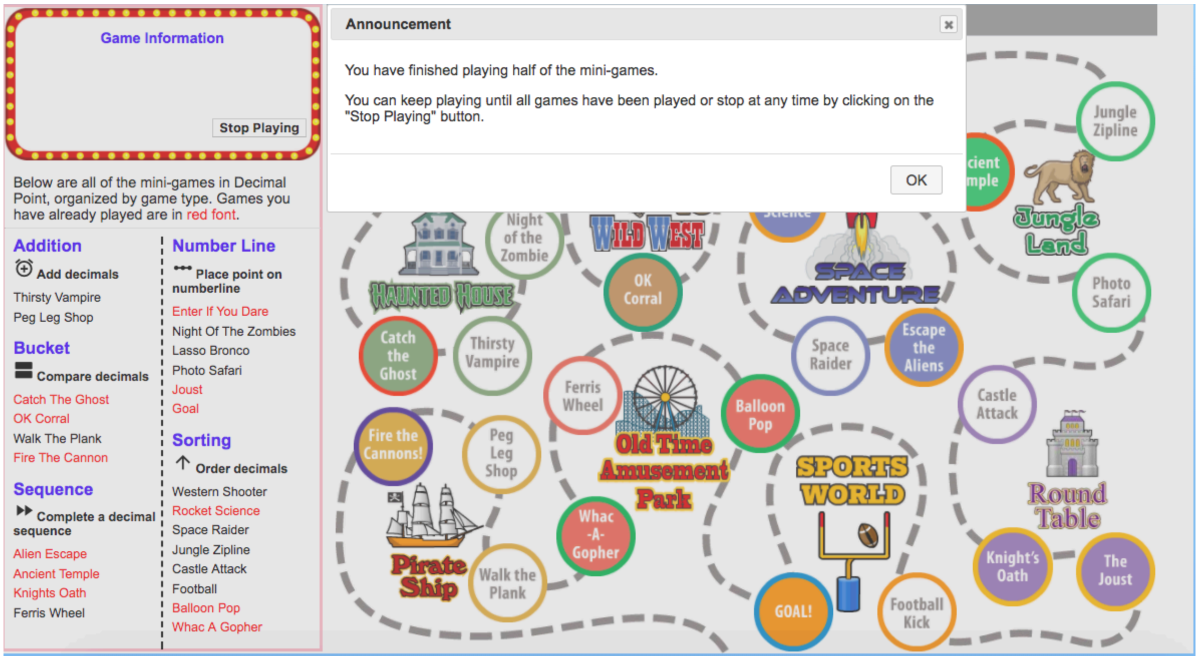
RQ1: *Is there a difference in learning outcomes between boys and girls?*
RQ2: *Is there a difference in problem-solving performance between boys and girls?*
RQ3: *Is there a difference in self-explanation performance between boys and girls?*
RQ4: *Is there a difference in enjoyment between boys and girls?*

To compare how boys and girls differ on the above metrics, I use the analysis of variance (ANOVA) test and include η_p^2 as the indicator of effect size. According to Cohen (2013), the η_p^2 benchmarks for small, medium and large effects are 0.01, 0.06 and 0.14 respectively. Across all studies, I also removed outlier students whose learning gains are more than 2.5 standard deviations away from the mean.

Fall 2017 Study on Student Agency

Reported by Nguyen et al. (2018), this study was motivated by whether agency – the capability for students to make their own decisions in how they play, and a key aspect in many computer games – is helpful to learning. Many learning platforms have given students agency over instructionally irrelevant choices – such as customizing game icons (Cordova & Lepper, 1996) and personalizing the interface (Snow et al., 2015) – as a simple way of applying gamification. The Fall 2017 study, however, sought to examine agency in a more meaningful context, for both learning and playinng, by letting students decide which order of mini-games to play and when to stop playing. In particular, the study involved two conditions: Low Agency and High Agency. The Low Agency condition featured the base game from a prior study by McLaren et al. (2017), where students played through all 24 mini-games in a fixed order, with two rounds of each mini-game (Figure 1). On the other hand, the High Agency condition gave students the option to play the mini-games in any order, and to finish the game any time after having completed 24 mini-game rounds (Figure 4).

Results from this study indicated that there were no significant differences in learning outcomes and enjoyment between the Low Agency and High Agency conditions (Nguyen et al., 2018). A post-hoc analysis also showed that most students in the High Agency condition still followed the canonical mini-game ordering, which might explain why their learning and game experience was similar to those of students in the Low Agency condition. The gender effects identified in this study, based on data from 81 boys and 77 girls, are reported as follows.



Game Information

Stop Playing

Below are all of the mini-games in Decimal Point, organized by game type. Games you have already played are in red font.

Addition

Add decimals

Thirsty Vampire
Peg Leg Shop

Bucket

Compare decimals

Catch The Ghost
OK Corral
Walk The Plank
Fire The Cannon

Sequence

Complete a decimal sequence

Alien Escape
Ancient Temple
Knights Oath
Ferris Wheel

Number Line

Place point on numberline

Enter If You Dare
Night Of The Zombies
Lasso Bronco
Photo Safari
Joust
Goal

Sorting

Order decimals

Western Shooter
Rocket Science
Space Raider
Jungle Zipline
Castle Attack
Football
Balloon Pop
Whac A Gopher

Announcement

You have finished playing half of the mini-games.

You can keep playing until all games have been played or stop at any time by clicking on the "Stop Playing" button.

OK

Figure 4.1: The main game map in the High Agency condition, after the student has played one-half of the mini-games and is given the option to stop. A "Stop Playing" button appears on the dashboard in the upper left.

Study Data

This study involved 197 students from two middle schools. Among them, 32 participants were excluded from analyses because they did not fully complete all materials and measures in the study. Additionally, 7 participants were removed due to having learning gains or delayed learning gains that are 2.5 standard deviations away from the mean. The remaining 158 students, including 81 boys and 77 girls, had a mean age of 11.15 (*SD* = 0.60).

Gender Comparisons

RQ1: *Is there a difference in learning outcomes between boys and girls?*
Figure 5 shows the test score comparison by gender. While the average test score of girls remained lower than that of boys throughout, a smaller gap between these two groups was observed in the posttest and delayed posttest than in the pretest. To examine which part of the tests led to this shrinking gap, Table 4.1 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains by gender at each transfer level. There was a trend of boys outperforming girls at all three transfer levels at pretest, especially at the near transfer level, where the difference was statistically significant. However, after playing the game, girls achieved significantly higher learning gains and delayed learning gains at the far transfer level.

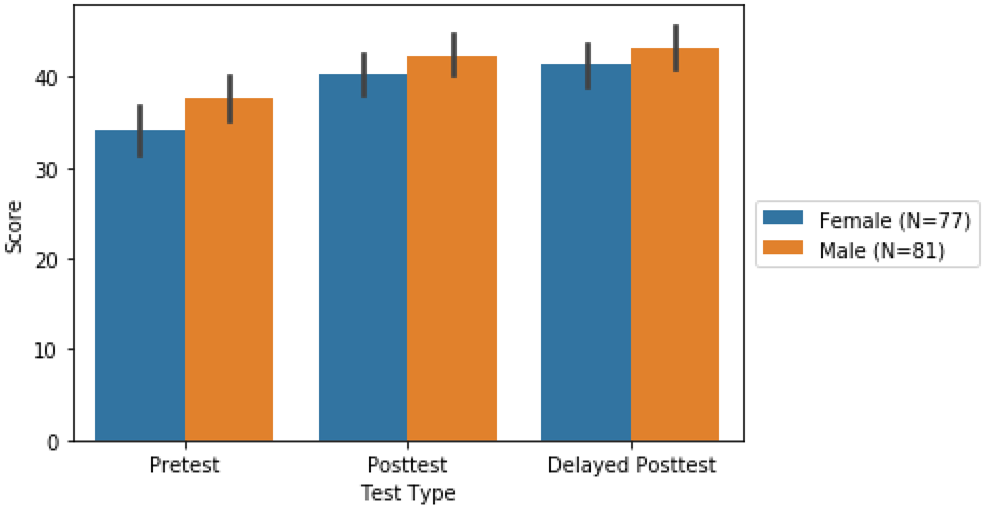


Figure 4.2: The performance of boys and girls in each test in Fall 2017. Error bars denote the 95% confidence interval around the mean.

Table 4.1: Comparison of test performance and learning gains by gender at each transfer level.

Category	**Transfer**	**Male M (*SD*)**	**Female M (*SD*)**	**Statistical result**
Pretest score	Near (*)	13.642 (4.978)	11.935 (5.247)	*F*(1, 156) = 4.403, *p* = .037, η_p^2 =.027
	Middle	4.580 (2.024)	4.403 (2.363)	*F*(1, 156) = 0.258, *p* = .612, η_p^2 =.002
	Far (†)	13.642 (5.283)	12.195 (4.888)	*F*(1, 156) = 3.185, *p* = .076, η_p^2 =.020
Learning gains	Near	3.099 (3.942)	3.909 (4.265)	*F*(1, 156) = 1.540, *p* = .216, η_p^2 =.010
	Middle	0.407 (1.523)	0.299 (1.598)	*F*(1, 156) = 0.192, *p* = .662, η_p^2 =.001
	Far (*)	0.840 (2.648)	1.818 (3.077)	*F*(1, 156) = 4.507, *p* = .033, η_p^2 =.029
Delayed learning gains	Near	2.938 (4.041)	3.896 (3.926)	*F*(1, 156) = 2.280, *p* = .133, η_p^2 =.014
	Middle (†)	0.630 (1.427)	0.156 (1.679)	*F*(1, 156) = 3.666, *p* = .057, η_p^2 =.023
	Far (*)	1.593 (3.089)	2.714 (3.634)	*F*(1, 156) = 4.384, *p* = .038, η_p^2 =.027

(*) p < .05; (†) p < .1

RQ2: *Is there a difference in problem-solving performance between boys and girls?*
A one-way ANOVA showed a significant gender difference in game duration per round in minutes, *F*(1, 156) = 11.727, *p* = .001, η_p^2 = .086, where boys (*M* = 0.786, *SD* = 0.475) spent less time playing the game than girls (*M* = 1.131, *SD* = 0.766). There was also a significant gender difference in the number of game errors per round, *F*(1, 156) = 7.16, *p* = .008, η_p^2 = .044, where boys (*M* = 1.784, *SD* = 1.382) had fewer errors than girls (*M* = 2.538, *SD* = 2.101).

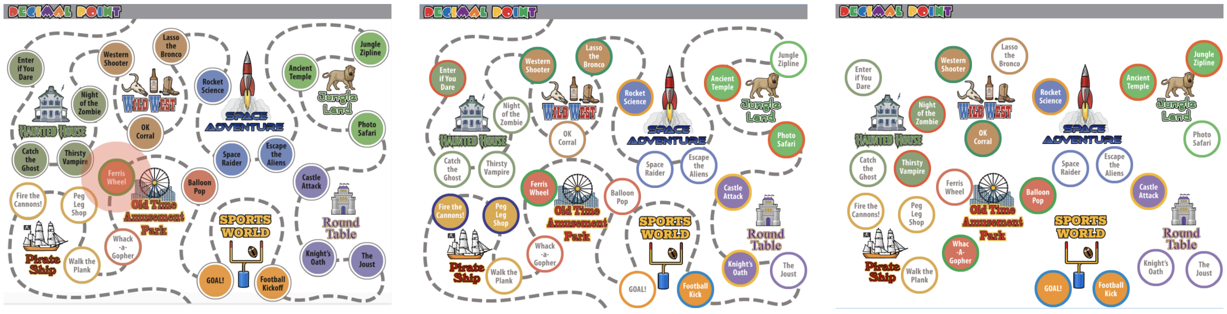
RQ3: *Is there a difference in self-explanation performance between boys and girls?*

A one-way ANOVA showed a marginally significant gender difference in self-explanation duration per round in minutes, *F*(1, 156) = 3.072, *p* = .082, η_p^2 = .019, between male (*M* = 0.421, *SD* = 0.140) and girls (M = 0.458, *SD* = 0.120), with girls trending toward longer self-explanation times. Additionally, there was a significant difference in number of self-explanation errors, *F*(1, 156) = 5.735, *p* = .018, η_p^2 = .035, where boys (*M* = 0.661, *SD* = 0.458) made significantly more errors than girls (*M* = 0.505, *SD* = 0.354).

RQ4: *Is there a difference in enjoyment between boys and girls?*
As previously mentioned, the post-intervention evaluation survey in this study covered three enjoyment factors: lesson enjoyment, ease of interface use, and math attitude. A series of one-way ANOVA showed no significant gender differences in lesson enjoyment, *F*(1, 156) = 0.039, *p* = .844, η_p^2 < .001, math attitude, *F*(1, 156) = 1.629, *p* = .204, η_p^2 = .010, or ease of interface use, *F*(1, 156) = 0.046, *p* = .831, η_p^2 < .001.

Spring 2018 Study on Agency and Indirect Control

Reported by Harpstead et al. (2019), this study was conducted to further examine the effect of agency in *Decimal Point*, building on the concepts of self-determination (Reeve et al., 2003) and contextual autonomy (Deterding, 2016), which posit that situational contexts from unrelated design choices may diminish students' feeling of having control and, in turn, their agency. In the context of *Decimal Point*, the dashed line on the game map (Figure 4.3) may be an indirect control factor that prompted students to follow the canonical mini-game sequence, even when they were given agency over mini-game selection. To test this hypothesis, I and members of the McLearn Lab designed three study conditions: Low Agency, High Agency and High Agency without Line. The first two conditions were identical to those used in the Fall 2017 study, while the third was a variant of the High Agency condition without the dashed line on the map (Figure 6).



(a) (b) (c)

Figure 4.3: The game map versions based on the three agency conditions in the study: (a) Low Agency, (b) High Agency, (c) High Agency without Line.

Results from this study indicated that removing the dashed line led to students exercising more agency, measured by deviation from the canonical path, and achieving higher learning efficiency (Harpstead et al., 2019). At the same time, there was no condition effect on enjoyment, suggesting that indirect controls such as the dashed line does not directly impact learning outcomes or enjoyment while still influencing the overall learner experience. The gender effects identified in this study are reported as follows.

Study Data

287 students from two public middle schools participated in this study. Among them, 35 were excluded because they did not fully complete all the tests and materials. Additionally, I removed data from 13 students because of login errors, where students accidentally logged in to their classmates' accounts and used the materials. One student was excluded as an outlier due to having learning gains that are 2.5 standard deviations away from the mean. The remaining 238 participants included 107 boys, 130 girls, and one student whose gender data was missing. Because my analyses focus on gender difference, I also excluded the student with missing data; thus, my final sample consists of 237 participants, with a mean age of 11.86 (*SD* = 0.47).

Gender Comparisons

RQ1: *Is there a difference in learning outcomes between boys and girls?*
Figure 7 shows the test score comparison by gender. On average, girls had lower scores than boys on the pretest, but similar scores on the posttest and delayed posttest. To examine which part of the tests allowed girls to catch up with boys, Table 4 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between boys and girls at each transfer level. We observed that boys outperformed girls at all three transfer levels at pretest, especially at the near transfer level, where the difference was significant. However, after playing the game, girls had significantly better performance than boys on the near- and middle-level items of the posttest.

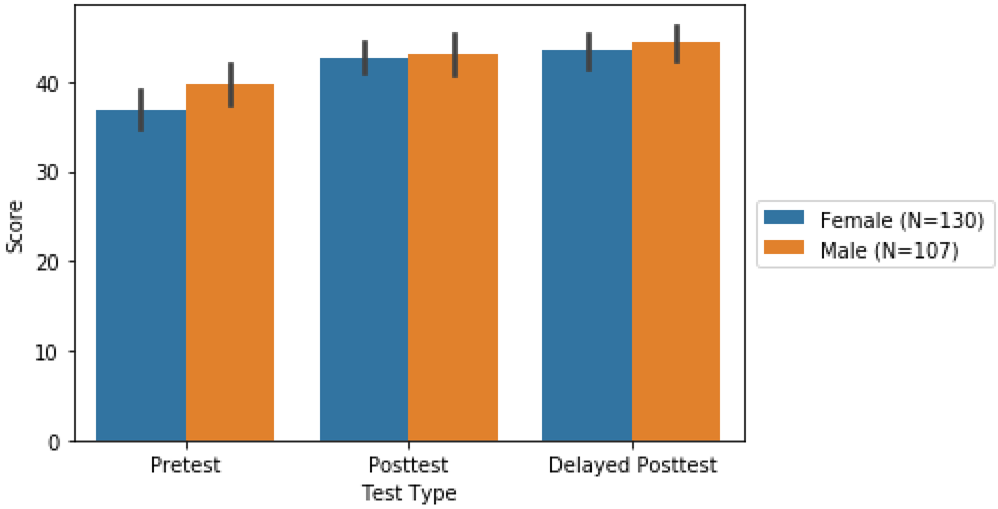


Figure 4.4: The performance of boys and girls in each test in Spring 2018. Error bars denote the 95% confidence interval around the mean.

Table 4.2: Comparison of test performance and learning gains by gender at each transfer level.

Category	**Transfer**	**Male M (*SD*)**	**Female M (*SD*)**	**Statistical result**
Pretest score	Near (*)	14.561 (4.717)	12.831 (4.863)	$F(1, 235) = 7.632$, $p = .006$, $\eta_p^2 =.031$
	Middle	5.299 (1.889)	5.008 (2.021)	$F(1, 235) = 1.293$, $p = .257$, $\eta_p^2 =.005$
	Far	13.738 (5.370)	13.215 (5.294)	$F(1, 235) = 0.565$, $p = .453$, $\eta_p^2 =.002$
Learning gains	Near (*)	2.364 (3.859)	3.615 (3.771)	$F(1, 235) = 6.322$, $p = .013$, $\eta_p^2 =.026$
	Middle (*)	-0.121 (1.821)	0.469 (1.653)	$F(1, 235) = 6.839$, $p = .009$, $\eta_p^2 =.028$
	Far	1.168 (3.374)	1.477 (3.346)	$F(1, 235) = 0.496$, $p = .482$, $\eta_p^2 =.002$
Delayed learning gains	Near (†)	2.860 (3.930)	3.877 (4.137)	$F(1, 235) = 3.711$, $p = .055$, $\eta_p^2 =.016$
	Middle	0.252 (1.828)	0.615 (1.668)	$F(1, 235) = 2.550$, $p = .112$, $\eta_p^2 = .011$
	Far	1.458 (3.653)	1.923 (3.523)	$F(1, 235) = 0.989$, $p = .321$, $\eta_p^2 =.004$

(†) p < .1; (*) p < .05

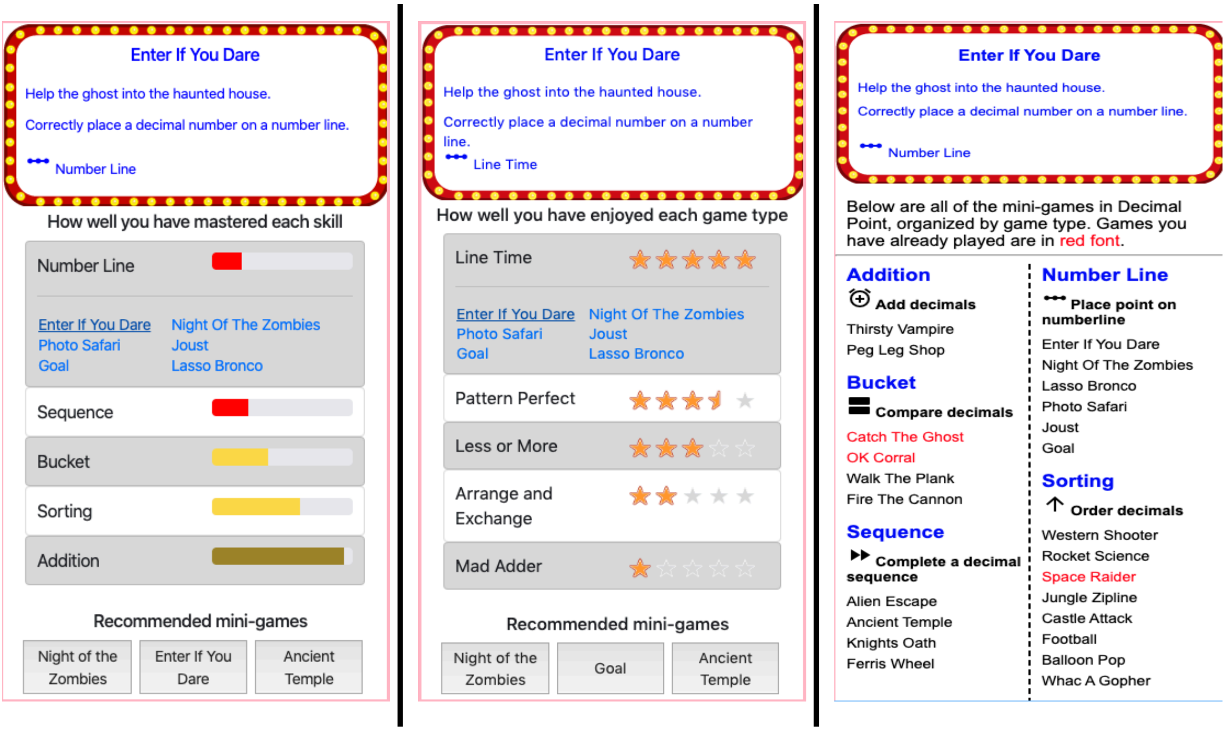
RQ2: *Is there a difference in problem-solving performance between boys and girls?*
A one-way ANOVA showed no significant gender differences in game duration per round in minutes, $F(1, 235) = 1.064$, $p = .303$, $\eta_p^2 = .007$. Similarly, there were no significant gender differences in number of game errors per round, $F(1, 235) = 0.235$, $p = .628$, $\eta_p^2 = .001$.

RQ3: *Is there a difference in self-explanation performance between boys and girls?*
A one-way ANOVA showed no significant gender differences in self-explanation duration per round in minutes, $F(1, 235) = 0.636$, $p = .426$, $\eta_p^2 = .003$. However, there was a significant gender difference in self-explanation errors per round, $F(1, 235) = 11.391$, $p = .001$, $\eta_p^2 = .046$, where boys ($M = 0.692$, $SD = 0.518$) made significantly more errors than girls ($M = 0.495$, $SD = 0.381$).

RQ4: *Is there a difference in enjoyment between boys and girls?*
As previously mentioned, the post-intervention evaluation survey in this study covered two enjoyment factors: enjoyment of content and enjoyment of interface. A one-way ANOVA showed a marginally significant difference in enjoyment of content, $F(1, 235) = 3.509$, $p = .062$, $\eta_p^2 = .015$, where boys (M = 3.309, SD = 1.194) reported lower enjoyment than girls (M = 3.568, SD = 0.924). There were no significant gender differences in enjoyment of interface, $F(1, 235) = 0.561$, $p = .455$, $\eta_p^2 = .002$.

Fall 2019 Study on Learning versus Enjoyment

Reported by Hou et al., (2022), this study was designed to examine the adoption of open learner models (Bodily et al., 2018; Bull, 2020), which are commonly used in intelligent tutoring systems to promote self-regulated learning. Towards understanding whether maximizing enjoyment is helpful to learning, the study also introduced a novel concept of an open enjoyment model. In particular, the study involved a learning-oriented version and an enjoyment-oriented version of *Decimal Point.* In the learning-oriented version (Figure 4.5a), students saw an open learner model that displayed their current mastery of each of the five decimal skills in Table 2; this data was computed based on their performance on the mini-game rounds completed so far. Students were also recommended to select a mini-game corresponding to their least mastered skill to play next. In the enjoyment-oriented version (Figure 4.5b), students instead saw a dashboard that showed how much they enjoyed the mini-games associated with each decimal skill; this data was computed based on the enjoyment rating (from 1 to 5 stars) that they submitted after completing each mini-game, using an established survey format called the “fun-o-meter” (Read & MacFarlane, 2006). Students were also recommended to select a mini-game in their most enjoyed game type (i.e., one of the five game types in Table 2.1) to play next. In addition, a control condition identical to the High Agency version used in the Fall 2017 and Spring 2018 studies (Figure 4.5c).


Enter If You Dare

Help the ghost into the haunted house.

Correctly place a decimal number on a number line.

••• Number Line

How well you have mastered each skill

Number Line

Enter If You Dare
Night Of The Zombies
Photo Safari
Joust
Goal
Lasso Bronco

Sequence

Bucket

Sorting

Addition

Recommended mini-games

Night of the Zombies | Enter If You Dare | Ancient Temple

Enter If You Dare

Help the ghost into the haunted house.

Correctly place a decimal number on a number line.

••• Line Time

How well you have enjoyed each game type

Line Time ★★★★★

Enter If You Dare
Night Of The Zombies
Photo Safari
Joust
Goal
Lasso Bronco

Pattern Perfect ★★★½☆

Less or More ★★★☆☆

Arrange and Exchange ★★☆☆☆

Mad Adder ★☆☆☆☆

Recommended mini-games

Night of the Zombies | Goal | Ancient Temple

Enter If You Dare

Help the ghost into the haunted house.

Correctly place a decimal number on a number line.

••• Number Line

Below are all of the mini-games in Decimal Point, organized by game type. Games you have already played are in red font.

Addition

Add decimals

Thirsty Vampire
Peg Leg Shop

Bucket

Compare decimals

Catch The Ghost
OK Corral
Walk The Plank
Fire The Cannon

Sequence

Complete a decimal sequence

Alien Escape
Ancient Temple
Knights Oath
Ferris Wheel

Number Line

Place point on numberline

Enter If You Dare
Night Of The Zombies
Lasso Bronco
Photo Safari
Joust
Goal

Sorting

Order decimals

Western Shooter
Rocket Science
Space Raider
Jungle Zipline
Castle Attack
Football
Balloon Pop
Whac A Gopher

(a) (b) (c)

Figure 4.5: The dashboards shown along the game map in the Learning (a), Enjoyment (b) and Control (c) condition. The skills in the Enjoyment condition are renamed to appear more playful, e.g., *Addition* becomes *Mad Adder*.

Results from the study indicated no differences in learning between students in the three conditions – Learning-oriented, Enjoyment-oriented, and Control – although there were differences in game play patterns, where students exposed to the learning-oriented dashboard replayed more mini-game rounds than those in the enjoyment-oriented version (Hou et al., 2020a). Additionally, while students in the Enjoyment- and Learning-oriented conditions followed the open models’ recommendations at similar rates (about 50% of the times), following the open learner model led to better in-game learning and post-game performance, while following the open enjoyment model did not (Hou et al., 2022a). The gender effects identified in this study are reported as follows.

Study Data

196 students from two public schools participated in the study. Among them, 35 students were removed from analyses due to not finishing all of the tests and study materials in time. Additionally, two outlier students whose learning gain scores were more than 2.5 standard deviations away from the mean were excluded. Thus, the final sample includes 159 students (82 boys, 77 girls), with a mean age of 10.93 (*SD* = 0.64).

Gender Comparisons

RQ1: *Is there a difference in learning outcomes between boys and girls?*
Figure 9 shows the test score comparison by gender. On average, girls had lower test scores than boys in the pretest, but slightly higher scores in the posttest and delayed posttest. To examine which part of the tests allowed girls to catch up with boys, Table 5 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between boys and girls at each transfer level. Boys performed marginally better than girls on the near transfer level of the pretest, but girls demonstrated significantly larger learning gains on the near- and middle-level items on the immediate posttest and near-level items on the delayed posttest.

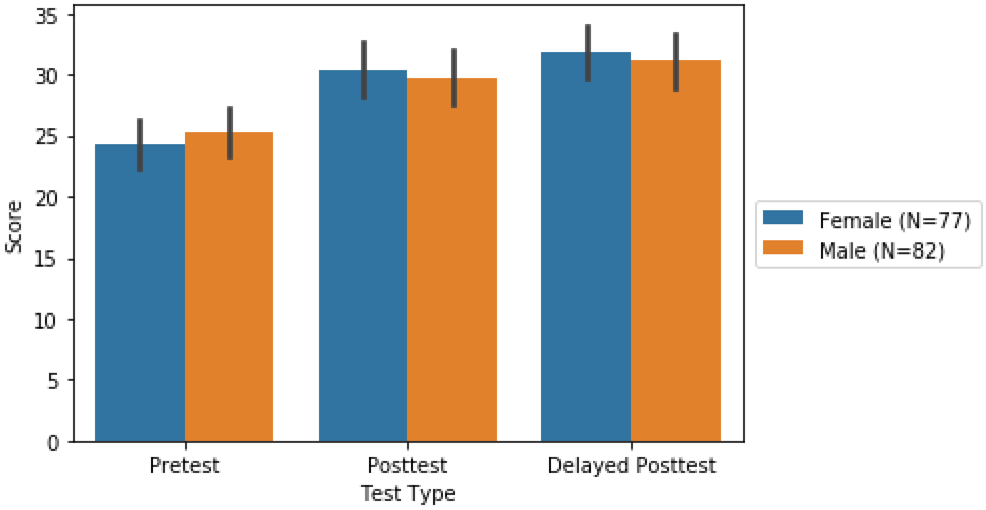


Figure 4.6: The performance of boys and girls in each test in Fall 2019. Error bars denote the 95% confidence interval around the mean.

Table 4.3: Comparison of test performance and learning gains by gender at each transfer level.

Category	**Transfer**	**Male M (*SD*)**	**Female M (*SD*)**	**Statistical result**

Pretest score	Near (†)	12.049 (4.693)	10.649 (4.542)	$F(1, 157) = 3.643$, $p = .058$, $\eta_p^2 = .023$
	Middle (*)	3.500 (2.074)	2.831 (1.902)	$F(1, 157) = 4.474$, $p = .036$, $\eta_p^2 = .028$
	Far	9.793 (4.786)	10.740 (4.747)	$F(1, 157) = 1.569$, $p = .212$, $\eta_p^2 = .010$
Learning gains	Near (*)	2.354 (3.368)	3.419 (3.530)	$F(1, 157) = 4.541$, $p = .035$, $\eta_p^2 = .028$
	Middle (*)	0.280 (2.405)	1.065 (2.142)	$F(1, 157) = 4.695$, $p = .032$, $\eta_p^2 = .029$
	Far	1.683 (2.893)	1.636 (3.967)	$F(1, 157) = 0.007$, $p = .932$, $\eta_p^2 < .001$
Delayed learning gains	Near (*)	3.061 (2.954)	4.091 (3.514)	$F(1, 157) = 4.020$, $p = .047$, $\eta_p^2 = .025$
	Middle	0.232 (2.593)	0.883 (2.606)	$F(1, 157) = 2.495$, $p = .116$, $\eta_p^2 = .016$
	Far	2.488 (3.639)	2.714 (3.821)	$F(1, 157) = 0.147$, $p = .702$, $\eta_p^2 = .001$

(†) p < .1; (*) p < .05.

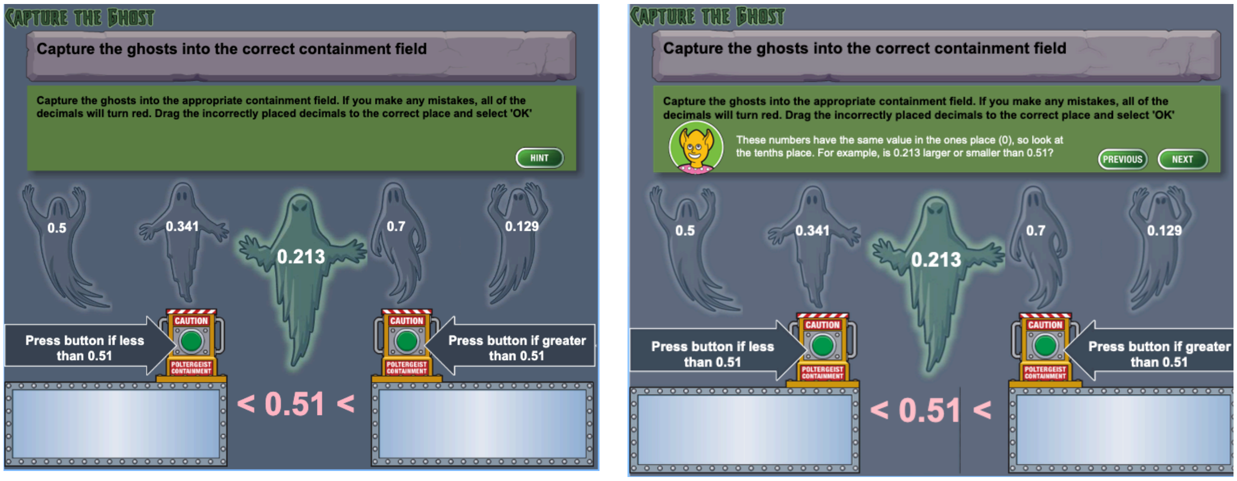
RQ2: *Is there a difference in problem-solving performance between boys and girls?*
A one-way ANOVA showed no significant gender difference in game duration per round in minutes, $F(1, 157) = 1.215$, $p = .272$, $\eta_p^2 = .019$. There were likewise no significant differences in average game errors per round, $F(1, 157) = 0.148$, $p = 0.701$, $\eta_p^2 = .001$.

RQ3: *Is there a difference in self-explanation performance between boys and girls?*
A one-way ANOVA showed a significant gender difference in self-explanation duration per round in minutes, $F(1, 157) = 14.355$, $p < .001$, $\eta_p^2 = .084$, where boys ($M = 0.369$, $SD = 0.109$) spent less time on self-explanation questions than girls ($M = 0.449$, $SD = 0.153$). There was also a significant gender difference in self-explanation errors per round, $F(1, 157) = 8.204$, $p = .005$, $\eta_p^2 = .050$, with boys ($M = 0.868$, $SD = 0.397$) making more errors than girls ($M = 0.681$, $SD = 0.428$).

RQ4: *Is there a difference in enjoyment between boys and girls?*
As previously mentioned, the post-intervention evaluation survey in this study covered three enjoyment factors: multidimensional engagement (6 items), game engagement (5 items) and achievement emotions (6 items). A series of one-way ANOVA showed no significant gender difference in multimendional engagement, $F(1, 157) = 0.110$, $p = .740$, $\eta_p^2 = .001$, game engagement, F(1, 157) = 2.073, p = .152, $\eta_p^2 = .013$, or achievement emotions, $F(1, 157) = 0.224$, $p = .637$, $\eta_p^2 = .001$.

Spring 2020 Study on Hints and Error Messages

Reported by McLaren et al., (2022c), this study was conducted to measure the effect of on-demand hints and error messages on students' learning and enjoyment outcomes in *Decimal Point*. The goal of the study was to shed light on whether instructional support mechanisms – such as hints and error messages – could be helpful in a digital learning game environment, given that these mechanisms have supported student learning in other platforms (VanLehn, 2006, 2016; Xu et al., 2019) but may disrupt the flow (Csikszentmihalyi, 1990) and engagement (Bouvier et al., 2013) of game play. In addition, prior research on hint provision in digital learning games has reported mixed results about its benefits (Drey et al., 2020; Easterday et al., 2017; O'Rourke et al., 2014), while more often focusing on how hints are perceived and used than whether they lead to learning (Conati et al., 2013; Lee & Chiou, 2020; Melero et al., 2012). To address this gap, I added on-demand hints to *Decimal Point* via a hint request button that is available at all times during the problem-solving stage (Figure 4.7a). After clicking on this button, the student can go forward and backward through three levels of hints with the Previous and Next buttons (Figure 4.7b). The hints at higher levels go into more detailed suggestions on how to solve the problem, with level 3 being the bottom-out hint that explicitly reveals the answer – this is a common hint pattern in educational technology and tutoring systems (Aleven et al., 2016). In addition, I implemented error messages that appear whenever the student makes a common decimal error, such as forgetting to carry across the decimal point (Figure 4.8). In this study, the game version with these hints and error messages was compared to the base game, where the only instructional feedback was whether the student's answer was correct or incorrect.


CAPTURE THE GHOST
Capture the ghosts into the correct containment field
HINT
0.5
0.341
0.213
0.7
0.129
CAUTION
POLTERGEIST CONTAINMENT
Press button if less than 0.51
Press button if greater than 0.51
< 0.51 <
PREVIOUS
NEXT

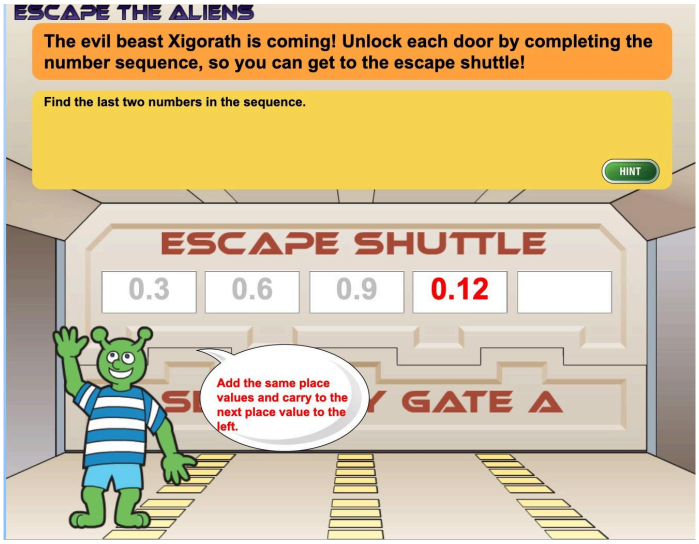
Capture the ghosts into the appropriate containment field. If you make any mistakes, all of the decimals will turn red. Drag the incorrectly placed decimals to the correct place and select 'OK'

These numbers have the same value in the ones place (0), so look at the tenths place. For example, is 0.213 larger or smaller than 0.51?

(a)

(b)

Figure 4.7: The hint request button (a) and a second-level hint message (b), which reads "*These numbers have the same value in the ones place (0), so look at the tenths place. For example, is 0.213 larger or smaller than 0.51?*"



The evil beast Xigorath is coming! Unlock each door by completing the number sequence, so you can get to the escape shuttle!

Find the last two numbers in the sequence.

Add the same place values and carry to the next place value to the left.

Figure 4.8: An example error message targeting a common decimal addition error.

Due to the COVID pandemic which started in the spring of 2020, this study was conducted in-person at two schools and remotely at three schools. As reported in McLaren et al. (2022), this difference in instructional context has yielded significant differences in completion rates, learning outcomes, and hint usage. In particular, in-class students completed the materials at a significantly higher rate than remote students. Furthermore, students learned better without hints and error messages and girls learned more than boys in-class, but these differences were not present in the remote setting. These results point to how the instructional context can have a large influence on the effectiveness of educational technology. In light of this insight, when examining the gender effects, I also investigate the in-person sample and remote sample separately.

Study Data

Two middle schools with 170 students in total participated in the in-class portion of this study. 17 students were excluded from the analysis because they did not complete all of the materials and tests in time. Thus, the final classroom sample includes 151 students (78 boys, 73 girls), with a mean age of 11.06 (*SD* = 0.86).

Three middle schools with 223 students in total participated in the remote portion of this study. 96 students were excluded from the analysis for failing to complete all the materials within the study duration. Additionally, two outlier students were excluded because their learning gains were more than 2.5 standard deviations away from the mean. Thus, the final remote sample includes 124 students (61 boys, 63 girls), with a mean age of 11.79 (*SD* = 0.56)[3].

Gender Comparisons

RQ1: *Is there a difference in learning outcomes between boys and girls?*

[3] This final sample size is different from the size reported in McLaren et al. (2022a), where two outlier students were not removed.

For the in-class study, Figure 4.9 shows the test score comparison by gender. On average, girls had lower scores than boys in the pretest, but higher scores in the posttest and delayed posttest. To examine which part of the tests led to girls' better performance, Table 4.4 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between boys and girls at each transfer level. There were no statistically significant gender differences in pretest performance, but girls had significantly higher learning gains and delayed learning gains than boys at both the near- and far-transfer level.

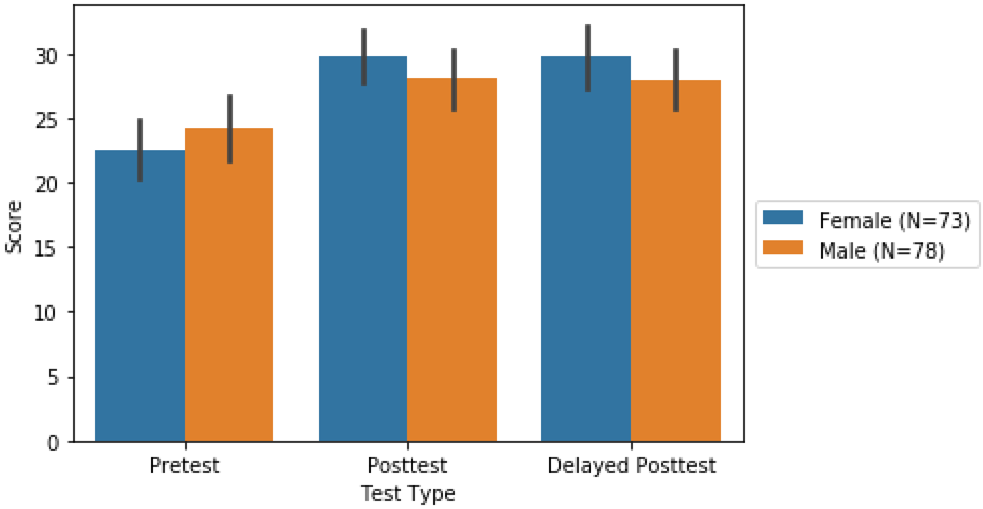


Figure 4.9: The performance of boys and girls in each test in the Spring 2020 classroom study. Error bars denote the 95% confidence interval around the mean.

Table 4.4: Comparison of test performance and learning gains by gender at each transfer level.

Category	**Transfer**	**Male M (*SD*)**	**Female M (*SD*)**	**Statistical result**
Pretest score	Near (*†)*	11.101 (5.261)	9.595 (5.835)	F(1, 151) = 2.820, p = .095, η_p^2 = .010
	Middle	2.392 (1.822)	2.595 (1.930)	F(1, 151) = 0.444, p = .506, η_p^2 = .003
	Far	10.595 (4.963)	10.189 (4.381)	F(1, 151) = 0.286, p = .594, η_p^2 = .002
Learning gains	Near (*)	2.646 (3.697)	4.622 (4.437)	F(1, 151) = 8.999, p = .003, η_p^2 = .056
	Middle	0.873 (2.047)	0.635 (2.475)	F(1, 151) = 0.423, p = .516, η_p^2 = .003
	Far (*)	0.190 (3.134)	2.230 (3.394)	F(1, 151) = 14.937, p < .001, η_p^2 = .090
Delayed learning gains	Near (*)	2.797 (3.891)	4.378 (4.631)	F(1, 151) = 5.251, p = .023, η_p^2 = .034
	Middle	0.392 (2.180)	0.784 (2.484)	F(1, 151) = 1.076, p = .301, η_p^2 = .007
	Far (*)	0.443 (3.177)	2.324 (3.928)	F(1, 151) = 10.670, p = .001, η_p^2 = .066

(*) p < .05; (†) p < .1

For the remote study, Figure 4.10 shows the test score comparison by gender. On average, girls performed slightly better than boys at all of the pretest, posttest and delayed posttest. Table 4.5 further breaks down this result by transfer level. Girls had higher learning gains and delayed learning gains than boys at the near-transfer level, but not the mid- or far-transfer level. Notably, there were no significant gender differences in any learning outcome measure.

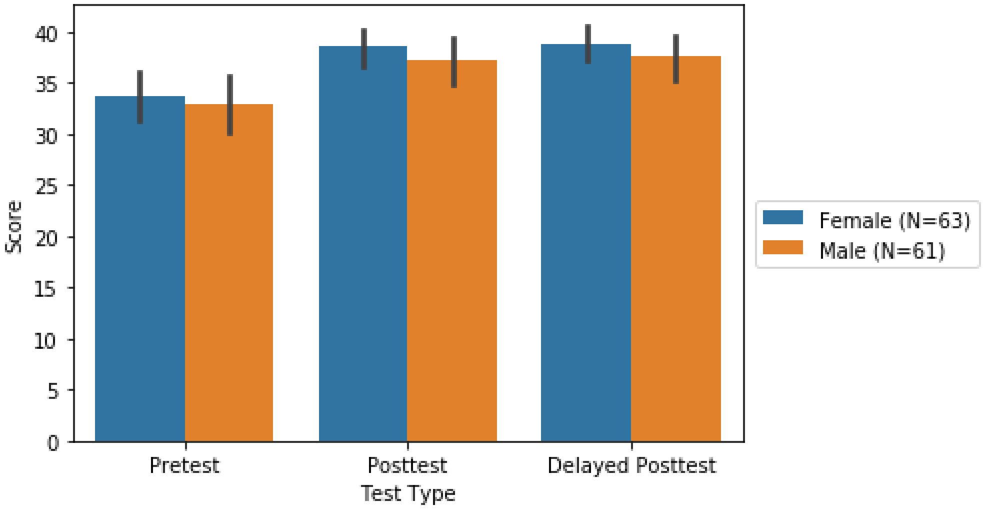


Figure 4.10: The performance of boys and girls in each test in the Spring 2020 remote study. Error bars denote the 95% confidence interval around the mean.

Table 4.5: Comparison of test performance and learning gains by gender at each transfer level.

Category	**Transfer**	**Male M (*SD*)**	**Female M (*SD*)**	**Statistical result**
Pretest score	Near	14.525 (5.117)	13.969 (4.598)	*F*(1, 123) = 0.409, *p* = .542, η_p^2 = .003
	Middle	4.656 (2.287)	4.703 (2.280)	*F*(1, 123) = 0.013, *p* = .908, η_p^2 < .001
	Far	13.738 (5.036)	14.906 (4.773)	*F*(1, 123) = 1.774, *p* = .185, η_p^2 = .014
Learning gains	Near	2.344 (3.903)	3.188 (4.246)	*F*(1, 123) = 1.333, *p* = .251, η_p^2 = .011
	Middle	0.410 (2.895)	0.234 (2.943)	*F*(1, 123) = 0.113, *p* = .738, η_p^2 = .001
	Far	1.475 (2.998)	1.188 (3.398)	*F*(1, 123) = 0.252, *p* = .616, η_p^2 = .002
Delayed learning gains	Near	2.508 (3.618)	3.312 (3.788)	*F*(1, 123) = 1.471, *p* = .227, η_p^2 = .012
	Middle	0.246 (3.448)	0.094 (3.095)	*F*(1, 123) = 0.068, *p* = .795, η_p^2 = .001
	Far	1.984 (3.217)	1.594 (3.467)	*F*(1, 123) = 0.424, *p* = .516, η_p^2 = .003

RQ2: *Is there a difference in problem-solving performance between boys and girls?*
For the in-class study, a one-way ANOVA showed a significant gender difference in average game duration per round in minutes, *F*(1, 151) = 5.425, *p* = .021, η_p^2 = .035, where boys (*M* = 0.947, *SD* = 0.623) spent less time than girls (*M* = 1.234, *SD* = 0.8886). There were no

significant gender differences in average game errors per round, $F(1, 151) = 0.976$, $p = 0.325$, $\eta_p^2 = .006$.

For the remote study, a one-way ANOVA showed no significant gender differences in average game duration per round in minutes, $F(1, 123) = 0.002$, $p = .966$, $\eta_p^2 < .001$. Similarly, there were no significant gender differences in average game errors per round, $F(1, 123) = 0.201$, $p = .654$, $\eta_p^2 = .002$.

RQ3: *Is there a difference in self-explanation performance between boys and girls?*
For the in-class study, a one-way ANOVA showed a significant gender difference in self-explanation duration per round in minutes, $F(1, 151) = 5.378$, $p = .022$, $\eta_p^2 = .034$, where boys ($M = 0.345$, $SD = 0.098$) spent less time on self-explanation questions than girls ($M = 0.381$, $SD = 0.094$). There was also a significant gender difference in self-explanation errors per round, $F(1, 151) = 6.086$, $p = .015$, $\eta_p^2 = .039$, with boys ($M = 0.897$, $SD = 0.425$) making more errors than girls ($M = 0.729$, $SD = 0.414$).

For the remote study, a one-way ANOVA showed a significant gender difference in self-explanation duration per round in minutes, $F(1, 123) = 3.978$, $p = .048$, $\eta_p^2 = .031$, where boys ($M = 0.481$, $SD = 0.210$) spent less time on self-explanation questions than girls ($M = 0.570$, $SD = 0.282$). There was also a significant gender difference in self-explanation errors per round, $F(1, 123) = 7.744$, $p = .006$, $\eta_p^2 = .059$, with boys ($M = 0.707$, $SD = 0.461$) making more errors than girls ($M = 0.496$, $SD = 0.384$).

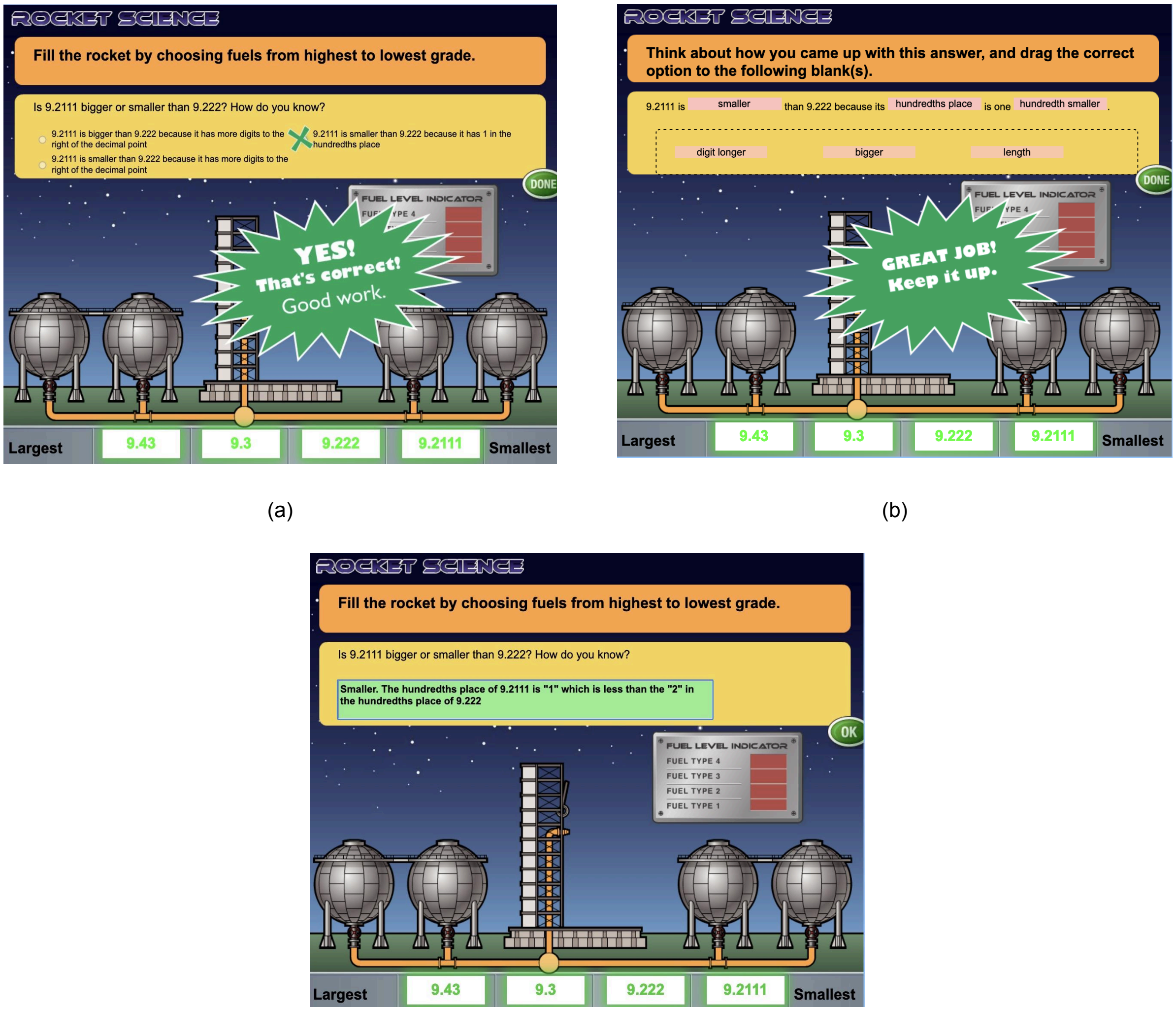
RQ4: *Is there a difference in enjoyment between boys and girls?*
As previously mentioned, the evaluation survey for this study involved three enjoyment factors: multidimensional engagement, achievement emotions, and player experience. For the in-class study, a series of one-way ANOVA showed no significant gender differences in multidimensional engagement, $F(1, 151) = 1.691$, $p = .195$, $\eta_p^2 = .011$, or in achievement emotions, $F(1, 151) = 1.808$, $p = .181$, $\eta_p^2 = .012$. However, there was a significant difference in player experience, $F(1, 151) = 5.558$, $p = .02$, $\eta_p^2 = .036$, where boys ($M = 3.473$, $SD = 0.694$) reported higher levels of enjoyment than girls ($M = 3.206$, $SD = 0.706$).

For the remote study, there were no significant gender differences in multidimensional engagement, $F(1, 123) = 2.674$, $p = .105$, $\eta_p^2 = .021$, achievement emotions, $F(1, 123) = 1.940$, $p = .166$, $\eta_p^2 = .016$, or player experience, $F(1, 123) = 0.792$, $p = .375$, $\eta_p^2 = .006$.

Spring 2021 Study on Types of Prompted Self-Explanation

Reported by McLaren et al. (2022a; 2022b), This study was conducted to measure the impact of different types of prompted self-explanation on students' learning outcomes and enjoyment. The study was motivated by whether less constrained self-explanations, such as open-ended responses, could promote learning in a digital game context, given that they facilitate active and constructive engagement that induces robust knowledge (Wylie & Chi, 2014) but may also induce extraneous cognitive load (D. M. Adams & Clark, 2014; Killingsworth et al., 2015).

Indeed, prompted self-explanation has been shown to support learning within games (Hsu & Tsai, 2011), although this effect has not been seen in all studies with learning games (D. M. Adams & Clark, 2014). To further examine this area, I and members of the McLearn Lab developed three versions of the self-explanation prompt in *Decimal Point*. The menu-based explanations (Figure 4.11a) involve multiple-choice questions with 3 or 4 options; this format has also been used in all prior *Decimal Point* studies. The scaffolded self-explanations (Figure 4.11b) prompted students to fill in the blanks using a given word bank of 4 or 5 possible phrases. Finally, with focused self-explanations (Figure 4.11c), students were tasked with typing their responses into an open-ended text box. To assure that students would expend at least minimal effort in self-explaining, the game required that their self-explanations contain at least four words, including at least one of the keywords from a relevant list (including common misspellings) that would legitimately be found in a correct explanation. This approach aligns with the focused self-explanations described in Wylie & Chi (2014) and is minimally constrained.



Fill the rocket by choosing fuels from highest to lowest grade.

Is 9.2111 bigger or smaller than 9.222? How do you know?

○ 9.2111 is bigger than 9.222 because it has more digits to the right of the decimal point

☒ 9.2111 is smaller than 9.222 because it has 1 in the hundredths place

○ 9.2111 is smaller than 9.222 because it has more digits to the right of the decimal point

(a)

Think about how you came up with this answer, and drag the correct option to the following blank(s).

9.2111 is smaller than 9.222 because its hundredths place is one hundredth smaller.

digit longer bigger length

(b)

Fill the rocket by choosing fuels from highest to lowest grade.

Is 9.2111 bigger or smaller than 9.222? How do you know?

Smaller. The hundredths place of 9.2111 is "1" which is less than the "2" in the hundredths place of 9.222

(c)

Figure 4.11: The different types of self-explanation prompts in the mini-game Rocket Science: (a) Menu-based, (b) Scaffolded, (c) Focused.

Results from this study indicated that focused self-explanations led to significantly higher delayed posttest performance and sense of mastery than menu-based self-explanations (McLaren et al., 2022b). There were no other differences in learning or enjoyment measures between the three conditions. Thus, these findings support the view of Chi and Wylie (2014) that constructive prompts are more beneficial for retention and deeper knowledge. Furthermore, the benefits of enjoyment often claimed for digital learning games do not appear to have been undercut with less constrained self-explanations, as there were no differences in enjoyment across the three conditions. In the analyses reported below, I further focus on how boys and girls perform at each type of prompted self-explanation.

Study Data

357 students from four middle schools participated in this study. Among them, 143 were dropped due to (a) failing to complete part of the learning materials or any tests or (b) having participated in a study with similar materials the previous year at one of the schools. Note that the relatively high attrition rate was due, at least in part, to running the study during the COVID-19 pandemic. Some students participated in person, some at home, and some in a hybrid format. Additionally, 6 students were excluded as outliers, due to having learning gains that are 2.5 standard deviations away from the mean. The remaining 208 students, with 97 boys and 111 girls, had a mean age of 11.58 (*SD* = 0.58)[4].

Gender Comparisons

RQ1: *Is there a difference in learning outcomes between boys and girls?*
Figure 4.12 shows the test score comparison by gender. On average, girls had lower test scores than boys in the pretest, but had similar posttest scores and slightly higher delayed posttest scores. To examine which part of the tests allowed girls to catch up with boys, Table 4.6 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between boys and girls at each transfer level. Male students performed marginally better than girls on the near transfer level of the pretest, but girls demonstrated significantly larger learning gains on the near- and middle-level items on the immediate posttest and near-level items on the delayed posttest.

[4] This final sample size is different from the size reported in McLaren et al. (2022b), where 6 outlier students were not removed.

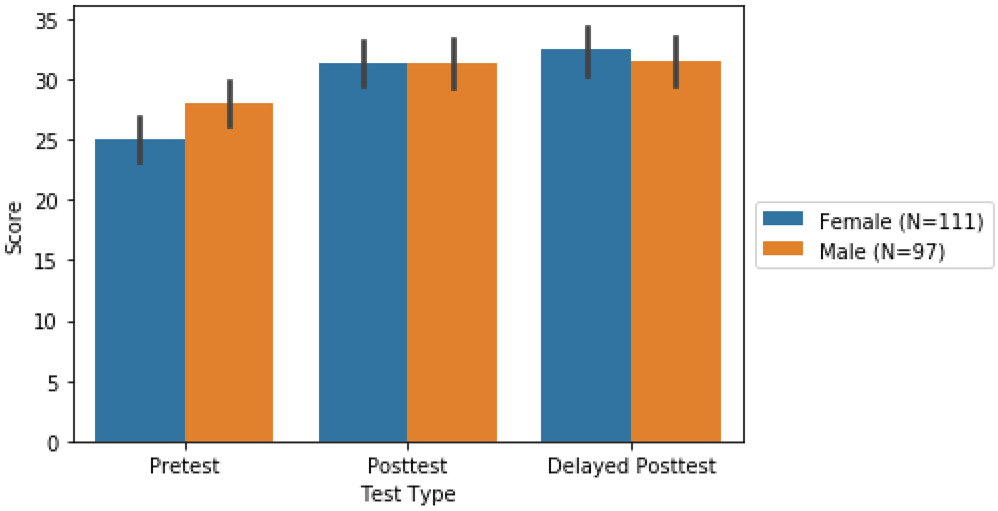


Figure 4.12: The test performance of boys and girls in each test type in Spring 2021. Error bars denote the 95% confidence interval around the mean.

Table 4.6: Comparison of test performance and learning gains by gender at each transfer level.

Category	**Transfer**	**Male M (*SD*)**	**Female M (*SD*)**	**Statistical result**
Pretest score	Near (*)	13.381 (4.700)	10.946 (5.045)	$F(1, 206) = 12.855$, $p < .001$, $\eta_p^2 = .059$
	Middle (*)	3.969 (2.234)	3.297 (2.243)	$F(1, 206) = 4.663$, $p = .032$, $\eta_p^2 = .022$
	Far	10.619 (5.159)	10.838 (4.902)	$F(1, 206) = 0.099$, $p = .754$, $\eta_p^2 < .001$
Learning gains	Near (*)	1.918 (3.108)	4.027 (3.748)	$F(1, 206) = 19.194$, $p < .001$, $\eta_p^2 = .080$
	Middle	0.041 (2.669)	0.360 (2.795)	$F(1, 206) = 0.704$, $p = .403$, $\eta_p^2 = .003$
	Far	1.402 (3.567)	1.955 (3.657)	$F(1, 206) = 1.211$, $p = .272$, $\eta_p^2 = .006$
Delayed learning gains	Near (*)	1.938 (3.204)	4.207 (3.710)	$F(1, 206) = 21.962$, $p < .001$, $\eta_p^2 = .096$
	Middle (*)	-0.021 (2.865)	0.919 (2.530)	$F(1, 206) = 6.308$, $p = .013$, $\eta_p^2 = .030$
	Far	1.629 (3.355)	2.225 (4.085)	$F(1, 206) = 1.301$, $p = .255$, $\eta_p^2 = .006$

(*) p < .05.

RQ2: *Is there a difference in problem-solving performance between boys and girls?*
A one-way ANOVA showed a significant gender difference in average game duration per round in minutes, $F(1, 206) = 9.598$, $p = .002$, $\eta_p^2 = .045$, where boys ($M = 1.025$, $SD = 0.556$) spent significantly less time than girls ($M = 1.291$, $SD = 0.668$). At the same time, there were no significant differences in average game errors per round, $F(1, 206) = 2.201$, $p = .139$, $\eta_p^2 = .011$, between boys and girls.

RQ3: *Is there a difference in self-explanation performance between boys and girls?*

Because students had different interactions with each type of prompted self-explanation, the duration and number of errors in each condition are measured on different scales. Generally, students needed more time to type an open-ended response (focused condition) than to fill in the blank (scaffolded condition) or complete a multiple-choice question (menu-based condition). Furthermore, in the menu-based condition, there are 3 or 4 multiple-choice options, so students can make at most 2 or 3 errors per mini-game. In the scaffolded condition, there are several blanks to fill in so students can make a much larger number of errors per mini-game. Finally, in the focused condition, the student responses were manually graded in a post-hoc manner, so students could make at most 1 error per mini-game. To account for these differences, I examined the gender effect separately in each condition.

In the menu-based condition (36 boys, 38 girls), there were no significant gender differences in average self-explanation duration, $F(1, 72) = 0.014$, $p = .907$, $\eta_p^2 < .001$, or in self-explanation errors, $F(1, 72) = 2.575$, $p = .113$, $\eta_p^2 = .035$. Similarly, the scaffolded condition (26 boys, 42 girls) did not yield significant gender differences in average self-explanation duration, $F(1, 66) = 0.125$, $p = .725$, $\eta_p^2 = .002$, or in self-explanation errors, $F(1, 66) = 0.136$, $p = .714$, $\eta_p^2 = .002$. In the focused condition (35 boys, 31 girls), boys and girls did not differ significantly in time spent on self-explanation, $F(1, 64) = 0.649$, $p = .423$, $\eta_p^2 = .010$, but boys ($M = 0.731$, $SD = 0.210$) made significantly more self-explanation errors per round than girls ($M = 0.610$, $SD = 0.222$), $F(1, 64) = 5.143$, $p = .027$, $\eta_p^2 = .074$.

RQ4: *Is there a difference in enjoyment between boys and girls?*
As previously mentioned, the evaluation survey for this study involved three enjoyment factors: multidimensional engagement, achievement emotions, and player experience. A one-way ANOVA showed no significant gender differences in multidimensional engagement, $F(1, 206) = 1.664$, $p = .198$, $\eta_p^2 = .008$, player experience, $F(1, 206) = 0.027$, $p = .869$, $\eta_p^2 < .001$, or achievement emotions, $F(1, 206) = 2.601$, $p = .108$, $\eta_p^2 = .012$.

Result Summary and Post-hoc Analyses

The above analyses across the five prior studies of *Decimal Point* have demonstrated consistent gender differences in playing and learning from the game, which are summarized in Table 4.7 [5]. They also answer the research questions I have posed as follows.

- For **RQ1** – whether boys and girls had different learning outcomes – boys tended to outperform girls at pretest, but girls often had higher learning gains and delayed learning gains. This pattern is especially consistent at the near transfer level, with the most frequent occurrences of significant gender differences.
- For **RQ2** – whether boys and girls had different problem-solving performance – girls consistently spent more time in the problem-solving activities than boys, while both groups made a similar number of errors.

[5] The Spring 2020 in-class and remote studies are considered separately due to their distinct instructional contexts.

- For **RQ3** – whether boys and girls had different self-explanation performance – there was a highly consistent trend where girls took more time and made significantly fewer errors than boys.
- Finally, for **RQ4** – whether boys and girls reported different levels of enjoyment with the game – while different enjoyment categories were surveyed in each study, boys and girls mostly reported similar levels of enjoyment.

Table 4.7: Summary of gender comparison across studies. The value in each cell indicates which gender had higher outcomes in the corresponding category (M for male and F for female). Light yellow cells denote small effect sizes ($\eta_p^2 < .06$) and orange cells denote medium effect sizes ($.06 \le \eta_p^2 < .14$).

Category	F17 (n = 158)	S18 (n = 237)	F19 (n = 159)	S20 In-class (n = 153)	S20 Remote (n = 125)	S21 (n = 208)
Learning						
Pretest - Near	M (*)	M (*)	M	M	M	M (*)
Pretest - Mid	M	M	M (*)	F	F	M (*)
Pretest - Far	M	M	F	M	F	F
Learning Gains - Near	F	F (*)	F (*)	F (*)	F	F (*)
Learning Gains - Mid	M	F (*)	F (*)	M	M	F
Learning Gains - Far	F (*)	F	M	F (*)	M	F
Delayed Gains - Near	F	F (*)	F (*)	F (*)	F	F (*)
Delayed Gains - Mid	M (†)	F	F	F	M	F (*)
Delayed Gains - Far	F (*)	F	F	F (*)	M	F
Problem-solving activities						
Problem-solving Duration	F (*)	F	F	F (*)	F	F (*)
Problem-solving Errors	F (*)	F	M	F	M	F
Prompted Self-explanation activities						
Menu-based SE Duration	F	F	F (*)	F (*)	F (*)	F
Menu-based SE Errors	M (*)	M (*)	M (*)	M (*)	M (*)	M
Scaffolded SE Duration	-	-	-	-	-	F
Scaffolded SE Errors	-	-	-	-	-	M
Focused SE Duration	-	-	-	-	-	M

Focused SE Error	-	-	-	-	-	M (*)
Post-intervention enjoyment ratings						
Enjoyment of content	F	F	-	-	-	-
Enjoyment of interface	M	F	-	-	-	-
Math attitude	M	-	-	-	-	-
Game engagement	-	-	F	-	-	-
Multidimensional engagement	-	-	M	M	F	F
Achievement emotions	-	-	F	M	F	M
Player experience	-	-	-	M (*)	F	F

(*) p < .05

From the above summary, I observed that the most consistent and significant gender differences across studies are in learning gains, especially at the near-transfer level, which includes test items similar to the in-game problems, and in self-explanation behaviors. This pattern led to the follow-up question: did girls learn more from the game than boys because they performed better (i.e., make fewer errors) in the prompted self-explanation activities? To test this hypothesis, I constructed two mediation models with gender as an independent variable (where male is coded as 1 and female as 0), average self-explanation errors per mini-game round as a mediator, pretest score as a covariate, and posttest / delayed posttest score as the dependent variable. The confidence interval of the indirect effect was estimated at the 0.05 significance level via bias-corrected non-parametric bootstrapping with 2000 iterations (Hayes & Rockwood, 2017; Vallat, 2018). The results of this analysis on each of the five *Decimal Point* studies are shown in Figures 4.13 - 4.20 below, where * indicates that the coefficients are significantly different from 0 at the α = 0.05 significance level.

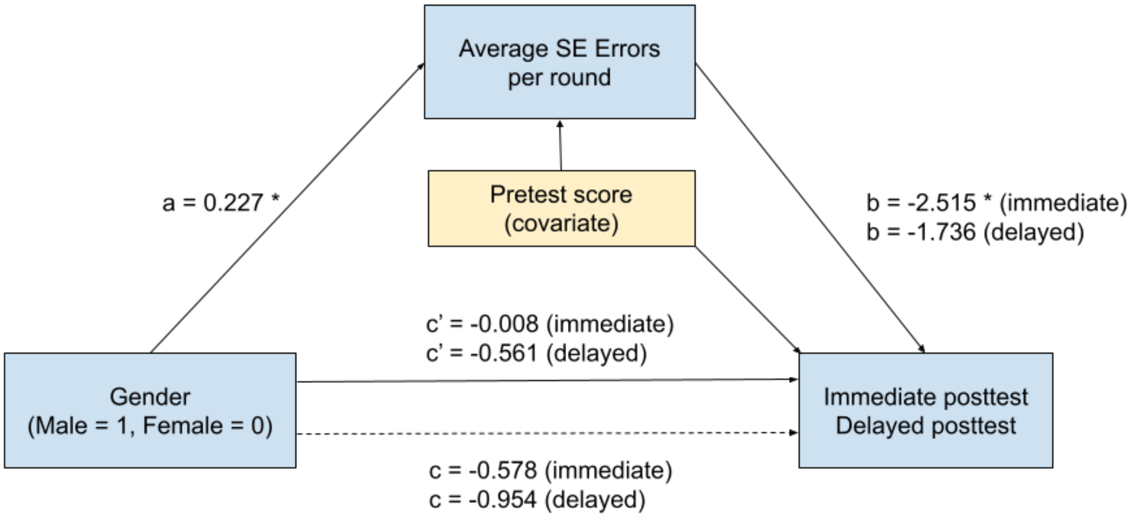


Figure 4.13: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Fall 2017 study.

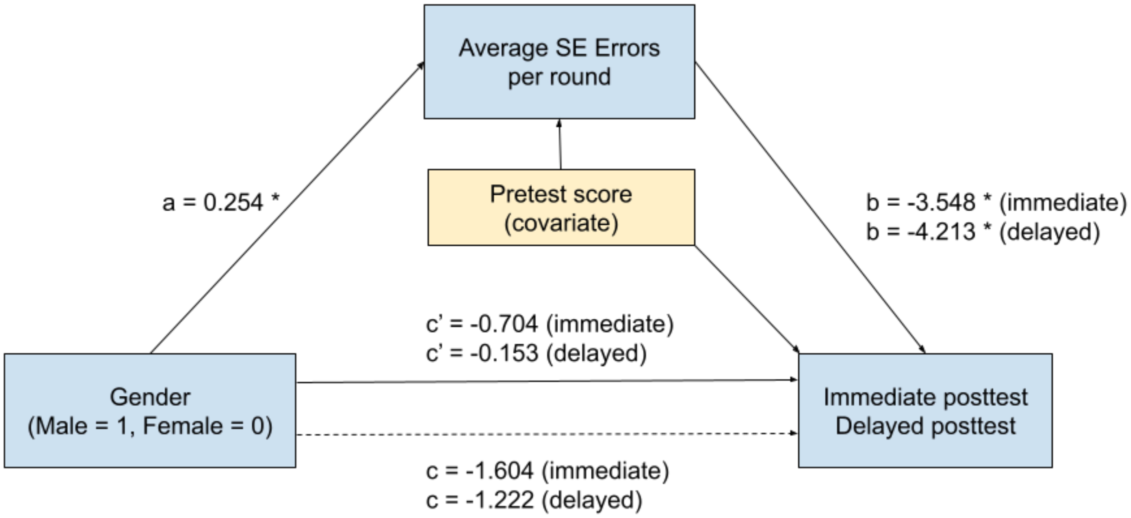


Figure 4.14: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Spring 2018 study.

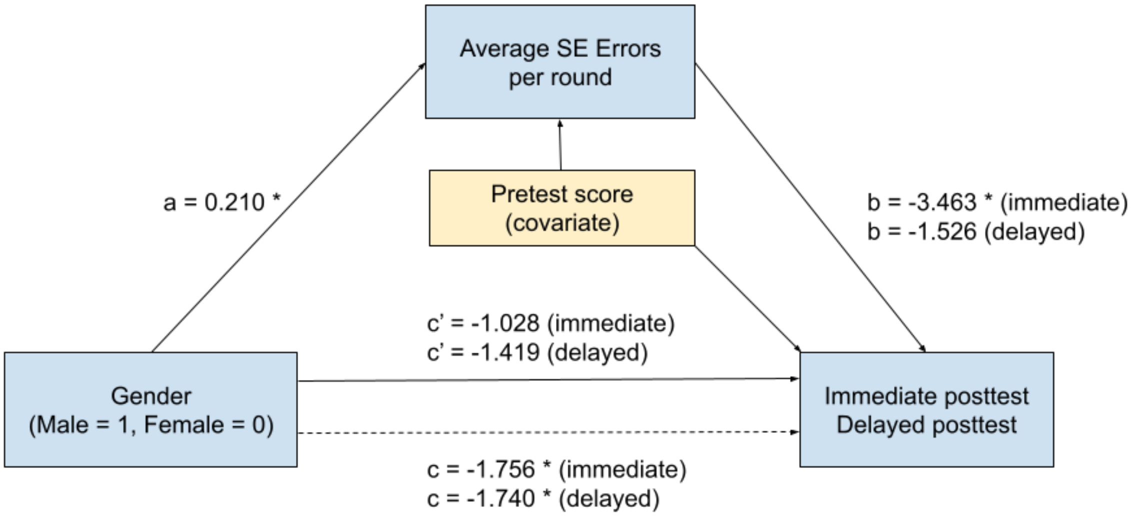


Figure 4.15: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Fall 2019 study.

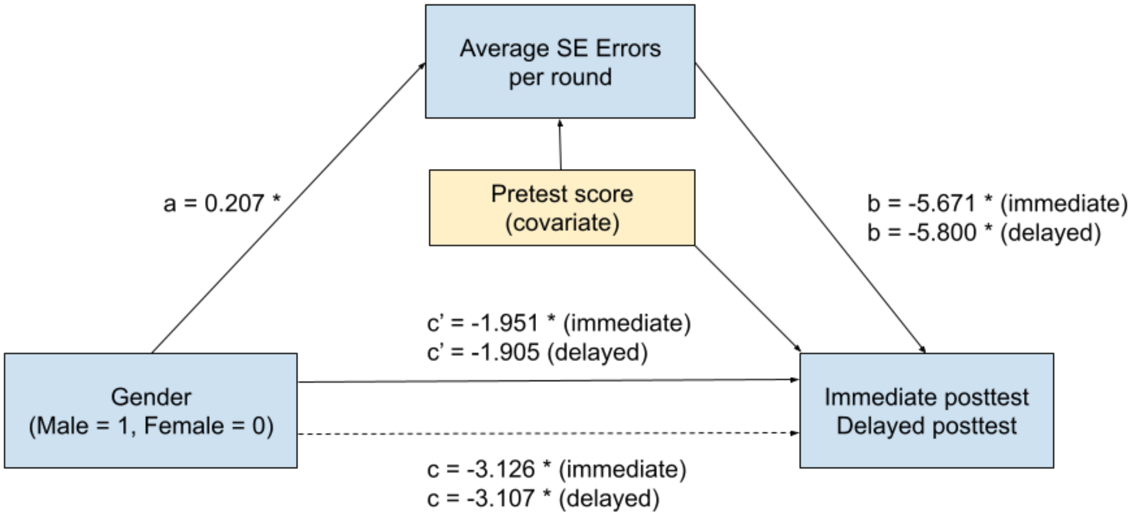

Average SE Errors per round
Pretest score (covariate)
a = 0.207 *
b = -5.671 * (immediate)
b = -5.800 * (delayed)
c' = -1.951 * (immediate)
c' = -1.905 (delayed)
Gender (Male = 1, Female = 0)
Immediate posttest Delayed posttest
c = -3.126 * (immediate)
c = -3.107 * (delayed)

Figure 4.16: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Spring 2020 in-class study.

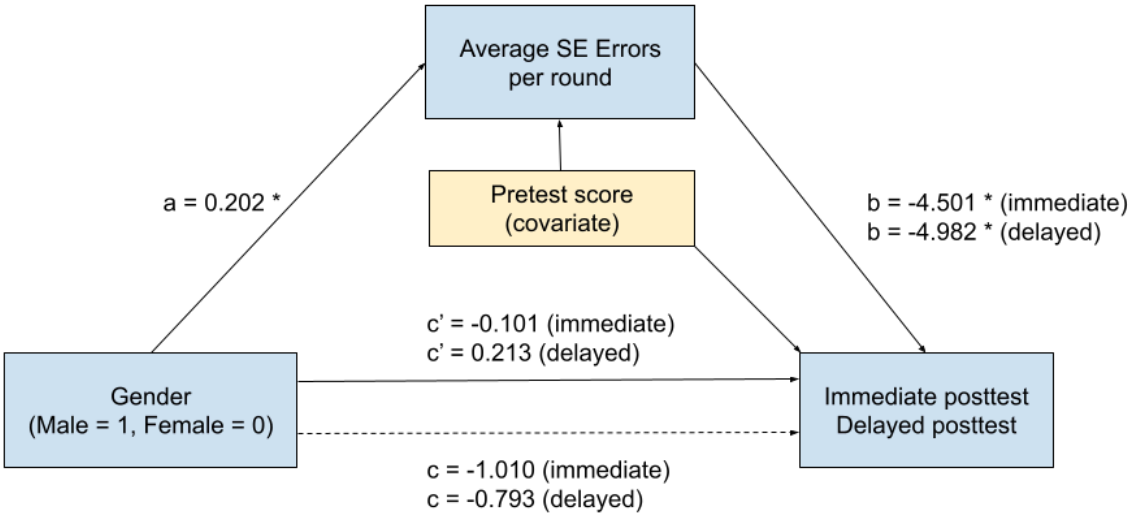


Figure 4.17: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Spring 2020 remote study.

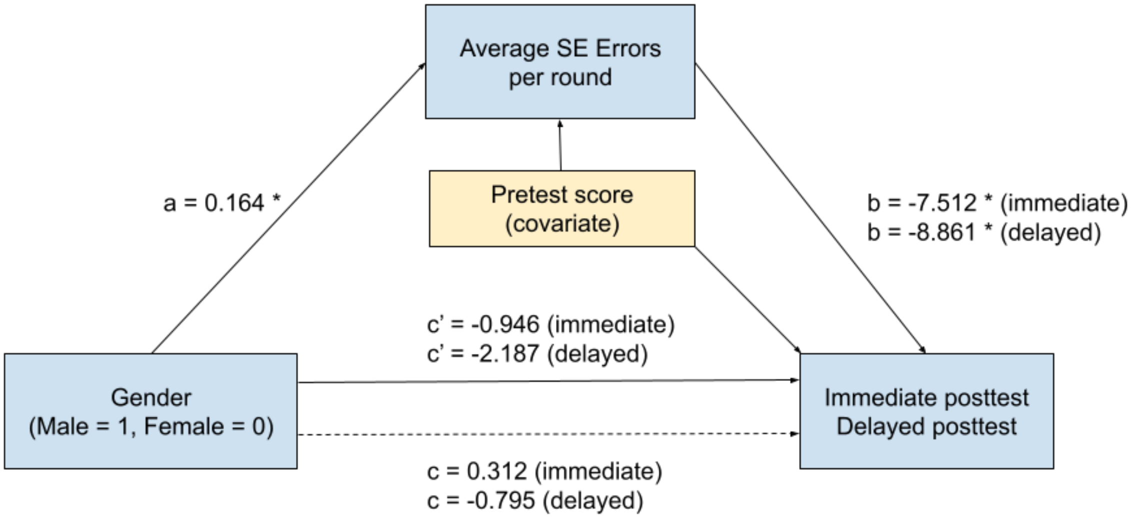


Figure 4.18: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Spring 2021 study with menu-based self-explanation activities.

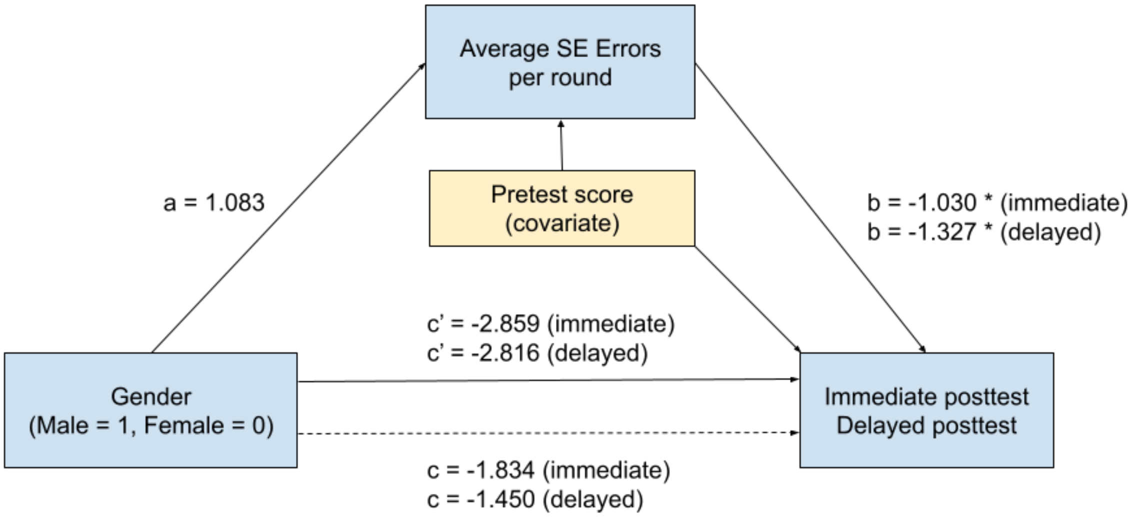


Figure 4.19: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Spring 2021 study with scaffolded self-explanation activities.

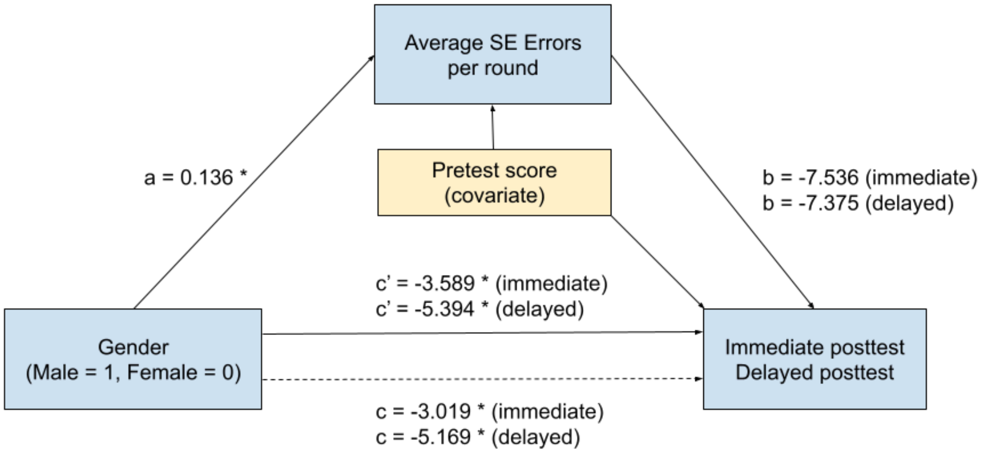


Figure 4.20: Diagram of the mediation pathway from gender to posttest and delayed posttest performance through self-explanation errors in the Spring 2021 study with focused self-explanation activities.

The following patterns emerged from the mediation models across all studies. First, when controlling for pretest scores, gender (where male is coded as 1 and female as 0) was positively associated with the average number of self-explanation errors, which was in turn negatively associated with posttest and delayed posttest performance. Second, the effect of gender on post-intervention test performance was mediated by the number of self-explanation errors. Results of bootstrapping procedures showed a significant indirect effect in six out of eight mediation models that predict posttest performance, and four out of eight mediation models that predict delayed posttest performance. Notably, the indirect effect was significant in all instances where students worked with multiple-choice self-explanation questions, including all the studies prior to Spring 2021 as well as the menu-based condition in the Spring 2021 study. Finally, the regression models predicting posttest and delayed posttest scores based on gender, pretest scores and self-explanation errors were able to explain about 65-75% of the variance in test scores across all studies, indicating that these factors are reasonably predictive of post-intervention performance.

Discussion

I have investigated the gender effects in various aspects of game play and learning across five prior studies of *Decimal Point*, with over 1,000 students in total. My analyses were motivated by why and how gender influences the process of playing and learning from digital learning games, given that there are clear gender differences in game preferences (Aleksić & Ivanović, 2017; Chou & Tsai, 2007; Greenberg et al., 2010; Romrell, 2014), yet insufficient empirical guidance on translating these differences into instructionally effective and inclusive learning game designs. This research topic is especially relevant in middle-school math domains, where there exists an established gender gap favoring boys (Arroyo et al., 2013; Breda et al., 2018; Wai et al., 2010) with long-lasting career implications (X. Huang et al., 2019). Beyond comparing boys and girls' measures of test performance, problem-solving and prompted self-explanation, I also examined mediation pathways that explain the connection between these factors. In turn, these results lay the foundation for my thesis work, which aims to expand the current analyses to a wider range of game contexts and gender dimensions.

Overall, the previous classroom studies with *Decimal Point* have identified consistent gender differences in students' learning outcomes and self-explanation behaviors. First, girls under-performed compared to boys at pretests but had higher learning gains and delayed learning gains. This result did not reach significance every year, or for every transfer level, but consistently emerged as a strong trend, especially at the near transfer level, which is closest to the game's learning content. Second, girls made fewer errors than boys on the self-explanation prompts, though not during the problem-solving portion of the game. The gender difference in self-explanation errors was significant in four out of five studies. In addition, when controlling for pretest scores, performance in the prompted self-explanation activities was found to explain the effect of gender on posttest and delayed posttest performance. While the effect sizes of the above gender comparisons range only from small to medium, based on Cohen (2013)'s general thresholds, they do match the average effect size values reported in educational interventions on individual students (Lipsey et al., 2012; Table 10), as well as those from prior surveys on gender differences in mathematics learning (Else-Quest et al., 2010; 2013) and in educational technology usage (Arroyo et al., 2013). Thus, the consistent gender differences reported are still noteworthy and could point to an important game design feature that may be leveraged in future work to support girls' learning and, in turn, bridge the gender gap in math education. These implications are further discussed below.

The observation of boys having higher pretest scores is consistent with prior literature demonstrating boys' tendency to perform better at math than girls in late elementary and early middle school (Robinson & Lubienski, 2011). However, the fact that girls had consistently higher learning gains and delayed learning gains is an important pattern. Notably, this pattern was not due to a ceiling effect, as both boys and girls' average posttest and delayed posttest scores were in the range of 30-40 (out of 52 possible points), indicating that they still had room for improvement. Rather, this result can be attributed to the game's learning benefits, which helped girls catch up with their boys in math performance after playing. In turn, this thesis work contributes to the body of research showing that digital learning games can lead to gender differences in learning outcomes that favor girls (Adamo-Villani et al., 2008; Chung & Chang, 2017; Joiner et al., 2011; Khan et al., 2017; Klisch et al., 2012). However, as other learning game studies have reported no gender differences (Chang et al., 2014; Clark et al., 2011; Dorji et al., 2015; Manero et al., 2016; Papastergiou, 2009), I also set out to explore why *Decimal Point,* in particular, was more beneficial for girls.

The first conjecture was that girls learned more because they approached the self-explanation prompts more carefully and deliberately. In contrast, boys spent less time and made significantly more errors with prompted self-explanation. As self-explanation is an established instructional technique for promoting deep learning and transfer (Chi et al., 1994; Wylie & Chi, 2014), it is not surprising that self-explanation behaviors are associated with differences in learning outcomes (Richey & Nokes-Malach, 2015). This connection is supported by a post-hoc analysis which reveals a significant mediation effect of self-explanation errors in six out of eight models predicting posttest scores and four out of eight models predicting delayed posttest scores. In other words, girls learned more because they were more careful when performing self-explanation than boys.

At the same time, girls' better self-explanation performance could be attributed to their faster development of verbal learning strategies, which would give them an advantage over boys in this type of activity (Nikolaenko, 2005; Stevenson et al., 2009). This distinction is most pronounced in the focused self-explanation condition in the Spring 2021 study, where students had to write open-ended responses that contain certain keywords (which are not revealed to them). In this setting, students could not use a trial-and-error approach, as they were able to with menu-based or scaffolded self-explanation, but had to rely on their own decimal understanding and generative language skills. Consequently, girls not only made significantly fewer errors but also spent less time than boys in the focused self-explanation activities; this was in fact the only instance, across all *Decimal Point* studies, where girls spent less time with self-explanation than boys (Table 4.7). These findings constitute a novel contribution to the existing literature examining self-explanation interventions, which typically did not report on gender differences. One prior study testing this idea with 7- to 9-year-olds found significant gender differences in learning through self-explanation, where girls performed better than boys if no feedback was provided (Stevenson et al., 2009), but more research is needed to understand whether this is a robust effect and whether it persists among older children and adults. Therefore, the findings from this thesis work raise the need to further explore the connection between gender, self-explanation behaviors and learning outcomes in future studies of *Decimal Point*, as well as learning games in general.

A second hypothesis is that learning math in a game context reduces the math saliency of the content, thus decreasing the likelihood of triggering anxiety about math performance in girls (Doyle & Voyer, 2016; Nguyen & Ryan, 2008; Picho et al., 2013; Spencer et al., 1999). By reducing girls' anxiety from stereotype threats, games may free up more working memory for learning about mathematical concepts and, as a result, allow girls to catch up to boys on the posttest despite typically attaining lower scores on the pretest (Gödöllei Lappalainen, 2017; Sitzmann, 2011). If the game affords girls a greater opportunity to correct misconceptions and build knowledge about decimal number operations than they experience with more typical instruction, this feature might explain why girls were more thoughtful and made fewer errors on self-explanation prompts. As part of my proposed work, described in the next section, I will test this hypothesis by measuring students' anxiety as a means of assessing the impact of stereotype threat.

An opposite trend was observed in the problem-solving activities in the game, where girls tended to spend more time and make more errors than boys, although not significantly so (Table 4.7). This difference can be attributed to girls' lower prior knowledge, causing them to struggle more with the learning content in the game. However, their struggles may turn out to be beneficial, as prior studies have reported that the emotions students feel while struggling, namely confusion and frustration, were positively correlated with learning outcomes (D'Mello et al., 2014; Lehman et al., 2013). Results from my analysis of prior *Decimal Point* studies did indeed show that girls were able to acquire higher learning gains after game play. When examining the role of the problem-solving activities in inducing this effect, we should note that these activities are where the game's fantasy settings and narratives emerge most strongly. For example, while playing the mini-game in Figure 3.3, students would get to interact with different objects representative of the Amusement Park theme and receive occasional feedback from their alien friends. This immersive experience could lead students to attribute any negative emotion while playing, such as anxiety and frustration, to the game environment, rather than the task content (Holmes et al., 2019). Thus, when facing similar tasks in the posttest and delayed posttest, without the surrounding game context, students – especially girls – could tackle them more comfortably than they did in the pretest.

Taken together, the gender comparisons in *Decimal Point* suggest several mechanisms through which learning games can bridge the gender gap in middle-school math education. First, girls demonstrated better learning with prompted self-explanation than boys, which could lead to their higher learning gains. Second, the informal game context could reduce the stereotype threat that girls face while studying math. Third, the immersive game themes and narratives could promote learning engagement and offset the negative emotions that students may experience during the learning process. Most notably, while these mechanisms appear to have stronger effects on girls than boys, they have the potential to benefit all students. In other words, promoting girls' math learning does not need to be at the expense of boys' learning.

Towards translating the above findings into more general learning game design guidelines, there are two key questions to address. First, how do the game environment and narrative impact the relationship between gender and learning outcomes? While prior *Decimal Point* studies have manipulated different aspects of the game, the overall narrative of traveling through an amusement park along with helpful alien friends (Figure 1.1) has remained constant. It is unclear if similarly robust gender effects can be replicated when the game presents a different narrative (e.g., hunting for treasures or fighting against a mastermind villain). Therefore, an important next step is to investigate how well the game's gender effects generalize to other narratives. Second, would additional dimensions of gender that extend upon the current binary classification provide a more nuanced understanding of how gender influences game play and learning? As recent research in gender studies has challenged the binary gender perspective and advocated for a multidimensional gender framework which is applicable even to young children (Fast & Olson, 2018; Gülgöz et al., 2019; Hyde et al., 2019; Olson & Gülgöz, 2018; Perry et al., 2019; Rae et al., 2019), there can be benefits to examining the gender effects in *Decimal Point* through these multidimensional lenses. Furthermore, given that prior work in digital learning games has exclusively employed the binary gender classification, the potential value of analyzing additional gender dimensions, as demonstrated through *Decimal Point* studies, would also be of interest to the broader research community.

In turn, these two key questions and the research directions that they inspire have helped shape the trajectory of my proposed work, which I describe in details below.

Overview of the Next Steps

Results from five prior studies of *Decimal Point* have revealed a highly consistent pattern, where girls perform better than boys at the prompted self-explanation activities, which leads to better learning after game play. As the next step, I have conducted two additional studies which expand on previous research along two axes – (1) incorporating multiple dimensions of gender and (2) examining the role of the learning environment in the identified relationships between gender, self-explanation and learning outcomes. These studies aim to uncover other types of gender differences that could contribute to the gender effects in learning. A brief summary of the study settings is as follows.

The first follow-up study was a 2x2 randomized controlled experiment, where students learn from either the game *Decimal Point* or an equivalent tutoring system, and each learning platform either contains both problem-solving and self-explanation activities, or only problem-solving activities. This study was designed to untangle the effects of the learning platform (game versus tutor) and the effects of learning with versus without self-explanation. In the second study, I focused on the comparison between three learning platforms: the decimal tutor, the game *Decimal Point*, and a reskin of the game *Decimal Point* with a more masculine narrative. The goal of this study was to examine whether the observed gender effects generalize to a different game environment which is more aligned with boys' interests.

The two studies also incorporated additional survey items for a more comprehensive representation of the gender effects. In particular, the pre-intervention survey featured items adapted from the COAT-PM measures (Liben & Bigler, 2002) to capture multiple dimensions of gender beyond binary gender identity. In the post-intervention survey, rating items related to situational interest (Linnenbrink-Garcia et al, 2010), state anxiety (B. G. Chung et al., 2010; Veit & Ware, 1983), evaluation apprehension and self-efficacy (Spencer et al., 1999) were included to test different hypotheses about the pathway from gender to learning outcomes. In turn, the new studies and survey measures yielded further insights into when and how girls learn more than boys, which have important implications for digital learning game design and math education. The next three chapters will describe the study contexts and findings in detail.

5. Investigating the Effects of Gender and Self-explanation in Decimal Learning

There is emerging evidence that digital learning games can help bridge the gender gap in STEM learning, allowing girls to catch up with boys when girls are often disadvantaged in the classroom. However, it is unclear which learning game features can induce this effect, and whether such features also generalize to other learning games or tutoring systems. In an earlier analysis of *Decimal Point*, a digital learning game that teaches decimal numbers and operations to middle school students, we have identified a consistent gender difference favoring girls in learning outcomes, which can be explained by girls' better performance in the self-explanation activities. Our current work extends on these prior results through a 2x2 study with 386 students, where we manipulated whether students learned from *Decimal Point* or a conventional tutor, and whether they performed self-explanation or not. Our results showed that, across conditions, girls performed worse than boys at pretest but had similar posttest performance after the intervention. Additionally, in the two conditions with self-explanation, performance on the self-explanation activities could explain the gender differences in learning outcomes. These results point to self-explanation, rather than the game environment of *Decimal Point*, as the driving force behind the gender effects. At the same time, we found that the game led to more engagement and less anxiety than the tutor, suggesting that using the game is still more beneficial overall. Finally, through an analysis of multiple gender dimensions – namely gender-typed occupational interests, activities and traits – we found that masculine-typed behaviors are predictive of engagement, while feminine-typed behaviors are predictive of evaluation apprehension. These results yield important insights into the design of inclusive and instructionally effective digital learning games.

Introduction

There is an established gender gap in middle school math education, where girls report higher anxiety and lower engagement than boys, which negatively impacts their performance and even long-term career choices (Breda et al., 2018; Hill et al., 2016; C. Huang, 2013; Reilly et al., 2015). In addition, gender differences favoring boys still emerge when focusing on data representing top performers among students or in advanced areas of math (Breda et al., 2018; Wai et al., 2010). As an increasingly popular form of educational technology, digital learning games have been identified as a possible contributor to narrowing the gender gap, particularly in STEM education (Connolly et al., 2009; Hou et al., 2020a; Kinzie & Joseph, 2008; Law, 2010; Pezzullo et al., 2017; Steiner et al., 2009). However, there remains mixed evidence about the benefits of learning games across knowledge domains and student populations; while some prior works have shown that girls benefited more from learning games than boys (Khan et al., 2017; Klisch et al., 2012; B. McLaren et al., 2017; Tsai, 2017), others reported no gender differences in learning outcomes after game play (Chang et al., 2014; Clark et al., 2011; Dorji et al., 2015; Manero et al., 2016; Papastergiou, 2009).

Our studies of the learning game *Decimal Point* may shed light on the learning game factors that contribute to gender differences. In particular, across several studies of the game on over 1,000 students, we have identified a consistent trend of girls learning more from the game than boys (H. A. Nguyen et al., 2022). Additionally, we found that student's performance on the self-explanation activities in the game, which prompted them to explain how they arrived at the solution, mediated the relationship between gender and learning outcomes. In particular, girls tended to make fewer self-explanation errors than boys, which helped them learn more. At the same time, there may be other factors in play that contribute to the gender effects. First, the game's casual setting and slow pace may be more closely aligned with girls' gaming preferences (Arroyo et al., 2013; Chou & Tsai, 2007; Dele-Ajayi et al., 2018), leading to higher engagement and therefore better learning from girls (i.e., the *engagement hypothesis*). Second, the game environment may help alleviate the experience of stereotype threat in math learning, which in turn free up working memory space and allow girls to learn better (i.e., the *stereotype threat hypothesis* - Gödöllei Lappalainen, 2017; Sitzmann, 2011).

At the same time, recent research in gender studies has advocated for a more comprehensive representation of gender beyond gender identity, taking into account factors such as gender-typed occupational interests, activities and traits (Hyde et al., 2019; Liben & Bigler, 2002). Examining these attributes would clarify which gender dimensions and game features best predict learning outcome and how they interact (Egan & Perry, 2001). Furthermore, they will contribute to the development of more inclusive learning platforms across different age groups. Thus, our analysis of gender differences in Decimal Point aims to not only clarify the role of the potential contributing factors – namely self-explanation performance, engagement and experience of stereotype threat – but also investigate how different dimensions of gender interact with these factors. To this end, we conducted a 2x2 experiment manipulating the learning platform (Decimal Point versus conventional tutor) and presence of self-explanation activities (with versus without self-explanation). Our research questions for this experiment are as follows.

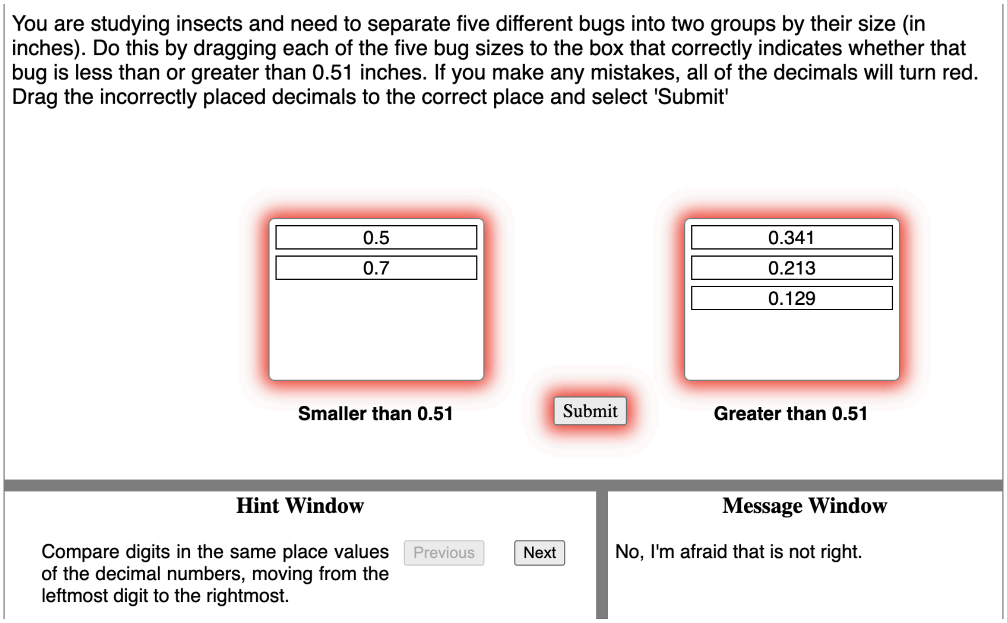
RQ1: *How do the learning platforms and self-explanation prompts impact students' learning outcomes?*

RQ2: *How do the learning platforms and self-explanation prompts influence the relationship between gender and learning outcomes?*

RQ3: *How do students' gender dimensions influence their enjoyment with the game and the tutor?*

The Learning Game *Decimal Point* and the *Decimal Tutor*

This study employed the *Decimal Point* game version from McLaren et al. (2022b), where students played through all 24 mini-games in a fixed order, and in each mini-game round, hints and error messages were provided. In each mini-game round, students could request up to three levels of hints for the problem-solving activity at any time, with the final hint level being the bottom-out hint that reveals the solution. On the other hand, error messages were triggered when students made errors that reflected specific decimal misconceptions (e.g., labeling 0.213 as larger than 0.51, which indicates the misconception that longer decimals are larger - Irwin, 2001). In these cases, a pre-defined message would appear in a pop-up window to remind students about the exhibited misconception and how to correct it.



You are studying insects and need to separate five different bugs into two groups by their size (in inches). Do this by dragging each of the five bug sizes to the box that correctly indicates whether that bug is less than or greater than 0.51 inches. If you make any mistakes, all of the decimals will turn red. Drag the incorrectly placed decimals to the correct place and select 'Submit'

Compare digits in the same place values of the decimal numbers, moving from the leftmost digit to the rightmost.

No, I'm afraid that is not right.

Figure 5.1. An example level in the tutor, corresponding to the mini-game in Figure 5.2.

The *Decimal Tutor* (Figure 5.1) is an Intelligent Tutoring System that features identical learning content and scaffolding mechanisms as in the *Decimal Point* game. The tutor also presents a series of levels similar to the *Decimal Point* mini-games, each with a problem-solving and a self-explanation activity. However, its interface only consists of basic widgets, without any game characters, colorful elements and fantasy narratives. An earlier version of the tutor, without hints and error messages, was used in a media comparison study by McLaren et al. (2017), which showed that the game led to significantly better learning outcomes and enjoyment than the tutor. A post hoc analysis of the study also revealed that girls learned more from the game than boys; however, it did not explore the potential mediating role of self-explanation performance (McLaren & Farzan et al., 2017). In the current study, we again experimented with a comparison of the game and the tutor, while also manipulating whether students performed self-explanation. In this setting, we aim to investigate which features of these learning platforms could lead to the observed gender differences in prior *Decimal Point* studies, and whether these features are specific to the game or also present in the tutor.

Methods

Participants and Design

576 students across eight elementary and middle schools in a mid-sized U.S. city participated in our study, which was conducted over six days during regular class time. On the first five days, students went through the pretest, demographic survey, intervention materials, evaluation questionnaire, and the immediate posttest. The final day of the study took place one week later, where students completed the delayed posttest. To avoid potential distractions that may occur when students sit next to each other but use different learning platforms, students were assigned by classroom to use either the game or the tutor. Within each classroom, they were randomly assigned into a condition with prompted self-explanations or a condition without.

90 students from one school were excluded from analysis, due to a technical issue that led to the study data at this school not being recorded. Among the remaining students, 386 students finished all study materials, with 105 in the game with self-explanation (G-SE) condition, 98 in the game without self-explanation (G-NSE) condition, 96 in the tutor with self-explanation (T-SE) condition, and 87 in the tutor without self-explanation (T-NSE) condition. These students ranged in age from 10 to 13 years (*M* = 10.85, *SD* = 0.65). In terms of gender identity, 52% (n = 201) of the students identified as male, 47% (n = 182) as female, 0.3% (n = 1) as trans or non-binary, and 0.6% (n = 2) preferred not to disclose their gender. Due to the small sample size of the last two categories, we excluded the three students from analyses of gender identity, but still included them in analyses of multidimensional gender scales.

Materials

Students completed all study materials on a web-based learning environment (Aleven et al., 2009b). The materials included three versions of the test, a pre-intervention and post-intervention survey, in addition to the two learning platforms mentioned above.

Pretest, Posttest and Delayed Posttest. Each test features 43 items that each range from 1 to 3 points, for a total of 52 points. The test items were designed to either assess the decimal skills and procedures practiced in the intervention materials (e.g., "place 0.4 on a number line from -1 to 1") or probe for high-level conceptual understanding (e.g., "is a longer decimal always larger than a shorter decimal?"). There were three isomorphic versions of the test that were randomly assigned to each student's pretest, posttest and delayed posttest.

Demographic and Gender-Typed Behaviors Survey. Before doing the pretest, students completed a demographic survey asking them about their age, grade level, self-identified gender identity and race. Then, they were assigned a 54-item survey, adapted from the Children's Occupational Interests, Activities, and Traits - Personal Measure (COAT-PM - Liben & Bigler, 2002), which assesses their interests, activities and traits in relation to gender-stereotyped norms. All survey items were labeled as either masculine-typed or feminine-typed[6], rated on a Likert scale, and belonged to three domains. The *occupational interests* domain measures the degree of interest in pursuing certain professions, with 18 items rated from 1 (not at all) to 4 (very much). These items include occupations like “hairstylist” or “nursed” (feminine) and “construction worker” or “engineer” (masculine). The *activity* domain evaluates the frequency of engaging in particular activities, with 18 items rated from 1 (never) to 4 (often or very often). Examples of these activities are "making jewelry" or "taking dance lessons" (feminine) and "playing basketball" or "going fishing" (masculine). The *traits* domain gauges self-perceptions of personal characteristics, with 18 items rated from 1 (not at all like me) to 4 (very much like me). These items encompass qualities such as “gentle” or “neat” (feminine) and "adventurous" or "confident" (masculine).

We note that the survey items were chosen to portray *stereotypical* perceptions of gender differences, rather than actual gender differences. Our objective in examining the students' ratings of these items is to measure the extent to which they align their behaviors with conventional gender norms and stereotypes (Liben & Bigler, 2002). To this end, we computed two scales of feminine-typed behaviors (α = 0.85) and masculine-typed behaviors (α = 0.81) by averaging the corresponding items from all three dimensions. In other words, in addition to their gender identity, each student’s gender is represented by a feminine-typed scale and a masculine-typed scale, which are continuous measures ranging from 1 to 4.

Table 5.1. The dimensions of enjoyment covered in the post-intervention questionnaire.

Dimension	Example statement
Affective engagement	I felt frustrated or annoyed.
Behavioral / cognitive engagement	I tried out my ideas to see what would happen.
Situational interest	The game (tutor) was exciting.
Achievement emotion	I enjoyed the challenge of learning the material.
Experience of meaning	The game [tutor] felt relevant to me.
Experience of mastery	I felt capable while playing the game [learning from the tutor].
Experience of appropriate challenge	The game [tutor] was challenging but not too challenging
Evaluation apprehension	If I did poorly on this activity, people would look down on me.
Test self-efficacy	I could handle this activity.
State anxiety	During the activity, I felt very nervous.

[6] This labeling was performed in the back end for data analysis and was not visible to the students when they took the survey.

Evaluation Questionnaire. After completing the game or the tutor, students were asked to rate several statements about their learning experience on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). These statements cover seven dimensions of enjoyment: multidimensional engagement (Ben-eliyahu et al., 2018) with the affective subscale (3 items, α = .70) and behavioral / cognitive subscale (3 items, α = .54); situational interest (3 items, α = 0.83 - Linnenbrink-Garcia et al., 2010); enjoyment dimension of achievement emotion (6 items, α = 0.89 - Pekrun, 2005); evaluation apprehension (4 items, α = 0.86 - Spencer et al., 1999); test self-efficacy (5 items, α = 0.74 - Spencer et al., 1999); state anxiety (3 items, α = 0.70 - Chung et al., 2010; Veit & Ware, 1983). Table 5.1 includes example items for each of these constructs. For our analysis, we excluded the behavioral / cognitive engagement subscale due to its low reliability.

Results

First, a series of repeated-measures ANOVA showed a significant difference for all students between pretest and posttest score, F = 97.88, p < .001, η^2_p = .20, as well as between pretest and delayed posttest score, F = 128.33, p < .001, η^2_p = .25. In other words, students' performance improved after learning from both the game and the tutor.

We also examined the correlations between gender dimensions. Our results showed that gender identity, where “female” was coded as 1 and “male” coded as 0, was positively correlated with feminine-typed scale (r = 0.58, p < .001) and negatively correlated with masculine-typed scale (r = -0.34, p < .001). In addition, feminine-typed scale was positively correlated with masculine-typed scale (r = 0.20, p < .001). Given the correlation coefficients, while the three gender dimensions were moderately correlated, they were not redundant.

Next, we investigated our research questions as follows. In the first question, we focused on the differences among the four study conditions, expressed as two factors of learning platform (Game versus Tutor) and self-explanation prompt (With Self-explanation and Without Self-explanation). In the latter two questions, we further examined the interaction between gender and the condition factors to identify the similarities and differences between gender identity and gender-typed scales.

RQ1: *How do the learning platforms and self-explanation prompts impact students' learning outcomes?*

Table 5.2. Descriptive statistics of students' test performance and number of hint requests across the four study conditions, reported in M (SD) format.

Condition	N	Pretest	Posttest	Delayed Posttest	Hint request
G-SE	105	21.35 (11.56)	24.61 (12.14)	25.07 (13.29)	102.45 (83.82)
G-NSE	98	23.35 (12.63)	26.26 (12.25)	27.26 (12.45)	84.84 (83.15)

T-SE	96	20.93 (10.67)	23.43 (10.36)	24.67 (10.73)	77.07 (77.26)
T-NSE	87	20.72 (9.71)	23.21 (11.03)	23.99 (11.53)	93.01 (75.07)

The mean scores on the pretest, immediate posttest and delayed posttest by condition are shown in Table 5.2. With pretest scores as covariate, a two-way ANCOVA comparing immediate posttest scores by condition showed that neither the learning platform (F = 0.58, p = .45, η^2_p = .002) nor the self-explanation prompts (F = 0.05, p = .83, η^2_p = .000) had a significant main effect. The interaction between the learning platform and self-explanation prompts was likewise not significant (F = 0.03, p = .89, η^2_p = .000). When comparing delayed posttest scores by condition, we observed similar results: the main effects of the learning platform (F = 0.72, p = .40, η^2_p = .002) and self-explanation prompts (F = 0.13, p = .72, η^2_p = .000) were not significant, neither was their interaction (F = 0.32, p = .57, η^2_p = .001). Thus, our hypothesis that the game and self-explanation prompts would lead to better learning outcomes was not confirmed.

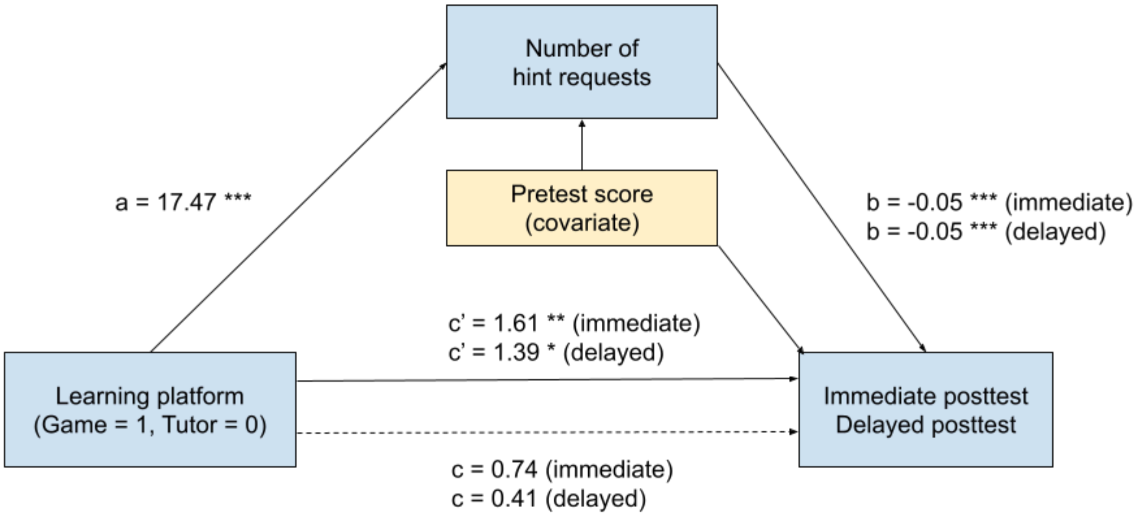


Figure 5.2. Diagram of the mediation pathway from learning platform to posttest and delayed posttest performance through hint usage behavior. (*) p < .05, (**) p < .01, (***) p < .001.

We conducted a post hoc analysis on students' performance during the intervention to better understand the lack of condition effect. In particular, we investigated how frequently students requested hints in each condition and how this behavior impacted their learning outcomes. To this end, we built a mediation model with the learning platform as an independent variable, the number of hint requests during intervention as a mediator, and the posttest / delayed posttest score as the dependent variable. The confidence interval of the indirect effect was estimated at the 0.05 significance level via bias-corrected non-parametric bootstrapping with 2000 iterations (Hayes & Rockwood, 2017; Vallat, 2018). Based on the mediation results (Figure 5.2), we found that the effect of the learning platform on posttest performance was mediated by the number of hint requests. The regression coefficient between the learning platform (where the game is coded as 1) and number of hint requests was positive and significant, while the coefficient between the number of hint requests and posttest score was negative and significant. The bootstrap procedures also indicated a significant indirect effect (*ab* = -0.87, 95% CI [-1.45, -0.29], $p < .001$). Similar findings were observed in the mediation model predicting delayed posttest scores, with a significant indirect effect of the number of hint requests (*ab* = -0.97, 95% CI [-1.65, -0.41], $p < .001$). On the other hand, the direct effect of the game on test performance, without considering the mediator, was positive and significant. In other words, while the game did lead to better learning when controlled for the number of hint requests, students playing the game also requested more hints than those using the tutor, which negatively impacted their learning. Due to these conflicting trends, the overall total effect of the learning platform on test performance was not significant.

RQ2: *How do the learning platforms and self-explanation prompts influence the relationship between gender and learning outcomes?*

Table 5.3. Descriptive statistics of test performance by gender, reported in M (SD) format.

Gender Identity	N	Pretest	Posttest	Delayed Posttest
Female	182	20.00 (9.70)	23.71	24.52
Male	201	23.27 (12.27)	25.28	26.20

Table 5.3 shows the descriptive statistics for the pretest, posttest and delayed posttest performance between boys and girls. For pretest performance, a one-way ANOVA revealed a significant effect of gender identity ($F = 8.27$, $p = .004$, $\eta^2_p = .021$), where boys had higher pretest scores than girls. When comparing posttest and delayed posttest performance, we also examined potential interactions between gender identity and condition factors, through a series of three-way ANCOVAs with pretest score as covariate. Our results showed that neither the main effect of gender identity ($F = 2.28$, $p = .13$, $\eta^2_p = .006$) nor its interactions with the learning platform ($F = 0.63$, $p = .43$, $\eta^2_p = .002$) and self-explanation prompts ($F = 1.25$, $p = .26$, $\eta^2_p = .003$) were significant. With delayed posttest performance, there was likewise no significant main effect of gender identity ($F = 3.03$, $p = .08$, $\eta^2_p = .008$) or interaction effect between gender identity and self-explanation prompts ($F = 2.86$, $p = .09$, $\eta^2_p = .008$).

We also analyzed pretest scores with multiple dimensions of gender. Using a regression model predicting pretest score based on masculine-typed and feminine-typed scales, we found that masculine-typed scale was a significant and positive predictor ($\beta = 3.64$, $p = .003$), while feminine-typed scale was a significant and negative predictor ($\beta = -4.42$, $p < .001$). Another model predicting pretest score based on gender-typed scales and binary gender identity (with “female” coded as 1 and “male” coded as 0) showed that masculine-typed scale ($\beta = 4.01$, $p = .008$) and feminine-typed scale (β = -4.96, $p = .004$) remained significant predictors, while gender identity was not ($\beta = 0.75$, $p = .66$)

Next, we built a regression model with pretest score as covariate and the following predictor variables: learning platform (Game versus Tutor), self-explanation prompt (With Self-explanation versus Without Self-explanation), masculine-typed scale, feminine-typed scale, and their interactions with the learning platform and self-explanation prompt. Our results showed that, when predicting posttest scores and delayed posttest scores, none of the predictors were significant.

Following previous analyses of the gender effects in the game *Decimal Point* (Nguyen et al., 2022), we also examined the self-explanation performance by gender, as well as the mediating role of self-explanation performance in the relationship between gender and learning outcomes. In this context, we considered only students who were prompted to perform self-explanation, i.e., those in the G-SE (n = 105) and T-SE (n = 96) conditions. A two-way ANCOVA assessing the effects of the learning platform and gender identity on the number of self-explanation errors, with pretest score as covariate, showed a significant main effect of gender (F = 7.53, p = .007, η^2_p = .037), with girls (M = 30.35, SD = 14.35) making fewer self-explanation errors than boys (M = 35.74, SD = 13.12). The effects of the learning platform (F = 0.38, p = .54, η^2_p = .002) and its interaction with gender identity (F = 0.23, p = .63, η^2_p = .001) were not significant.

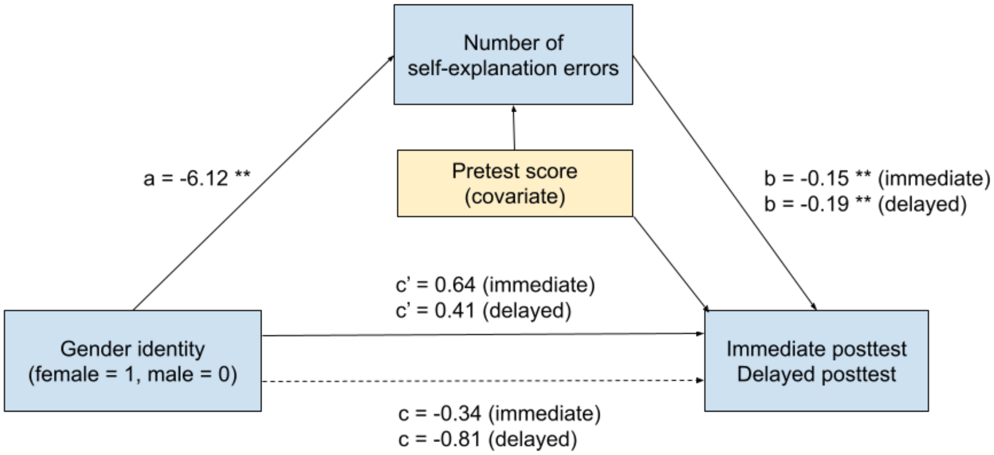


Figure 5.3. Diagram of the mediation pathway from gender identity to posttest and delayed posttest performance, through self-explanation performance. (**) p < .01.

To identify how self-explanation performance may mediate the relationship between gender identity and learning outcomes, we constructed two mediation models with pretest score as covariate, gender identity as an independent variable, self-explanation error as a mediator, and posttest / delayed posttest score as the dependent variable (Figure 5.3). Our results revealed a negative association between gender identity (where “female” was coded as 1) and self-explanation error, as well as between self-explanation error and test performance. Bootstrapping procedures indicated a significant indirect effect of self-explanation performance on the relationship between gender identity and posttest performance (*ab* = 0.97, 95% CI [0.48, 1.90], p = .004), as well as between gender identity and delayed posttest performance (*ab* = 1.21, 95% CI [0.57, 2.20], p < .001).

In summary, we found that girls and those with greater feminine-typed behaviors performed worse at pretest than boys and those with greater masculine-typed behaviors. However, there were no gender differences in posttest and delayed posttest performance. In addition, our results showed that girls and those with greater feminine-typed behaviors made fewer self-explanation errors than boys and those with greater masculine-typed behaviors. Notably, self-explanation performance was identified as a significant mediator in the relationship between gender and test performance when controlling for pretest.

RQ3: *How do students' gender dimensions influence their enjoyment with the game and the tutor?*

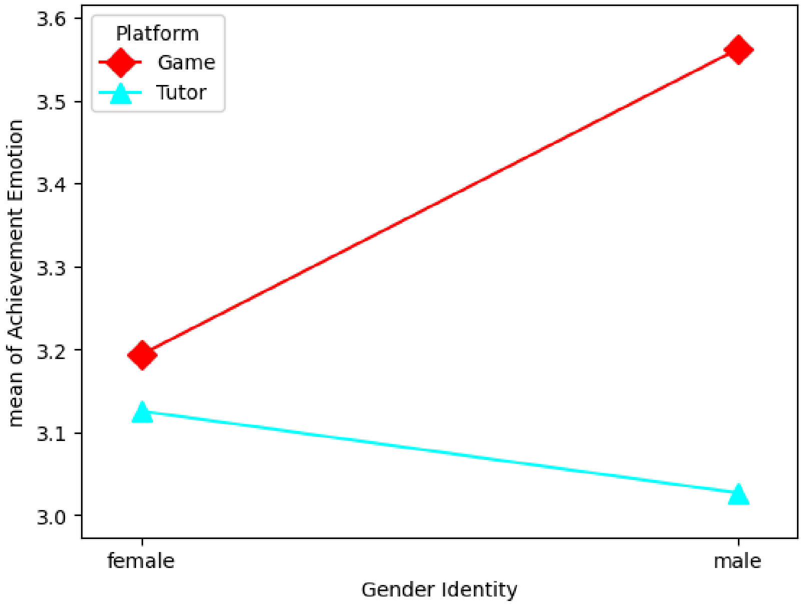


Figure 5.4: The interaction effect between gender identity and learning platform on achievement emotions.

We first performed a series of two-way ANOVAs assessing the effects of gender identity and the learning platform on each enjoyment dimension in Table 5.1. To examine the engagement hypothesis, we performed a series of two-way ANOVA on the dimensions of affective engagement, situational interest and achievement emotions. Our results showed that boys (*M* = 3.51, *SD* = 1.04) reported higher levels of affective engagement than girls (*M* = 3.17, *SD* = 0.96), *F* = 9.39, *p* < .01, η^2_p = .024, and the game (*M* = 3.54, SD = 0.96) led to higher levels of affective engagement than the tutor (*M* = 3.13, *SD* = 1.03), *F* = 14.44, *p* < .001, η^2_p = .024. Similarly, students playing the game (*M* = 3.25, *SD* = 1.17) reported more situational interest than those using the tutor (*M* = 2.85, *SD* = 1.02), *F* = 11.68, *p* < .01, η^2_p = .030. Finally, both the main effect of the learning platform (*F* = 8.79, p < .01, η^2_p = .023) and the interaction effect between gender identity and learning platform (F = 5.27, *p* = .02, η^2_p = .014) on achievement emotions were significant. As shown in Figure 5.4, the game led to higher levels of achievement emotions than the tutor, and boys reported higher levels of achievement emotions than girls in the game (*F* = 6.82, *p* = .01, η^2_p = .033), but not in the tutor (*F* = 0.45, *p* = .50, η^2_p = .003).

Finally, for the three player experience subscales, there were no significant main or interaction effects of gender identity and learning platform.

To examine the stereotype threat hypothesis, we investigated the dimensions of evaluation apprehension, test efficacy and state anxiety. Our results showed that girls (*M* = 2.43, *SD* = 1.05) reported more evaluation apprehension than boys (*M* = 2.17, *SD* = 0.99), *F* = 6.07, *p* = .01, η^2_p = .016, while the effects of the learning platform on evaluation apprehension was not significant. On the other hand, the game (*M* = 3.67, *SD* = 0.96) led to higher test efficacy than the tutor (*M* = 3.47, *SD* = 0.80), *F* = 5.09, p = .03, η^2_p = .013, while there was no significant gender effect on test efficacy. Finally, for state anxiety, there was a significant main effect of gender (*F* = 9.49, *p* < .01, η^2_p = .025) and a marginally significant main effect of the learning platform (*F* = 3.59, *p* = .06, η^2_p = .009). Girls (*M* = 2.80, *SD* = 1.07) reported higher state anxiety than boys (*M* = 2.45, *SD* = 1.07), while the game (*M* = 2.51, *SD* = 1.16) led to lower state anxiety than the tutor (M = 2.74, SD = 0.98). Across all three dimensions, there were no significant interaction effects between gender identity and learning platform.

Table 5.4. Regression models predicting enjoyment ratings based on gender-typed scales and their interactions with the learning platform.
(*) *p* < .05

	Masculine-typed scale	Feminine-typed scale	Platform (Tutor = 1)	Masculine-typed scale x Tutor	Feminine-typed scale x Tutor
Affective engagement	0.17	-0.21	-0.84	0.03	-0.16
Experience of meaning	0.39 (*)	0.08	-0.44	-0.02	-0.30
Experience of mastery	0.46 (*)	-0.10	-1.78 (*)	0.10	-0.56 (*)
Experience of appropriate challenge	-0.02	-0.06	-2.07 (*)	0.32	-0.53 (*)
Situational interest	0.38 (*)	0.13	-0.56	0.05	-0.06
Achievement emotion	0.37 (*)	0.10	-1.00	0.01	-0.32
Evaluation apprehension	-0.31	0.44 (*)	0.31	0.18	0.30
Test efficacy	0.19	-0.28	-0.67 (*)	0.09	-0.12
State anxiety	-0.37 (*)	0.58	0.79 (**)	0.13	0.39

To investigate the relationship between gender-typed scales and enjoyment measures, we built regression models predicting each of the enjoyment dimensions based on masculine-typed scale, feminine-typed scale and their interactions with the learning platform (Table 5.4). Our results showed that masculine-typed scale was a significant positive predictor of experience of meaning, experience of mastery, situational interest, achievement emotion, and a significant negative predictor of state anxiety. In addition, feminine-typed scale was a significant positive predictor of evaluation apprehension. We also found that learning from the tutor (versus playing the game) was a significant negative predictor of experience of mastery, experience of appropriate challenge and test efficacy, as well as a significant positive predictor of state anxiety. Finally, the interaction between feminine-typed scale and learning platform was a significant predictor of experience of mastery and experience of appropriate challenge. In particular, feminine-typed scale was positively associated with these measures in the tutor, but not in the game.

Discussion

In this work, we conducted a 2x2 experiment which manipulated whether students learned from the game *Decimal Point* or a conventional tutor, and whether they performed self-explanation following each problem-solving activity Following up on the consistent gender effects observed in past *Decimal Point* studies (Nguyen et al., 2022), our goal was to examine how these factors impacted learning outcomes and whether their influence interacted with gender dimensions. Based on our results, across all four study conditions, girls had lower pretest scores than boys but had similar performance in the posttest and delayed posttest. We also found that, compared to the tutor, the game did not lead to better learning but promoted higher levels of engagement and enjoyment. Finally, we identified several nuances in the relationship between gender dimensions and learning outcomes, as well as engagement, which gender identity alone did not reveal. We discuss these results in-depth as follows.

First, our findings regarding differences by gender identity are consistent with those from Nguyen et al. (2022) – girls performed worse than boys at pretest but were able to catch up after the intervention. Furthermore, among the conditions with self-explanation, we were able to replicate the mediation effect of self-explanation performance: girls tended to make fewer self-explanation errors than boys, leading to higher posttest and delayed posttest scores. Notably, this trend was present not only in the game, but also in the decimal tutor, suggesting that self-explanation performance is the driving factor behind the gender differences in learning. This result is similar to the findings from Baker et al. (under review), which shows that the frequency of gaming the system in self-explanation activities also mediated the relationship between gender and learning outcomes. In the context of our learning platforms, gaming the multiple-choice self-explanation questions involves clicking through all of the available options in quick succession without thinking through the problem. Under this definition, students who gamed the self-explanation questions would likely have a higher number of self-explanation errors, which could explain the similar findings between our mediation analysis and Baker et al. (under review). One possible reason for boys' worse self-explanation performance is that, unlike the problem-solving activities embedded in playful game contexts, the multiple-choice self-explanation questions more closely resembled typical math instructions. As boys were more engaged with the game experience, they were also more likely to become disengaged during the less playful self-explanation activities, thereby making more errors than girls.

With regards to enjoyment, we identified several gender differences across enjoyment dimensions. In particular, boys reported higher levels of affective engagement across conditions, as well as higher levels of achievement emotion only in the game conditions. Furthermore, girls reported higher levels of evaluation apprehension and state anxiety than boys across conditions. Based on these results, our engagement hypothesis was not supported – while the game's features appeared to align with girls' gaming preferences, based on results from prior surveys (Arroyo et al., 2013; Chou & Tsai, 2007; Dele-Ajayi et al., 2018; Nguyen et al., 2023), we found that boys still reported higher engagement overall. On the other hand, we found evidence for the stereotype threat hypothesis, whereby girls reported higher levels of evaluation apprehension and state anxiety than boys. However, this gender difference was present across conditions, contrary to our expectation it would only manifest in the tutor due to the game's effects on reducing the experience of stereotype threat in girls. One explanation for this outcome is that, even though the game *Decimal Point* features playful game narratives and characters, the math content is still salient and likely remains the focus of the students. Thus, it would be worth investigating whether a more immersive game environment – for instance, one that offers a sandbox experience, which is highly popular to young players (H. A. Nguyen et al., 2023) – could be more effective at alleviating the stereotype threat, and whether its effects translate to gender differences in learning outcomes.

When considering additional dimensions of gender, namely gender-typed occupational interests, activities and traits (Liben & Bigler, 2002), we observed several trends that complemented our binary gender identity analyses. First, while boys outperformed girls at pretest, we found that gender identity was not a significant predictor of pretest scores, when controlled for masculine-typed and feminine-typed scales. This result implies that a more nuanced representation of gender is a more powerful predictor of math performance. In addition, masculine-typed scale was a significant predictor of several dimensions of enjoyment, including experience of meaning, experience of mastery, situational interest, achievement emotion and state anxiety, while feminine-typed scale was a significant predictor of evaluation apprehension. While our gender identity analyses showed that boys were more engaged and girls were more anxious, this result provides a more nuanced understanding of how different gender dimensions can predict different measures of enjoyment (Hyde et al., 2019).

We should also note the comparison between the game and the tutor. While McLaren et al. (2017) had shown that the game *Decimal Point* led to significantly more learning and enjoyment than the tutor, these results were not replicated in the current study, where we found no significant differences in learning or enjoyment between the two platforms. One key difference between the platforms in our study and those in McLaren et al. (2017)'s study is the inclusion of hints. Through a post hoc analysis, we also identified that the new hint system was a contributing factor to the lack of differences between the game and the tutor. In particular, *Decimal Point*'s advantages over the tutor was countered by students' higher number of hint requests in the game, which led to worse learning outcomes. This result is consistent with the findings from McLaren et al. (2022), whereby the game version with hints led to worse learning outcomes than the game version without hints when deployed in the classroom. Overall, these trends suggest that the current hint system is not effective, as students could quickly go through all of the hint levels to reach the bottom-out hint without thinking through the problem (Aleven et al., 2016). Thus, a revision to the hint mechanisms that provides more metacognitive support (Aleven et al., 2006; Roll et al., 2007) is an important next step for future *Decimal Point* studies.

Additionally, while the self-explanation activities helped girls learn more than boys, we observed no significant differences in learning between the conditions with and without self-explanation. While our integration of self-explanation prompts into the decimal game and tutor were informed by prior research showing the benefits of multiple-choice self-explanation (Aleven & Koedinger, 2002), even in a learning game context (Mayer, 2019), it is possible that these activities did not provide significant learning benefits beyond the problem-solving questions that were present in all four conditions. In this case, it would be interesting to examine whether an open-ended self-explanation format, which allows for more active and constructive learning (Wylie & Chi, 2014), would have a more pronounced effect on student learning. Given the results from a prior study of *Decimal Point* indicating that open-ended self-explanation led to better learning than multiple-choice self-explanation (McLaren et al., 2022), we expect performing open-ended self-explanations would also be more beneficial than not performing self-explanations.

Conclusion

In summary, our findings have revealed important insights into the two main hypotheses for the gender effects in *Decimal Point*. The engagement hypothesis, which states that girls learned more than boys because they were more engaged with the game's features, was not supported, as we observed higher levels of engagement from boys and those with strong masculine-typed behaviors. The stereotype threat hypothesis, which states that girls learned more than boys because the game helped reduce their experience of stereotype threat when learning math, was likewise not supported. While we observed that girls reported higher levels of anxiety and evaluation apprehension than boys, this difference was present in both the game and the tutor, implying that the game did not reduce the experience of stereotype threat more effectively than the tutor. Instead, the primary driver of the gender differences in learning was self-explanation performance, whereby girls consistently did better than boys in previous studies (Nguyen et al., 2022) and in the current study. The robust effect of self-explanation is a novel and significant result which learning game researchers and designers could utilize to bridge the gender gap in other knowledge domains.

As the next step, we would like to examine the engagement hypothesis and stereotype threat hypothesis in a broader context. While these hypotheses were not supported by the findings from the current study, we should note that the game features of *Decimal Point* were designed to appeal to all young students, rather than align with specific gender preferences (Forlizzi et al., 2014). Thus, it would be interesting to compare the effects of *Decimal Point* to another learning game with a more gendered narrative, catering to either boys or girls' preferences. This follow-up study will shed light on whether gender-aligned game features can produce differences in learning outcomes and enjoyment, as well as how their effects interact with the identified self-explanation effects. To this end, we deployed a survey to understand how student’s gender identity and gender-typed behaviors shape their gaming preferences, as described in the next chapter.

6. Gender Differences in Learning Game Preferences: Results Using a Multidimensional Gender Framework

Prompted by findings of gender differences in learning game preferences and outcomes, education researchers have proposed adapting games by gender to foster learning and engagement (Kinzie & Joseph, 2008; Steiner et al., 2009). However, such recommendations typically rely on intuition, rather than empirical data, and are rooted in a binary representation of gender. On the other hand, recent evidence from several disciplines indicates that gender is best understood through multiple dimensions, including gender-typed occupational interests, activities, and traits (Hyde et al., 2019; Liben & Bigler, 2002). Our research seeks to provide learning game designers with empirical guidance incorporating this framework in developing digital learning games that are inclusive, equitable, and effective for all students. To this end, we conducted a survey study among 333 5th and 6th grade students in five urban and suburban schools in a mid-sized U.S. city, with the goal of investigating how game preferences differ by gender identity or gender-typed measures. Our findings uncovered consistent differences in game preferences from both a binary and multidimensional gender perspective, with gender-typed measures being more predictive of game preferences than binary gender identity. We also report on preference trends for different game genres and discuss their implications on learning game design. Ultimately, this work supports using multiple dimensions of gender to inform the customization of learning games that meet individual students' interests and preferences, instead of relying on vague gender stereotypes.

Introduction

While digital learning games have been shown to be a promising form of instruction thanks to their motivational and learning benefits (Hussein et al., 2021), designing effective games requires a clear understanding of the preferences of different player populations. For instance, there are consistent gender differences in game preferences, such that boys tend to prefer faster paced, action-style games, while girls tend to prefer games with puzzle and social interaction elements (Aleksić & Ivanović, 2017; Chou & Tsai, 2007). With digital learning games specifically, girls tend to rank goal clarity and social interaction as more important than boys, while boys tend to prefer challenge, progress feedback, and visual appeal (Dele-Ajayi et al., 2018). These gendered preferences can produce meaningful differences in learning behaviors and outcomes. For example, girls have sometimes been shown to enjoy learning games more (Adamo-Villani et al., 2008) and have greater learning outcomes (Khan et al., 2017; H. A. Nguyen et al., 2022) than boys. Different features of learning games have also been shown to induce gendered effects, such as girls benefiting more from a digital learning companion (Arroyo et al., 2013).

Prompted by these findings, researchers have proposed adapting digital learning games based on gender to create more inclusive and equitable learning experiences (Kinzie & Joseph, 2008;

Steiner et al., 2009). However, such recommendations for gender-based adaptation typically rely on game designers' intuitions, stereotypes, or preferences observed through playtesting and focus groups (Farrell & Moffat, 2014; Seaborn & Fels, 2015), rather than experimental studies. Moreover, such efforts are limited by a grounding in the gender binary, which views gender as one of two discrete categories, male and female, framed as biologically-based, apparent at birth, and stable over time (Hyde et al., 2019). Conflating gender with binary, birth-assigned sex is not only imprecise, but may also contribute to gender-stereotyped interests and gender disparities in academic achievement (B. G. Chung et al., 2010; Galdi et al., 2014). In contrast, evidence from multiple disciplines has demonstrated that gender is complex, fluid and dynamic, comprising multiple interrelated but separate dimensions (Hyde et al., 2019).

Thus, a more sophisticated and nuanced approach to understanding gender differences in learning game preferences should take into account not only self-reported gender identity (e.g, male, female, non-binary, trans), but also other gender dimensions – such as gender-typed occupational interests, activities, and traits (Liben & Bigler, 2002) – which are continuous and more fine-grained than gender identity. Such an approach is consistent with best practices in gender studies research, including with late elementary and middle school youth (Fast & Olson, 2018; Hyde et al., 2019), among whom these dimensions of gender are only modestly correlated with one another (Cook et al., 2019; Perry et al., 2019). Notably, prior work has shown that middle school children can be differentiated along these dimensions of gender and can reliably report on their gender-typed behaviors (Cook et al., 2019; Liben & Bigler, 2002; Martin et al., 2017).

Motivated by this multidimensional approach, our work seeks to better understand students' digital game preferences and how they relate to different dimensions of gender, through a survey deployed to 333 young students. The first half of the survey asked students to rank their preferred game genres (e.g., *action*, *strategy*, *sandbox*) and game narratives (the overarching game world and story – e.g., fighting pirates, hunting treasures), while the second half queried about dimensions of gender identity and gender-typed occupational interests, activities, and traits. With this survey design, our primary research questions are as follows.

RQ1*: Are there significant gender differences – based on gender identity or gender-typed interests, activities, and traits – in game genre preferences?*

RQ2: *Are there significant gender differences – based on gender identity or gender-typed interests, activities, and traits – in game narrative preferences?*

By addressing these questions, our work contributes to research on young students' preferences in digital games. We also demonstrate, through statistical testing and qualitative analyses, how additional dimensions of gender can better reflect individual preferences than binary gender identity. In turn, this knowledge can enable the design and development of more inclusive and effective learning games.

Methods

Participants

Our sample comprises *n* = 333 students who participated in a classroom study in 5th (*n* = 100) and 6th grades (*n* = 233) across five urban and suburban public schools in a mid-sized U.S. city. Students ranged in age from 10 to 13 years (*M* = 11.06, *SD* = .69). In terms of self-reported gender identity, 52.0% (*n* = 173) described themselves as male, 47.0% (*n* = 156) as female, 0.3% (*n* = 1) as trans or nonbinary, and 0.6% (*n* = 2) preferred not to disclose their gender. Because the subsample of students in the last two categories was small, these 3 students were excluded from statistical analyses of gender identity.

Materials and Procedures

Surveys were administered as part of a classroom study of a digital learning game in mathematics. Students first filled out a demographic questionnaire that queried about their age, grade level, and an open-ended item to self-identify their gender. Next, they completed two sets of surveys related to gender dimensions and game preferences.

Gender-Typed Behaviors. This survey is based on the Children's Occupational Interests, Activities, and Traits-Personal Measure (COAT-PM - Liben & Bigler, 2002), which assesses youth's interests, activities, and traits in relation to gender-stereotyped norms. The COAT-PM has two scale scores comprising masculine- and feminine-typed occupational interests, activities, and traits. For brevity, we adapted the measure by removing 9 gender-neutral items, 6 masculine-typed items, and 6 feminine-typed items, such that it consisted of 54 items in total (out of the original 75), rated on Likert scales from 1 to 4. The *occupational interests* domain indicates how much one wants to pursue certain jobs, and includes 18 items rated from 1 (not at all) to 4 (very much), such as "hairstylist" or "nurse" (feminine) and "construction worker" or "engineer" (masculine). The *activity* domain indicates how often one performs certain activities, and includes 18 items rated from 1 (never) to 4 (often or very often), such as "make jewelry" or "take dance lessons" (feminine) and "play basketball" or "go fishing" (masculine). The *traits* domain indicates how much one would describe themselves, and includes 18 items rated from 1 (not at all like me) to 4 (very much like me), such as "gentle" or "neat" (feminine) and "adventurous" or "confident" (masculine). Note that these items do not reflect actual gender differences (e.g., we do not assume "confidence" is a male-only trait); rather, they only portray *stereotypical* perceptions of gender differences. Our goal in analyzing students' ratings of these items is to quantify how much they shape their behaviors around traditional gender norms and stereotypes (Liben & Bigler, 2002). Each domain also included two gender-neutral items for filler which are excluded from analysis (e.g., "YouTuber," "practice an instrument," "friendly"). The internal consistency (Cronbach's α) was good for both the masculine-typed occupational interests, activities, and traits scale (α = .80) and the feminine-typed occupational interests, activities, and traits scale (α = .85).

Game Genre and Narrative Preferences. We examined digital learning game preferences in terms of game genres and narratives. First, students were asked to indicate their top three preferred game genres from a list of seven options, selected to capture the most popular genres among young players (GameTree Team, 2019; Yee, 2017), and then rank them from most to least liked. The game genres (and example games) listed were: *action* (e.g., Fortnite, Splatoon); *sports & racing* (e.g., Rocket League, FIFA); *strategy* (e.g., Age of Empires, Civilization); *sandbox* (e.g., Roblox, Minecraft); *music & party* (e.g., Pianista, Just Dance); *role-playing* (e.g., Stardew Valley, Legend of Zelda); and *casual* (e.g., Bejeweled, Animal Crossing).

Next, students were given the descriptions of four game narratives: *Amusement Park* (lead new alien friends around an amusement park), *Treasure Hunt* (search for hidden treasure among ocean landmarks, racing against an arch-nemesis), *Helping a Sea Friend* (help a sea creature save an underwater city that has lost power before it's too late), and *War at Sea* (fight criminal naval masterminds and disable their secret weapon to save the world). The narratives were brainstormed by the research team to vary along multiple dimensions – such as world building, goal focus and presence of competition or cooperation – that were shown to be differentially preferred by boys and girls in prior work (Arroyo et al., 2013; Dele-Ajayi et al., 2018). Based on the provided descriptions, students were asked to rank the four game narratives from most to least interesting.

Results

Game Genre Preferences

We focused our analyses on students' first and second choices among game genres. A series of chi-square analyses of gender identity (boy or girl) by preference for each game genre revealed significant binary gender differences in most game preferences, such that boys tended to prefer the genres of *action* and *sports & racing*, whereas girls tended to prefer *sandbox*, *music & party*, *role-playing*, and *casual* games. Boys and girls were similarly likely to prefer *strategy* games. Table 6.1 displays the percentage of boys and girls who ranked each game genre as their 1st or 2nd ranked options. Following Cohen's guidelines (Cohen, 2013), the reported significant gender identity differences on game preferences all have small to medium effect sizes (.10 < ϕ < .50).

Table 6.1. Gender differences in game genre preferences according to gender identity. Percentages sum to 200 to represent first and second rank choices.
(***) *p* < .001, (**) *p* < .01, (*) *p* < .05.

Game Genre	% Boys	% Girls	χ^2 (1)	ϕ
Action	70.5	39.1	32.80***	-.32
Sports & Racing	45.1	14.1	37.22***	-.34
Strategy	11.6	5.8	3.42	-.10
Sandbox	46.8	59.0	4.86*	.12
Music & Party	5.2	37.8	53.23***	.40
Role-Playing	12.7	21.2	4.19*	.11

Casual	8.1	23.1	14.29***	.21

To test for significant differences in digital game genre preferences based on masculine- and feminine-typing, we ran a series of one-way between-subject ANOVAs comparing the masculine- and feminine-typed behaviors of students who ranked specific game genres as their 1st or 2nd options. We found that students who preferred the genres of *action* and *sports & racing* reported significantly higher masculine-typed behaviors, whereas students who preferred the genres of *sandbox*, *music & party*, and *casual* games reported significantly lower masculine-typed behaviors. By contrast, students who preferred the genres of *action* and *sports & racing* reported significantly lower feminine-typed behaviors, and students who preferred the *music & party* genre reported significantly higher feminine-typed behaviors. See Table 6.2 for the descriptive statistics, effect sizes and *F*-statistics; here a game genre is considered not preferred (Not P.) if it is not within the top two ranked options. Following Cohen (Cohen, 2013), most of the significant gender-typing differences in game preferences can be interpreted as medium (*d* ≅ .50), with the exception of the large gender difference in preferring *music & party* games (*d* > .80).

Table 6.2. Gender differences in game genre preferences according to feminine- and masculine-typed occupational interests, activities, and traits.

Game Genre	Feminine-typed				Masculine-typed			
	Pref.	Not P.	*F*	*d*	Pref.	Not P.	*F*	*d*
	M (SD)	*M (SD)*			*M (SD)*	*M (SD)*		
Action	2.17 (.47)	2.32 (.48)	8.04**	-.31	2.42 (.47)	2.23 (.42)	14.64 ***	.41
Sports & Racing	2.09 (.42)	2.30 (.49)	14.40 ***	-.44	2.52 (.39)	2.26 (.46)	24.14 ***	.57
Strategy	2.08 (.42)	2.25 (.48)	3.34	-.35	2.38 (.45)	2.33 (.46)	.29	.10
Sandbox	2.24 (.49)	2.23 (.47)	.01	.01	2.25 (.44)	2.44 (.46)	15.40 ***	-.42
Music & Party	2.56 (.43)	2.15 (.45)	45.62 ***	.86	2.22 (.49)	2.37 (.44)	6.41*	-.34
Role-Playing	2.31 (.42)	2.22 (.49)	1.775	.19	2.26 (.43)	2.36 (.46)	2.16	-.21
Casual	2.32 (.51)	2.22 (.47)	2.03	.22	2.22 (.44)	2.36 (.46)	3.96*	-.30

Next, we computed the correlations between gender dimensions. Binary gender identity (with “female” coded as 1 and “male” coded as 0) was positively correlated with feminine-typed scales (*r* = .62) and negatively correlated with masculine-typed scales (*r* = -.33). In addition, feminine-typed scales were positively correlated with masculine-typed scales (*r* = .19). Thus, these three gender dimensions were moderately correlated but not redundant. Because there is no evidence of multicollinearity, our next analysis involves a series of logistic regressions predicting each game genre preference (i.e., whether it is included in the top two choices) with binary gender, masculine-typed scale, and feminine-typed scale (Table 3). This allowed us to assess how much each variable predicted preferences when controlling for the other variables. For the *action, sports & racing*, *sandbox*, and *music & party* genres, masculine- and feminine-type scales were significant predictors of preference while binary gender was not. For only one genre - *casual* games - was binary gender a significant predictor while masculine- and feminine-types were not. Models predicting *role-playing* and *strategy* genres were not significant and are excluded from Table 6.3.

Table 6.3. Logistic regressions predicting genre preferences according to gender identity (female = 0), masculine-typed scale, and feminine-typed scale. *Exp(B)* represents the change in odds corresponding to a unit change in the coefficient.

Game Genre	Constant	Binary gender	Feminine-typed scale	Masculine-typed scale
Action	*B* = -2.10*, *Exp(B)* = .12	*B* = .27, *Exp(B)* = 1.31	*B* = -.68*, *Exp(B)* = .51	*B* =1.07***, *Exp(B)* = 2.91
Sports & Racing	*B* = -1.87, *Exp(B)* = .15	*B* = .60, *Exp(B)* = 1.82	*B* = -1.48***, *Exp(B)* = .23	*B* = 1.22**, *Exp(B)* = 3.37
Sandbox	*B* = .044, *Exp(B)* = 1.05	*B* = -.13, *Exp(B)* = .88	*B* = .72*, *Exp(B)* = .2.06	*B* = -1.03***, *Exp(B)* = .36
Music & Party	*B* = -2.69*, *Exp(B)* = .07	*B* = -1.22, *Exp(B)* = .30	*B* = 1.54*, *Exp(B)* = .4.67	*B* = -1.33*, *Exp(B)* = .27
Casual	*B* = -1.73, *Exp(B)* = .18	*B* = -2.75**, *Exp(B)* = .064	*B* = -.74, *Exp(B)* = .48	*B* = -.58, *Exp(B)* = 1.78

Game Narrative Preferences

A series of chi-square analyses of gender identity (boy or girl) by preference for each game narrative revealed significant gender differences in two of the four narrative preferences, such that the *War at Sea* narrative was preferred by boys and *Help a Sea Friend* narrative was preferred by girls (Table 6.4). Following Cohen (Cohen, 2013), these effect sizes are medium. On the other hand, boys and girls were similarly likely to prefer the *Amusement Park* and *Treasure Hunt* narratives.

Table 6.4. Gender differences in game narrative preferences according to gender identity.

Game Narrative	% Boys	% Girls	χ^2 (1)	ϕ

Amusement Park	42.8	50.0	1.72	.07
Treasure Hunt	50.3	56.4	1.23	.06
Help a Sea Friend	32.9	53.2	13.77***	.21
War at Sea	74.0	40.4	38.04***	-.34

Table 6.5. Gender differences in game narrative preferences according to feminine- and masculine-typed occupational interests, activities, and traits.

Game Genre	Feminine-typed				Masculine-typed			
	Pref.	Not P.	*F*	*d*	Pref.	Not P.	*F*	*d*
	M (SD)	*M (SD)*			*M (SD)*	*M (SD)*		
Amusement Park	2.25 (.48)	2.23 (.48)	.18	.05	2.33 (.45)	2.35 (.47)	.13	-.04
Treasure Hunt	2.25 (.46)	2.22 (.50)	.21	.05	2.34 (.43)	2.34 (.50)	.01	.01
Help a Sea Friend	2.36 (.52)	2.13 (.42)	18.07***	.46	2.28 (.50)	2.38 (.42)	4.59*	-.24
War at Sea	2.12 (.44)	2.39 (.49)	27.44***	-.56	2.39 (.46)	2.27 (.45)	5.80*	.27

To test for significant differences in game narrative preferences based on masculine- and feminine-typing, we ran four one-way between-subjects ANOVAs comparing the masculine- and feminine-typed behaviors of students who ranked specific game narratives as their 1st or 2nd option. With regards to feminine-typed occupational interests, activities, and traits, there were medium effects of students who preferred the *Help a Sea Friend* narrative reporting significantly higher feminine-typed behaviors, and those who preferred the *War at Sea* narrative reporting significantly lower feminine-typed behaviors. In addition, there were small effects of students who preferred the *War at Sea* narrative reporting significantly higher masculine-typed behaviors, and those who preferred the *Help a Sea Friend* narrative reporting significantly lower masculine-typed behaviors. See Table 6.5 for descriptive statistics, effect sizes (Cohen's *d*), and *F*-statistics.

Next, we conducted a series of logistic regressions predicting each game narrative preference with binary gender, masculine-typed scale, and feminine-typed scale in a single model (Table 6.6). For the *Help a Sea Friend* narrative, masculine- and feminine-type scales were significant predictors of preference, while binary gender was not. For the *War at Sea* narrative, both binary gender and the feminine-type scale were significant predictors. Models predicting the *Amusement Park* and *Treasure Hunt* narratives were not significant and are excluded.

Table 6.6. Logistic regressions predicting game preferences according to gender identity (female = 0), masculine-typed scale, and feminine-typed scale. (*) *p* < .05, (**) *p* < .01, (***) *p* < .001.

Game Narrative	Constant	Binary gender	Feminine-typed scale	Masculine-typed scale
Help a Sea Friend	*B* = -.82, *Exp(B)* = .44	*B* = -.24, *Exp(B)* = .79	*B* = .98**, *Exp(B)* = 2.67	*B* = -.68*, *Exp(B)* = .51
War at Sea	*B* = .52, *Exp(B)* = 1.69	*B* = .84**, *Exp(B)* = 2.33	*B* = -.82*, *Exp(B)* = .44	*B* = .53, *Exp(B)* = 1.70

Post-hoc Analyses

As a follow-up, we analyzed whether the additional gender dimensions, reflected by the masculine- and feminine-typed scores, would reveal more nuances about students' game preferences. We identified 13 boys who had stronger feminine-typed (*M* = 2.18, *SD* = 0.29) than masculine-typed (*M* = 1.95, *SD* = 0.30) behaviors and 31 girls who had stronger masculine-typed (*M* = 2.52, *SD* = 0.47) than feminine-typed (*M* = 2.31, *SD* = 0.41) behaviors. In each of these groups, the *action* and *sandbox* genres were most popular. Additionally, *Treasure Hunt* was the most preferred narrative among the 13 boys, and *War at Sea* was the most preferred narrative among the 31 girls.

Discussion

In this work, we examined young students' game preferences through the lens of a multidimensional gender framework. Our research was motivated by the need to characterize gender differences in game preferences while considering additional gender dimensions beyond binary gender identity, including gender-typed occupational interests, activities and traits (Hyde et al., 2019; Liben & Bigler, 2002). This topic is especially relevant to the design of inclusive learning games that match students' gaming interests without relying upon vague or outdated gender stereotypes. Beyond reporting students' preferred game genres and narratives, our work also supports the analysis of multiple gender dimensions in educational research. We discuss the insights from these results below.

Overall, we observed consistent gender differences in game preferences when representing gender as binary categories and as continuous gender-typed scores. In particular, boys and those with strong masculine-typed behaviors tended to prefer the *action* and *sports & racing* genres, as well as the battle-oriented game narrative, *War at Sea*. Meanwhile, girls and those with strong feminine-typed behaviors reported more interest in the *casual* and *music & party* genres, and the co-operative game narrative, *Help a Sea Friend*. These patterns are consistent with those reported in past surveys (Aleksić & Ivanović, 2017; Chou & Tsai, 2007), indicating that gender-based game preferences have remained stable over the years.

However, our results suggest two significant advantages of using a multidimensional gender representation over the traditional binary one. First, based on Cohen's interpretation guidelines (Cohen, 2013), we observed gender differences with medium to large effect sizes using gender-typing measures, but only small to medium effect sizes using gender identity, suggesting that the former approach better reflects the influence of gender on children's game interests. This observation is further supported by logistic regression analyses showing that gender-typed scales predicted game preferences better than binary gender identity, when both were included in the same models. In short, a more nuanced assessment of gender is a more powerful predictor of game preferences. Second, our post-hoc analysis revealed distinct preferences from boys with stronger feminine-typed behaviors and from girls with stronger masculine-typed behaviors, compared to other students in their respective gender identity groups. Thus, when the students' gender identity and gender-typed behaviors do not completely align, their gaming interests can be better distinguished by measures of gender-typed interests, activities, and traits. That is, the multidimensional assessment of gender is more precise. Based on these findings, we encourage AIED researchers to adopt multidimensional gender measures in their future studies to more closely examine how students' background factors may influence their learning interest and experience.

The findings from our survey also have implications for learning game design. First, we observed that the *strategy*, *role-playing* and *casual* genres were preferred by only 5-25% of students. Notably, these three genres often feature game elements – such as slow pace, reflection and study of the environment – that are frequently used in learning games (Amory, 2001) because they can foster a playful experience without inducing high cognitive load and interfering with learning. At the same time, they may cause the identified genres to be less favored by students, for several possible reasons. First, many modern games, regardless of genre, tend to incorporate elements that require fast and accurate decision making from the players (Dale et al., 2020). If children are used to this style of game play, games that feature slower and more methodical play may be less engaging. Alternatively, prior work has shown that middle-school students may omit reporting that they enjoy games which do not fit with the gamer identity they are attempting to project (Higashi et al., 2021). Further work is needed to disambiguate these conjectures, but in both cases, increasing children's exposure to slow and reflective entertainment games is likely to improve their reaction to those features in digital learning games.

The remaining four game genres – *action*, *sports and racing*, *sandbox*, *music and party* – were all highly favored by both boys and girls. Among them, the *sandbox* genre was the most universally popular, being included in the top two genres of close to half of the boys and girls in the study. This attribute may be explained by the rapid rise in popularity of sandbox games, most notably *Minecraft* and *Roblox*, during the COVID-19 pandemic (Cowan et al., 2021; Hjorth et al., 2020). These platforms allow players to not only play games created by others, but also to freely design their own activities. In this way, they are effective at fostering player connections (Cowan et al., 2021) and have also been used for educational purposes (Dundon, 2019). Digital learning game researchers can take advantage of this trend by introducing sandbox elements, such as player agency (Ryan et al., 2006), into their games to promote engagement. Incorporating player agency has also been shown to result in better learning efficiency (Harpstead et al., 2019b) and outcomes (Taub et al., 2020) in learning games. Additionally, there are successful examples of full-fledged sandbox digital learning games, such as *Physics*

Playground, which utilize stealth assessment techniques to measure student learning without disrupting the sandbox experience (Shute et al., 2021). In turn, this prior work provides a solid foundation for investigating how digital learning games can take advantage of the sandbox genre's increasing popularity among young students.

One limitation of this study is that students may not have a clear picture of all the game genres and narratives included in the survey, given their brief text descriptions. Providing graphical illustrations of the example games in each genre and the storylines in each narrative would likely help students make more informed selections. At the same time, the findings thus far suggest several novel research directions. First, we could evaluate the role of additional gender dimensions, such as gender typicality (i.e., perceived similarity to other children of one's own and other gender - Martin et al., 2017), in explaining students' game preferences. In addition, future research should investigate the connection between gender differences in game preferences and in learning outcomes, for example by manipulating a learning game's narrative to be either masculine- or feminine-oriented while retaining the game mechanics and instructional materials. This comparison would be particularly interesting for learning games that have been shown to produce gender differences in learning outcomes (Arroyo et al., 2013; Khan et al., 2017; H. A. Nguyen et al., 2022). For instance, the math game *Decimal Point* (B. M. McLaren et al., 2017b), which features an amusement park metaphor, has yielded consistent learning benefits for girls over boys across several studies (Hou et al., 2020a, 2022b; H. A. Nguyen et al., 2022). To follow up, one could investigate the outcomes of an alternate version of the game with a more masculine-oriented narrative, such as *War at Sea* – in this case, would gender differences in learning favoring girls still emerge, or would boys have an advantage instead? More generally, the study of how gender-based preferences interact with the game features and instructional materials is an important step towards understanding the nuanced role of gender in shaping students' playing and learning experience.

Conclusion

In conclusion, our results indicate that a nuanced approach to gender is more predictive of game preferences than binary gender. That is, assessing multiple dimensions of gender is more precise, particularly for students whose activities, interests, and traits are less stereotypical of their binary gender identity. To our knowledge, this is the first research that utilizes a multidimensional gender framework for examining children's game preferences, and it suggests that this approach is more appropriate for game designers seeking to customize learning games to better suit individual students' interests and preferences. Ultimately, we envision that these customizations may also be driven by AI techniques which construct accurate and inclusive models of students' learning and preferences, using our identified multidimensional gender features as a starting point.

7. How Does Gameplay Narrative Impact the Relationship between Gender and Learning in *Decimal Point*?

Digital learning games have been shown to help girls learn more than boys in STEM subjects, which can help bridge the gender gap in U.S. classrooms where girls are traditionally disadvantaged. However, more research is needed to identify which game features induce this effect and how well it generalizes to other learning environments. To this end, the current study examines *Decimal Point*, a digital learning game for decimal number and operations which has shown consistent learning benefits for girls over boys. In particular, we compared *Decimal Point* to a conventional tutoring system, as well as a new learning game, *Ocean Adventure*, which features a more masculine game narrative and environment, based on the survey findings from Chapter 6. Our results showed that, in all three conditions, girls learned more than boys thanks to their better performance in the self-explanation activities. However, girls felt more anxious and were less engaged than boys. In addition, boys experienced more mastery only in the *Ocean Adventure* condition. Through an analysis of students' gender-typed behaviors, we also found that those with stronger feminine-typed behaviors reported higher levels of evaluation apprehension and state anxiety in the tutor, but not in the game. These results showed, yet again, that self-explanation can be a robust driving factor of gender differences in learning outcomes, and that students' gender can influence their engagement with different learning environments and game narratives.

Introduction

Given their ability to engage students and promote learning, digital learning games could potentially reduce girls' anxiety with STEM subjects and help them achieve similar learning outcomes as boys across grade levels. This proposition is supported by recent evidence showing that learning games can be effective for girls, often even more than for boys, in terms of both learning and affective outcomes (Arroyo et al., 2014; Hou et al., 2020; McLaren, Farzan et al., 2017), even if boys tend to spend more time playing games (Homer et al., 2012). In particular, with the learning game *Decimal Point*, which teaches decimal numbers and operations to middle school students (McLaren et al., 2017), our prior work has revealed consistent gender differences in learning outcomes favoring girls across six classroom studies (Chapter 4; Nguyen et al., 2022). In addition, we have identified that the driving factor for this robust effect was girls' better performance than boys in the self-explanation activities, which translated to girls' better learning outcomes post-intervention. However, our studies thus far have focused only on the game *Decimal Point* and an equivalent conventional tutor. To derive more actionable insights for learning game researchers and designers, there remains the need to understand how well the gender differences and self-explanation effect generalize to other game environments. Additionally, as recent works in gender studies research have advocated for a multidimensional representation of gender (Hyde et al., 2019), it would be important to examine how different gender dimensions play a role in shaping students' playing and learning experience.

We address this issue in our current work through the study of a new learning game for decimal numbers called *Ocean Adventure*, which was designed to align with boys' gaming preferences. In particular, following a prior survey study about boys and girls' game preferences (Chapter 7; Nguyen et al., 2023), we have identified that a game narrative centered around fighting pirate leaders to save the world was significantly more popular to boys than to girls. Based on this finding, we have developed the *Ocean Adventure* game following this narrative, while retaining all of the instructional materials from *Decimal Point* to ensure a valid comparison between the two games. Thus, the current study is a randomized controlled experiment with three conditions: the original *Decimal Point* game, the newly developed *Ocean Adventure* game, and a conventional tutor serving as the control condition. In this setting, we aimed to investigate whether and how gender differences in learning or enjoyment manifest in each condition, as well as whether self-explanation continued to explain the relationship between gender and learning outcomes. In particular, our research questions are as follows.

RQ1: *Are there differences in learning outcomes between the original Decimal Point, the masculine game Ocean Adventure, and an equivalent non-game tutor?*
RQ2: *How do students' gender dimensions influence their learning outcomes with the games and the tutor?*
RQ3: *How do students' gender dimensions influence their enjoyment with the games and the tutor?*

Decimal Point, *Decimal Tutor* and *Ocean Adventure*

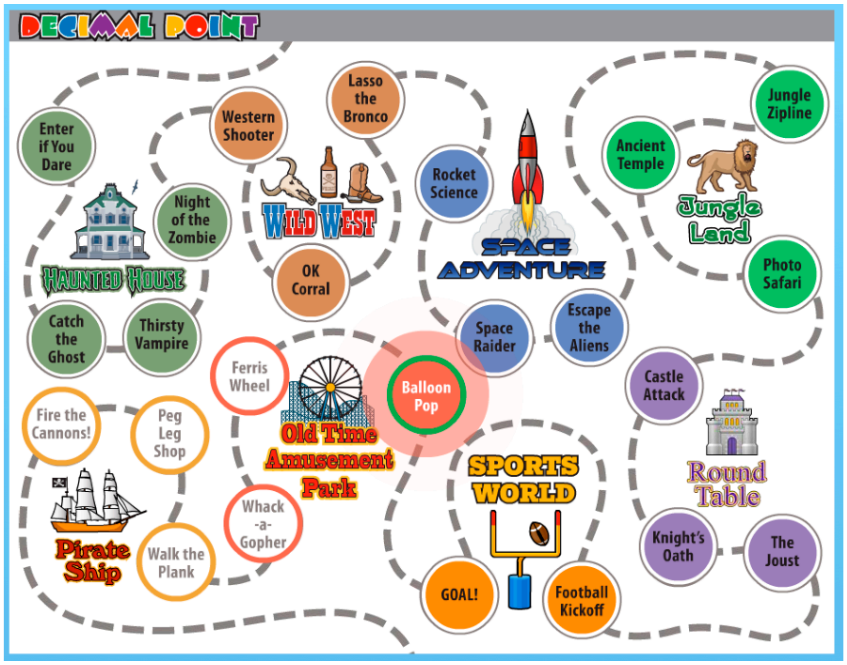
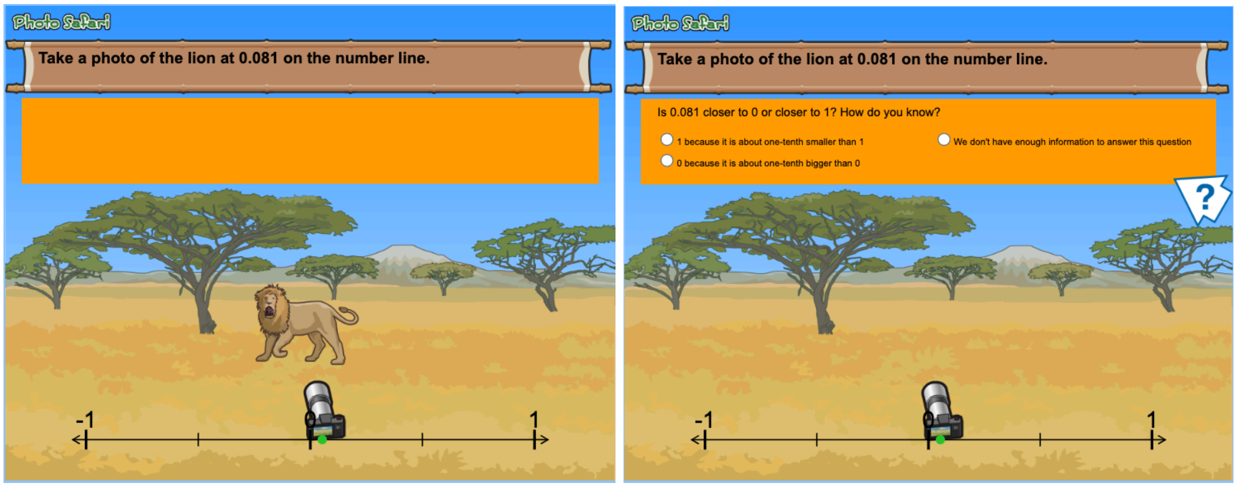


Figure 7.1. The main game map of *Decimal Point* which features an amusement park.



Take a photo of the lion at 0.081 on the number line.

Take a photo of the lion at 0.081 on the number line.

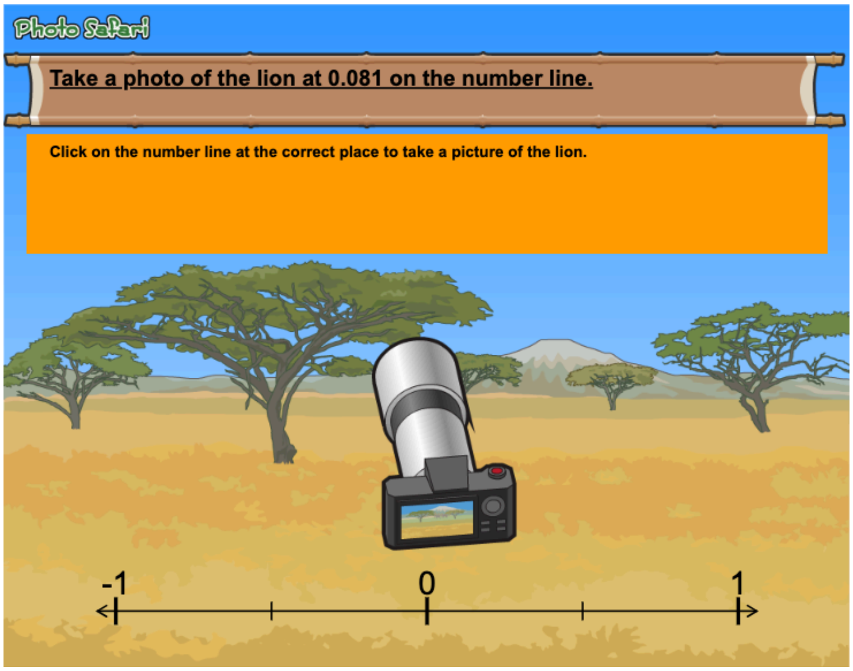
Is 0.081 closer to 0 or closer to 1? How do you know?

○ 1 because it is about one-tenth smaller than 1

○ We don't have enough information to answer this question

○ 0 because it is about one-tenth bigger than 0

Figure 7.2. The problem-solving activity (left) and self-explanation prompt (right) in the number line mini-game *Photo Safari*.



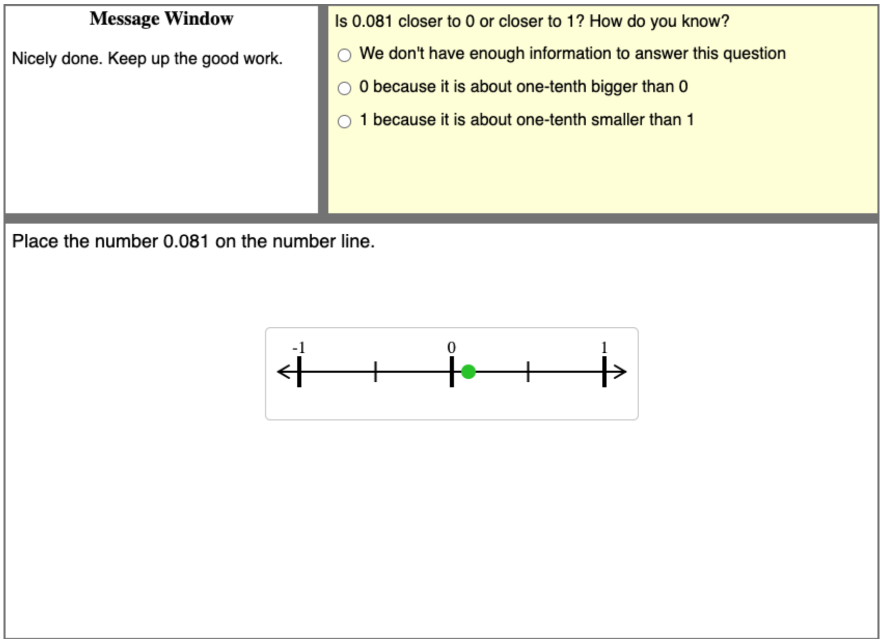
Take a photo of the lion at 0.081 on the number line.

Click on the number line at the correct place to take a picture of the lion.

Figure 7.3. After the student has been stuck for five minutes, the top text in the mini-game becomes underlined to let the experimenter know that they can provide the solution.

Decimal Point is a web-based digital learning game that teaches decimal numbers and operations to middle school students. The game features an amusement park metaphor (Figure 7.1), with 8 theme areas (e.g., *Haunted House*, *Wild West*) and 24 mini-games (e.g., *Night of the Zombies*, *Lasso the Bronco*). The student starts from the upper left corner of the amusement park (at the mini-game *Enter if You Dare*) and advances to the bottom left corner, playing two rounds of each mini-game along the way. Each mini-game round features a problem-solving activity (e.g., sort a sequence of numbers, place a number on the number line) followed by a self-explanation activity, prompting them to explain how they solved the problem-solving question (Figure 7.2). Students get immediate feedback about the correctness of their answer and can make any number of attempts to arrive at the correct answer, which is required for them to move to the next round. To prevent the scenarios where students get stuck at a problem, a subtle visual indicator will appear on their screen after five minutes of game play in each round (Figure 7.3). This indicator will prompt the experimenter to provide the solution to the student so that they could finish the current round.

Decimal Point has been the subject of many classroom studies on various learning game topics. An initial study by McLaren et al. (2017) showed that the game led to better learning outcomes and enjoyment than a conventional tutor with identical instructional materials. Subsequent studies have explored the topics of student agency (Nguyen et al., 2018), indirect control (Harpstead et al., 2019), learning versus enjoyment (Hou et al., 2020), hints and error messages (McLaren et al., 2022a), self-explanation format (McLaren et al., 2022b), and mindfulness interventions (Nguyen et al., 2022). Across these past studies, Nguyen et al. (2022) has uncovered a highly consistent trend of girls learning from the game more than boys, which can be explained by girls' better performance in the self-explanation activities. Following up on this finding, we conducted a 2x2 experiment which manipulates whether students played the game or used a conventional tutor, and whether they performed self-explanation after each problem-solving activity (Chapter 6). Our results showed that self-explanation helped girls learn more than boys, but there were no significant differences in learning outcomes between the game group and the tutor group. This result is inconsistent with the findings from McLaren et al. (2017), possibly due to the higher frequency of hint requests in the game, which led to worse learning outcomes.



Message Window

Nicely done. Keep up the good work.

Is 0.081 closer to 0 or closer to 1? How do you know?

○ We don't have enough information to answer this question

○ 0 because it is about one-tenth bigger than 0

○ 1 because it is about one-tenth smaller than 1

Place the number 0.081 on the number line.

Figure 7.4. An example level in the tutor, corresponding to the mini-game in Figure 7.2.

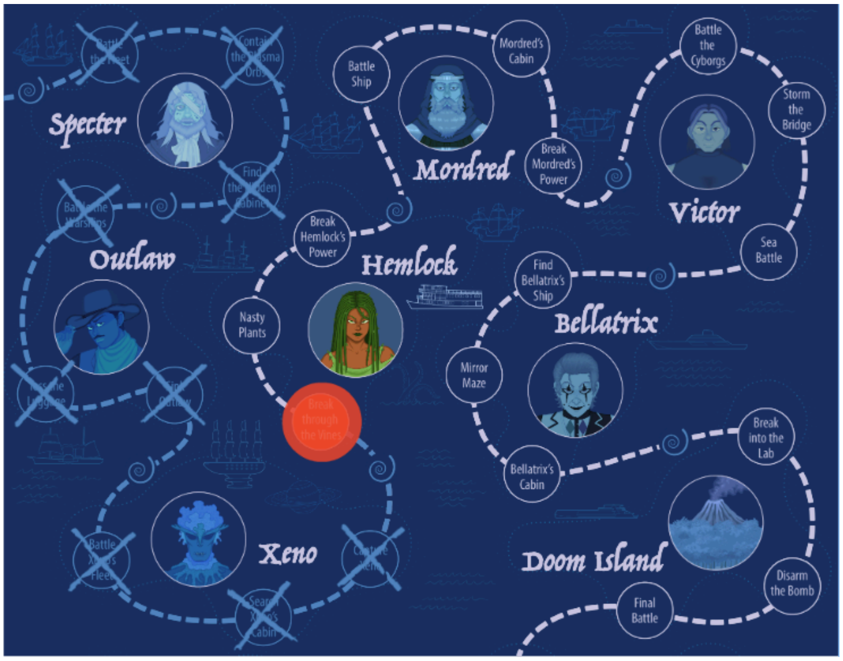
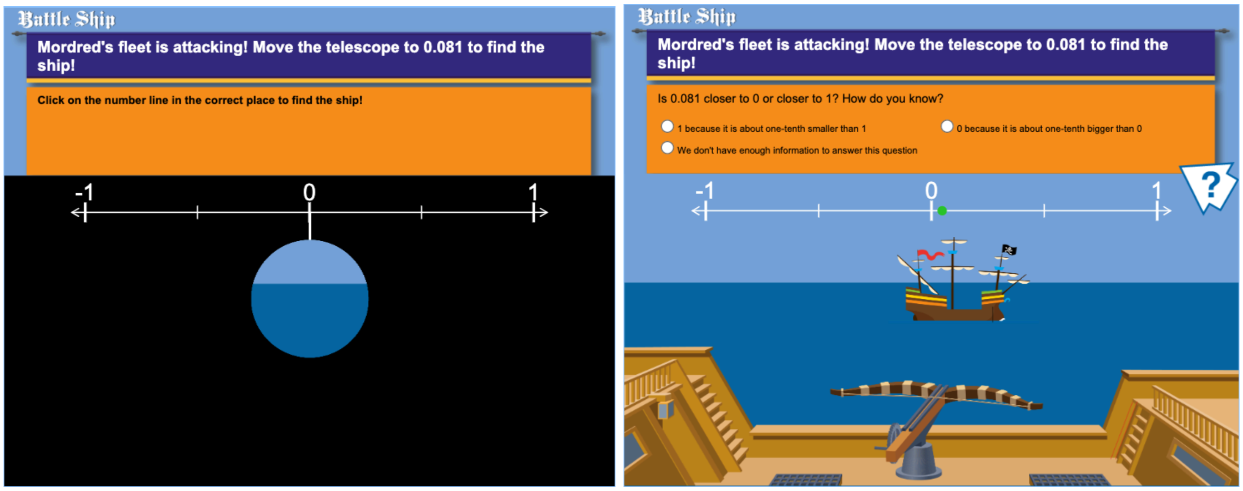


Figure 7.5. The main game map of *Ocean Adventure*, featuring eight pirate hideouts to conquer.



Battle Ship

Mordred's fleet is attacking! Move the telescope to 0.081 to find the ship!

Click on the number line in the correct place to find the ship!

Battle Ship

Mordred's fleet is attacking! Move the telescope to 0.081 to find the ship!

Is 0.081 closer to 0 or closer to 1? How do you know?

○ 1 because it is about one-tenth smaller than 1
○ 0 because it is about one-tenth bigger than 0
○ We don't have enough information to answer this question

Figure 7.6. The problem-solving activity (left) and self-explanation prompt (right) in the number line mini-game *Battle Ship*.

In this study, we aimed to follow up on the game versus tutor comparisons from McLaren et al. (2017) and Chapter 6, while also exploring the role of gender dimensions in shaping students' learning outcomes and enjoyment. To this end, in addition to *Decimal Point*, we utilized a conventional tutoring system and another learning game, Ocean Adventure, which both feature identical learning materials as those in *Decimal Point*. The tutor (Figure 7.4) is a web-based platform which contains a series of problem-solving and self-explanation activities, but without any graphical elements or game narratives. On the other hand, *Ocean Adventure* is a *reskin* of the game *Decimal Point* with a more masculine narrative, taking the player through space and time to combat several pirate leaders and save the world. This narrative was chosen based on the findings from Chapter 7 that a pirate battle theme was highly preferred by boys and those with strong masculine-typed behaviors, as opposed to the original amusement park narrative of *Decimal Point* with no clear gender preferences. Given that there were consistent learning outcome differences favoring girls in prior *Decimal Point* studies, we wanted to investigate how a game narrative aligned with boys' interests might influence these results. Thus, we redesigned the amusement park from Figure 7.1 to be a map of eight pirate hideouts, which the player must conquer (Figure 7.5). Each mini-game was also redesigned to be consistent with the overall narrative, while retaining the instructional materials. For instance, the original mini-game *Photo Safari* from *Decimal Point* (Figure 7.2) was turned into an exercise about aiming ballistas at an approaching enemy ship in *Ocean Adventure* (Figure 7.6). Unlike in the previous 2x2 study (Chapter 6), none of the three learning platforms in this study contained hint messages, to better replicate the game versus tutor comparison in McLaren et al. (2017).

Method

Participants and Design

The study involved 420 students across six elementary and middle schools in a northeastern U.S. city. Over the course of five days, during their regular class times, students went through a pretest, demographic survey, intervention materials, evaluation questionnaire, followed by an immediate posttest. One week after the posttest, students then completed a delayed posttest. To avoid potential distractions that may occur when students sit next to one another but use different learning platforms, students were assigned by classroom to use either one of the two games or the tutor.

81 students were excluded from analysis due to not finishing all of the study materials. Among the remaining 339 students, 122 played the game *Decimal Point*, 90 played the game *Ocean Adventure*, and 127 used the decimal tutor. These students ranged in age from 10 to 12 years old (*M* = 10.75, *SD* = 0.63). In terms of gender identity, 50% (n = 169) of the students identified as male, 48% (n = 164) as female, 0.9% (n = 3) identified as non-binary, and 0.9% (n = 3) preferred not to disclose their gender. Due to the small sample size of the last two categories, we excluded these six students from analyses of gender identity, but still included them in analyses of multidimensional gender scales.

Materials

Students completed all study materials on a web-based learning environment (Aleven et al., 2009). The materials included three versions of the test, a pre-intervention and post-intervention survey, in addition to the two learning platforms mentioned above.

Pretest, Posttest and Delayed Posttest. Each test features 43 items that each range from 1 to 3 points, for a total of 52 points. The test items were designed to either assess the decimal skills and procedures practiced in the intervention materials (e.g., “place 0.4 on a number line from -1 to 1") or probe for high-level conceptual understanding (e.g., "is a longer decimal always larger than a shorter decimal?"). There were three isomorphic versions of the test that were randomly assigned to each student's pretest, posttest and delayed posttest.

Demographic and Gender-Typed Behaviors Survey. Before doing the pretest, students completed a demographic survey asking them about their age, grade level, self-identified gender identity and race. Then, they were assigned a 58-item survey, adapted from the Children's Occupational Interests, Activities, and Traits - Personal Measure (COAT-PM - Liben & Bigler, 2002), which assesses their interests, activities and traits in relation to gender-stereotyped norms. All survey items were labeled as either masculine-typed or feminine-typed, rated on a Likert scale, and covered three domains. The *occupational interests* domain measures the degree of interest in pursuing certain professions, with 18 items rated from 1 (not at all) to 4 (very much). These items include occupations like "hairstylist" or "nursed" (feminine) and "construction worker" or "engineer" (masculine). The *activity* domain evaluates the frequency of engaging in particular activities, with 18 items rated from 1 (never) to 4 (often or very often). Examples of these activities are "making jewelry" or "taking dance lessons" (feminine) and "playing basketball" or "going fishing" (masculine). The *traits* domain gauges self-perceptions of personal characteristics, with 22 items rated from 1 (not at all like me) to 4 (very much like me). These items encompass qualities such as "gentle" or "neat" (feminine) and "adventurous" or "confident" (masculine).

An earlier version of this multidimensional gender survey, with 54 items, has been used in a prior classroom study of *Decimal Point* (H. A. Nguyen et al., 2023). The current survey incorporated four additional items to the *trait* subscale to enhance its reliability, resulting in a total of 58 items. We note that the survey items were chosen to portray *stereotypical* perceptions of gender differences, rather than actual gender differences. Our objective in examining the students' ratings of these items is to measure the extent to which they align their behaviors with conventional gender norms and stereotypes (Liben & Bigler, 2002). To this end, we computed two scales of feminine-typed behaviors ($\alpha = 0.81$) and masculine-typed behaviors ($\alpha = 0.83$) by averaging the corresponding items from all three dimensions. In other words, in addition to their gender identity, each student's gender is represented by a feminine-typed scale and a masculine-typed scale, which are continuous measures ranging from 1 to 4.

Table 7.1. The dimensions of enjoyment covered in the post-intervention questionnaire.

Dimension	Example statement
Affective engagement	I felt frustrated or annoyed.
Behavioral / cognitive engagement	I tried out my ideas to see what would happen.
Experience of meaning	The game [tutor] felt relevant to me.
Experience of mastery	I felt capable while playing the game [learning from the tutor].

Experience of appropriate challenge	The game [tutor] was challenging but not too challenging
Situational interest	The game (tutor) was exciting.
Achievement emotion	I enjoyed the challenge of learning the material.
Evaluation apprehension	If I did poorly on this activity, people would look down on me.
Test self-efficacy	I could handle this activity.
State anxiety	During the activity, I felt very nervous.

Evaluation Questionnaire. After completing the game or the tutor, students were asked to rate several statements about their learning experience on a Likert scale from 1 (strongly disagree) to 5 (strongly agree). These statements cover seven dimensions of enjoyment: multidimensional engagement (Ben-eliyahu et al., 2018) with the affective subscale (3 items, α = .71) and behavioral / cognitive subscale (3 items, α = .45); situational interest (3 items, α = 0.84 - Linnenbrink-Garcia et al., 2010); enjoyment dimension of achievement emotion (6 items, α = 0.91 - Pekrun, 2005); evaluation apprehension (4 items, α = 0.86 - Spencer et al., 1999); test self-efficacy (5 items, α = 0.72 - Spencer et al., 1999); state anxiety (3 items, α = 0.68 - Chung et al., 2010; Veit & Ware, 1983). Table 2 includes example items for each of these constructs. For our analysis, we excluded the behavioral / cognitive engagement subscale due to its low reliability.

Results

First, a series of repeated-measures ANOVA showed a significant difference for all students between pretest and posttest score, $F = 109.35$, $p < .001$, $\eta_p^2 = .244$, as well as between pretest and delayed posttest score, $F = 191.11$, $p < .001$, $\eta_p^2 = .361$. In other words, students' performance improved after learning in all three conditions.

We also examined the correlations between gender dimensions. Our results showed that gender identity, where "female" was coded as 1 and "male" coded as 0, was positively correlated with feminine-typed scale ($r = 0.53$, $p < .001$) and negatively correlated with masculine-typed scale ($r = -0.42$, $p < .001$). In addition, feminine-typed scale was positively correlated with masculine-typed scale ($r = 0.19$, $p < .001$). Given the correlation coefficients, while the three gender dimensions were moderately correlated, they were not redundant.

RQ1: *Are there differences in learning outcomes between the original Decimal Point, the masculine game Ocean Adventure, and an equivalent non-game tutor?*

Table 7.2. Descriptive statistics of students' test performance across the three study conditions, reported in M (SD) format.

Condition	N	Pretest	Postttest	Delayed Posttest
Tutor	127	24.48 (10.64)	27.89 (9.88)	29.17 (10.23)
Decimal Point	122	22.27 (10.32)	25.98 (9.39)	27.17 (9.08)
Ocean Adventure	90	22.63 (10.89)	27.09 (10.25)	27.97 (10.39)

The mean scores on the pretest, immediate posttest and delayed posttest by condition are shown in Table 7.2. We first note that, while students had similar pretest scores across the three conditions (F = 1.53, p = .219, η_p^2 = .008), there were three high-performing classes, with 7 students in the *Decimal Point* condition, 12 students in the Ocean Adventure condition and 44 students in the Tutor condition. A one-way ANOVA showed that the group of students in these three classes (M = 34.54, SD = 9.43) had significantly higher pretest scores than the remaining students (M = 20.61, SD = 9.06), F = 119.46, p < .001, η_p^2 = .262. Because two of these classes were assigned to use the tutor, their high prior knowledge could potentially skew the comparisons of learning by condition. Thus, to provide the full context, we present the results of our learning analyses both with and without the three high-performing classes.

We compared students' posttest and delayed posttest scores by condition using two-way ANCOVAs, using pretest scores as covariate. With the three high-performing classes included, our results showed no significant condition differences at posttest (F = 0.51, p = .599, η_p^2 = .003) and delayed posttest (F = 0.22, p = .803, η_p^2 = .001). Without the three high-performing classes, we identified significant condition differences at posttest (F = 3.46, p = .033, η_p^2 = .025) and at delayed posttest (F = 3.121, p = .046, η_p^2 = .022). Post hoc pairwise comparisons, with Bonferroni corrections, showed that the *Ocean Adventure* condition led to significantly higher posttest scores (t = -2.50, p = 0.04) and marginally higher delayed posttest scores (t = -2.31, p = .06) than the tutor condition.

RQ2: *How do students' gender dimensions influence their learning outcomes with the games and the tutor?*

Table 7.3. Descriptive statistics of test performance by gender identity, reported in M (SD) format.

Gender Identity	N	Pretest	Postttest	Delayed Posttest
Female	164	21.77 (9.70)	25.45 (8.88)	27.21 (9.12)
Male	169	24.83 (11.20)	28.66 (10.37)	29.31 (10.43)

Table 7.3 shows the descriptive statistics for the pretest, posttest and delayed posttest performance between boys and girls. For pretest performance, a one-way ANOVA revealed a significant effect of gender identity (F = 7.08, p = .008, η^2_p = .021), where boys had higher pretest scores than girls. When comparing posttest and delayed posttest performance, we also examined potential interactions between gender identity and condition, through a series of

two-way ANCOVAs with pretest score as covariate. With the three high-performing classes included, our results showed no significant effects of gender identity (*F* = 0.062, *p* = .804, η^2_p < .001) or gender - condition interaction (*F* = 0.282, *p* = .754, η^2_p = .002). On delayed posttest performance, the main effect of gender identity (*F* = 0.060, *p* = .807, η^2_p < .001) was not significant, neither was its interaction with condition (*F* = 0.574, *p* = .564, η^2_p = .004). Similar results were observed when dropping the three high-performing classes.

We also analyzed pretest scores with multiple dimensions of gender. Using a regression model predicting pretest scores based on masculine-typed and feminine-typed scales, we found that masculine-typed scale was a significant and positive predictor (*β* = 2.68, *p* = .047), while feminine-typed scale was a negative predictor (*β* = -2.51, *p* = .087). Another model predicting pretest score based on gender-typed scales and binary gender identity (with “female” coded as 1 and “male” coded as 0) showed that masculine-typed scale (*β* = 1.30, *p* = .442) was a positive predictor, while feminine-typed scale (*β* = -0.73, *p* = .711) and gender identity (*β* = 2.29, *p* = .181) were negative predictors; however, none of the predictors were significant.

Next, we built a regression model with pretest score as covariate and the following predictor variables: masculine-typed scale, feminine-typed scale, condition, and the interaction terms between gender-typed scales and condition. This model was used to predict posttest and delayed posttest performance. With the three high-performing classes included, being in the tutor condition was a significant and positive predictor of posttest scores (*β* = 12.56, *p* = .022). Without the high-performing classes, being in the *Ocean Adventure* condition was a significant and positive predictor of posttest scores (*β* = 13.28, *p* = .037). In both cases, there were no significant predictors of delayed posttest scores.

Following up on previous results about the role of self-explanation in helping girls learn more than boys (Chapter 4 and Chapter 6), we examined self-explanation performance by gender as follows. First, we conducted a two-way ANCOVA examining the effects of gender and condition on the number of self-explanation errors, with pretest scores as covariate. With the high-performing classes included, our result showed a significant main effect of gender identity (*F* = 7.29, *p* = .007, η^2_p = .022), with girls (*M* = 31.87, *SD* = 13.49) making fewer self-explanation errors than boys (*M* = 33.34, *SD* = 13.63). The effects of the condition and its interaction with gender identity were not significant. Without the high-performing classes, we still identified a significant main effect of gender (*F* = 6.42, *p* = .012, η^2_p = .024), with girls (*M* = 35.01, *SD* = 11.53) making fewer errors than boys (*M* = 37.44, *SD* = 10.90). At the same time, there was also a significant main effect of condition (*F* = 3.76, *p* = .024, η^2_p = .028). Post hoc pairwise comparisons, with Bonferroni corrections, showed that students in the tutor condition (*M* = 39.32, SD = 10.51) made significantly more self-explanation errors than those in the *Ocean Adventure* condition (*M* = 33.64, *SD* = 10.67), *t* = 3.38, *p* = .003.

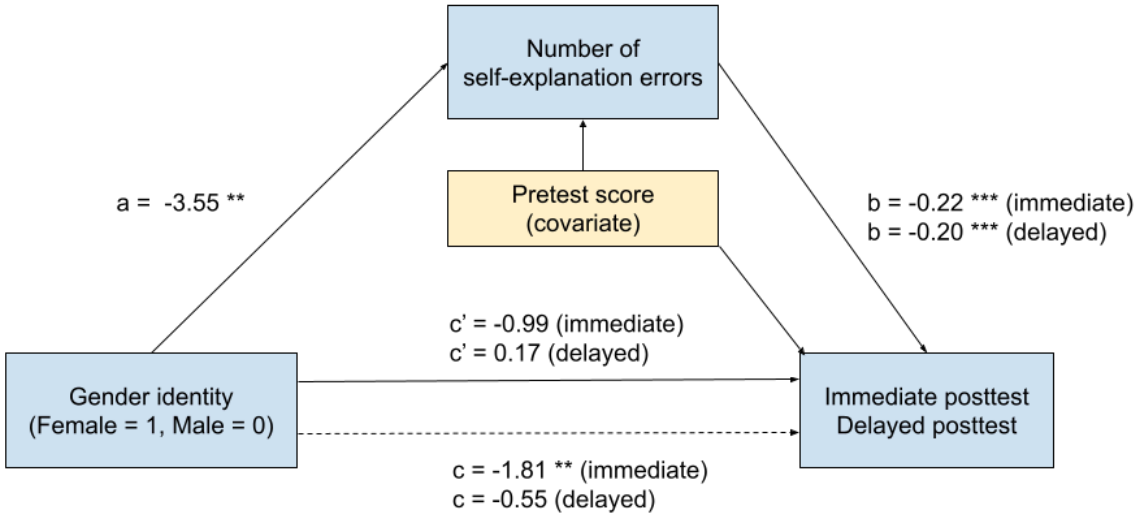


Figure 7.7. Diagram of the mediation pathway from gender identity to posttest and delayed posttest performance, through self-explanation performance. (**) *p* < .01, (***) *p* < .001.

To identify how self-explanation performance may mediate the relationship between gender identity and learning outcomes, we constructed two mediation models with pretest score as covariate, gender identity as an independent variable, self-explanation error as a mediator, and posttest / delayed posttest score as the dependent variable (Figure 7.7). Our results revealed a negative association between gender identity (where “female” was coded as 1) and self-explanation error, as well as between self-explanation error and test performance. Bootstrapping procedures indicated a significant indirect effect of self-explanation performance on the relationship between gender identity and posttest performance (*ab* = 0.81, 95% CI [0.26, 1.53], *p* < .001), as well as between gender identity and delayed posttest performance (*ab* = 0.72, 95% CI [0.25, 1.42], *p* < .001).

RQ3: *How do students’ gender dimensions influence their enjoyment with the games and the tutor?*

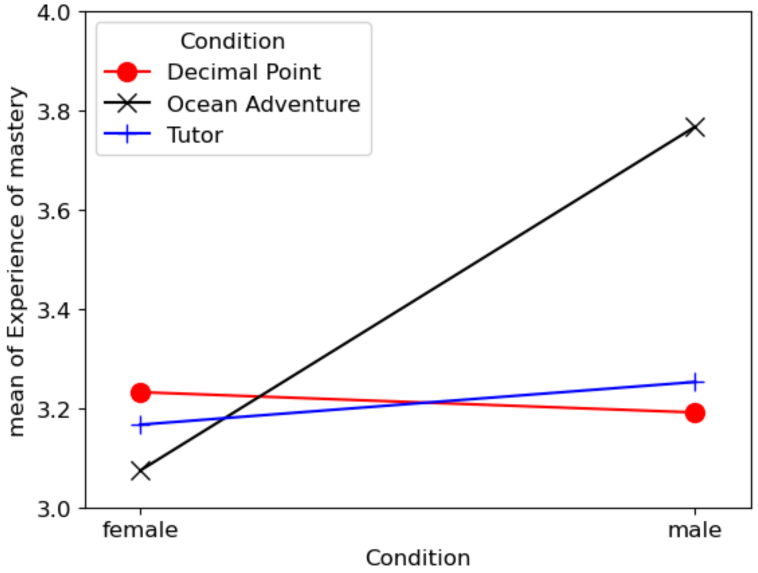


Figure 7.8. The interaction effect between gender identity and condition on experience of mastery.

We first performed a series of two-way ANOVAs assessing the effects of gender identity and condition on each enjoyment dimension in Table 7.1. To examine the engagement hypothesis, we performed a series of two-way ANOVA on the dimensions of affective engagement, situational interest, achievement emotions and player experience. Our results showed that boys (M = 3.26, SD = 1.10) reported higher levels of affective engagement than girls (M = 2.93, SD = 0.95), F = 9.00, p = .003, η^2_p = .027. There was also a significant condition effect on situational interest, F = 6.07, p = .003, η^2_p = .027, with both the *Decimal Point* and *Ocean Adventure* conditions leading to significantly higher levels of situational interest than the tutor condition. For experience of mastery, both the main effect of gender identity (F = 4.24, p = .040, η^2_p = .013) and the gender - condition interaction effect (F = 3.25, p = .040, η^2_p = .020) were significant. Pairwise comparisons showed that boys reported significantly more experience of mastery than girls in the *Ocean Adventure* condition (F = 9.94, p = .002, η^2_p = .104), but not in the other two conditions (Figure 7.8). For the remaining dimensions, there were no significant main or interaction effects

To examine the stereotype threat hypothesis, we investigated the dimensions of evaluation apprehension, test efficacy and state anxiety. Our results showed that girls (M = 2.48, SD = 0.98) reported higher levels of evaluation apprehension than boys (M = 2.11, SD = 0.97), F = 11.21, p = .001, η^2_p = .035. At the same time, girls (M = 3.37, SD = 0.80) reported lower levels of test efficacy than boys (M = 3.73, SD = 0.81), F = 15.284, p < .001, η^2_p = .046. For test anxiety, the main effects of gender identity (F = 14.16, p < .001, η^2_p = .043) and condition (F = 3.36, p = .036, η^2_p = .021) were both significant. In particular, girls (M = 2.86, SD = 0.95) reported higher levels of anxiety than boys (M = 2.45, SD = 1.01), and the *Decimal Point* condition (M = 2.48, SD = 0.95) led to lower levels of test anxiety than the tutor (M = 2.81, SD = 1.03).

To investigate the relationship between gender-typed behaviors and enjoyment measures, we built regression models predicting each of the enjoyment dimensions based on masculine-typed scale, feminine-typed scale and their interactions with the condition. Our results showed that masculine-typed scale was a significant predictor of affective engagement (β = 0.43, p = .043), situational interest (β = 0.51, p = .017), achievement emotion (β = 0.67, p = .002), experience of meaning (β = 0.59, p = .005), experience of mastery (β = 0.43, p = .042), and experience of appropriate challenge (β = 0.41, p = .027). In addition, feminine-typed scale was a significant predictor of situational interest (β = 0.51, p = .044) and experience of mastery (β = 0.60, p = .015). It was also a significant predictor of evaluation apprehension (β = 0.77, p = .018), test efficacy (β = -0.53, p = .051) and state anxiety (β = 0.66, p = .043) in the tutor condition.

Discussion

In this work, we examined how gender dimensions influenced students' learning and enjoyment with two digital learning games, *Decimal Point* and *Ocean Adventure*, and an equivalent conventional tutoring system. While *Decimal Point* features an amusement park narrative that was intended to appeal to all students (Forlizzi et al., 2014), *Ocean Adventure* was specifically designed to align with the interests of boys and those with strong masculine-typed behaviors (Nguyen et al., 2023). To our knowledge, this is the first study that compares two learning games with identical instructional materials but different narratives. In turn, we have identified several ways in which this narrative change interacted with gender dimensions in their influence on students' learning experience, which we discuss in detail as follows.

First, we found no significant differences in posttest and delayed posttest performance across conditions. However, when excluding 63 students in both the game and tutor conditions with very high pretest scores, compared to the general student population in the current study, we found that both game conditions led to better test performance than the tutor, with a significant difference between *Ocean Adventure* and the tutor. In addition, the differences in posttest and delayed posttest performance between the *Ocean Adventure* condition and the tutor condition were significant. This result is consistent with the advantages of the game over the tutor in promoting learning, as shown in McLaren et al. (2017). It also supports the conjecture from Chapter 6 that the gap between the game and the tutor was diminished when hints were introduced, as students playing the game tended to request hints more frequently, leading to worse learning outcomes. In terms of enjoyment, we found that the game conditions led to higher levels of situational interest and lower levels of state anxiety than the tutor. Overall, these results reinforce the advantages of digital learning games in promoting both learning and enjoyment over traditional instruction (Mayer, 2019). At the same time, they indicate that certain types of instructional scaffolds, such as hint messages, may not be effective in digital learning games, due to the hints potentially interrupting immersion and leading to unintended player behaviors. However, given the demonstrated benefits of hints in improving students' cognitive and metacognitive skills (Aleven et al., 2016), future research should continue to investigate how hints can be best incorporated into learning game contexts.

In terms of gender differences in learning, our results are consistent with those from prior studies in Chapter 4 and 6. Girls performed worse than boys at pretest, but this gender gap was erased in the posttest and delayed posttest, across all three conditions. Additionally, self-explanation performance continued to mediate the relationship between gender identity and learning outcomes, with girls making fewer self-explanation errors and therefore learning more in both the games and the tutor. As we discussed in earlier chapters, this effect could be due to girls' better language skills, which helped them interpret the self-explanation options better than boys (Nikolaenko, 2005; Stevenson et al., 2009). Alternatively, it could result from boys being more disengaged with the self-explanation activities because of their similarity to traditional instruction. Future research could incorporate survey measures of engagement during the self-explanation activities to better understand the cognitive and affective processes which take place as students perform self-explanation.

We also identified gender differences in enjoyment across several dimensions. In particular, compared to girls, boys reported higher levels of affective engagement across conditions, as well as more experience of mastery in the *Ocean Adventure* condition. These results imply that the engagement hypothesis was not supported – while only *Ocean Adventure* contained game features that aligned with boys' interest, boys were still more engaged than girls in all three conditions. On the other hand, boys reported lower levels of evaluation apprehension and test anxiety, in addition to higher levels of test efficacy, across conditions. These results are in line with the stereotype threat hypothesis, which posits that girls feel more anxious about learning math than boys. However, as these gender differences were observed after the intervention, they suggest that the game environments were not able to reduce girls' experience of stereotype threat. Given that girls were less engaged and felt more anxious than boys after the intervention, the gender differences in learning favoring girls could be attributed primarily to their better self-explanation performance. As self-explanation is an established and robust instructional technique for promoting deep learning and transfer (Richey & Nokes-Malach, 2015; Wylie & Chi, 2014), it is not surprising that girls would learn more when they outperformed boys in self-explanation. However, the finding that self-explanation activities remain effective in a learning game context, even if they are not highly gamified, is a novel and encouraging finding. This suggests that, as a first step, self-explanation prompts can be incorporated into learning games in other domains to promote girls' learning. Future work should also investigate ways to make the self-explanation activity more playful, so as not to break the game's immersion and to potentially help boys remain engaged.

Our analyses of multiple gender dimensions – i.e., gender-typed occupational interests, activities and traits (Liben & Bigler, 2002) further revealed several nuances to the gender differences in learning and enjoyment. First, similar to our finding from Chapter 6, gender identity was not a significant predictor of pretest scores, when controlled for masculine-typed and feminine-typed scales. This suggests that gender-typed behaviors are more predictive of math performance than gender identity alone. In addition, we found that masculine-typed scale was a significant and positive predictor of affective engagement, situational interest and achievement emotion, as well as experience of meaning, mastery and appropriate challenge. At the same time, feminine-typed scale was a significant and positive predictor of situational interest and experience of mastery. Notably, higher feminine-typed scale could predict higher evaluation apprehension and state anxiety, as well as lower test efficacy, but only in the tutor condition. In other words, the game conditions were more effective than the conventional tutor, which resembles traditional math instruction, in reducing the experience of stereotype threat for students with strong feminine-typed behaviors. While this effect did not influence learning outcomes as much as the self-explanation effect, it still provides evidence for the benefit of learning games in making learning content more accessible to marginalized student populations.

Conclusion

In conclusion, this research reveals important insights into the role of gender dimensions and the game environment in shaping students' experience with digital learning games. We replicated earlier results from Chapter 4 and 6, showing that girls learned more from the game than boys, thanks to girls' better performance in the self-explanation activities. With regards to the two main hypotheses, we again found that the engagement hypothesis was not supported, as boys were more engaged than girls overall. On the other hand, our multidimensional gender analysis provided evidence for the stereotype threat hypothesis; in particular, the game environment could reduce the experience of stereotype threat for students with strong feminine-typed behaviors more effectively than a conventional tutor. These findings yield important lessons into the design of inclusive and effective digital learning games, which we summarize in the next chapter.

8. Summary and Contributions

There is an established gender gap in middle school math education, where girls are disadvantaged in advanced areas of math, while also having lower confidence and higher math anxiety, which could lead to lower interest in math careers in the long term. Digital learning games may be well suited to address this issue, given their potential to engage young learners and the established learning benefits of modern learning games. However, the game industry remains male-dominated (Clement, 2021a), despite ongoing efforts to promote diversity and combat racism (Cole & Zammit, 2020; Hackney, 2017). With learning games specifically, designers often rely on intuition, rather than empirical guidance, for how to make games effective for different populations of learners, which could lead to reinforcing, rather than reducing, existing stereotypes. Furthermore, prior work on gender differences in game preferences and in learning with games has focused exclusively on distinctions between boys and girls, without considering additional gender dimensions that have been advocated by gender studies research. The limited perspective from the binary gender classification could result in alienating students whose gender identities do not fit the male - female dichotomy and creating biases in the data analysis or student modeling process. Towards addressing these issues, this thesis work examines why and how digital learning games can lead to gender differences in learning outcomes through seven classroom studies of the learning game *Decimal Point*, with over 1600 middle school students. The findings from each study and their contributions are summarized as follows.

In **Chapter 4**, I have reported on a consistent trend of girls having higher learning gains from *Decimal Point* than boys across five classroom studies. While there has been prior evidence of learning games benefitting girls more than boys (Arroyo et al., 2014; Khan et al., 2017; Klisch et al., 2012; Tsai, 2017), *Decimal Point* is the first game where this effect remains robust through several years of study. In addition, I have identified self-explanation as the driving factor behind the gender effects – girls made significantly fewer errors than boys in the self-explanation activities and therefore learned more from the game. Altogether, these results provide strong support for the use of digital learning games in promoting learning and bridging the gender gap in traditional classrooms. They also point to the potential of self-explanation in helping students, especially girls, achieve greater learning outcomes in a learning game context. This result is notable, as self-explanation prompts have rarely been employed in digital learning games, despite their established benefits in learning science literature (Chi et al., 1994; Wylie & Chi, 2014; Chi & Wylie, 2014). Our findings suggest that even self-explanation prompts in simple formats, such as multiple choice questions, can already help girls to catch up to boys in learning after game play. We expect that, when these activities are designed to be more playful (e.g., self-explaining to a non-player learning companion) and to foster more active learning (e.g., with a focused or open-ended format - McLaren et al., 2022), their effectiveness in bridging the gender gap will be more pronounced.

In **Chapter 5**, I conducted a follow-up study on the impact of gender and self-explanation in learning with the game *Decimal Point* and with a conventional tutor. We again saw that girls had greater learning gains post-intervention; notably, this effect was present in both the game and the tutor. Contrary to our initial hypotheses, the gender differences in learning outcomes favoring girls were not driven by the game environment, as we found that boys reported more engagement and girls reported more anxiety even after game play. Instead, girls' better performance in the self-explanation activities continued to explain their greater learning outcomes. In addition, our study was the first investigation of multiple gender dimensions in a learning game context, following recent trends in gender studies research which advocate for a more multi-faceted representation of gender beyond the male versus female dichotomy (Hyde et al., 2019). In particular, using the COAT-PM measures of gender-typed behaviors (Liben & Bigler, 2022), we found that these measures can predict learning and enjoyment better than binary gender. Notably, masculine-typed behavior was a significant predictor of enjoyment and engagement, while feminine-typed behavior was a significant predictor of anxiety and evaluation apprehension. Utilizing gender-typed measures not only provides a more nuanced understanding of the role of gender, but also allows for the inclusion of all students, including those who did not identify as male or female and would have been excluded from analyses of binary gender due to their small sample size.

In **Chapter 6**, I investigated students' preferences of game genres and game narratives, as well as how these preferences differ by gender. While prior survey studies have explored similar topics (Chun & Tsai, 2007; Dele-Ajayi et al., 2018; Arroyo et al., 2013), our work points to a novel finding that sandbox games, such as *Minecraft* and *Roblox*, have become highly popular to young children after the pandemic. This finding suggests that digital learning games should utilize features of sandbox games, such as allowing players a high degree of agency, to effectively engage young students. In terms of gender differences, our findings are consistent with those from past literature – boys had a preference for action and sports games, while girls were more interested in casual and music games. At the same time, through the use of multiple gender dimensions (Liben & Bigler, 2002), we again observed that gender-typed behaviors were more predictive of students' game preferences than binary gender identity and could reveal distinctions in game preferences among students with the same gender identity. We also identified the game narrative of fighting pirate leaders to save the world as one which was significantly preferred by boys and those with strong masculine-typed behaviors. This narrative is important for our next study of whether gender differences favoring girls still manifest in a game environment which is more aligned with masculine preferences.

In **Chapter 7**, I reported on a randomized controlled experiment with *Decimal Point*, a new game *Ocean Adventure* with a masculine narrative, and a conventional tutor. Our results showed that the game conditions, especially the newly developed *Ocean Adventure*, led to better learning and higher levels of enjoyment than the tutor, similar to the original comparison between the game and the tutor in McLaren et al. (2017). We also observed the effect of changing the game narrative, as boys reported higher levels of mastery experience than girls in the *Ocean Adventure* condition. However, there were no differences in the effects of *Decimal Point* and *Ocean Adventure* on learning or enjoyment. Instead, across all conditions, girls were able to catch up with boys post-intervention thanks to their better performance in the self-explanation activities, consistent with results from past studies. Another finding we replicated was that masculine-typed behavior positively predicted students' enjoyment. Notably, we also found that feminine-typed behavior positively predicted anxiety and evaluation apprehension only in the tutor, suggesting that the game conditions were able to reduce the experience of stereotype threat for students with strong feminine-typed behavior.

The above findings about gender and games, which were consistently replicated across seven classroom studies from 2017 to 2023, yielded several lessons for designing and researching learning games to bridge the gender gap in mathematics. First, self-explanation activities are easily integrated into learning games and can effectively help girls achieve better learning outcomes, while not hurting boys' performance. While this effect was observed in both the game *Decimal Point* and a conventional tutor, the game was more beneficial overall as it led to higher engagement and lower anxiety for all students. Second, representing gender as a multidimensional construct, comprising not just gender identity but also gender-typed behaviors, can allow for more inclusive data analysis and yield a more nuanced understanding of how gender impacts learning and enjoyment. Third, switching to a masculine game narrative increased boys' engagement but did not impact girls', suggesting that adapting games to specific gender preferences can help promote student engagement. However, to avoid subjecting students to gender stereotypes, this decision should be based on empirical data about students' game preferences. To this end, the survey introduced in Chapter 6 can serve as an initial template for researchers to explore gender-based game preferences with their own student population. Finally, we should note that the relationship between gender and game-based learning is a highly complex topic, due to the dynamic nature of gender and the constantly evolving learning technologies. Thus, we encourage conducting replication studies, in the same way *Decimal Point* studies were carried out, to identify robust gender effects that can inform data-driven design decisions and contribute to the development of equitable learning games.

As digital games become increasingly popular and accessible, the potential of games in education is more promising than ever. However, building games for a larger population of learners will require a more nuanced understanding of their backgrounds and perspectives. The study of human learning, much like that of machine learning, may run the risk of catering to the majority while alienating the minority, if care is not taken to preserve equity and inclusiveness (Mehrabi et al., 2021; Srinivasan & Chander, 2021). A meta-analysis of recent research in educational data mining, for example, has revealed that only 15% out of 385 reviewed papers considered students' demographic data in their analysis and predictive modeling (Paquette et al., 2020). In this context, I consider my thesis work, which advocates for extending beyond the traditional binary view of gender in games and learning games research, as a building block towards the development of more personalized and inclusive learning platforms.

References

Abeele, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies*, *135*, 102370.

Adamo-Villani, N., Wilbur, R., & Wasburn, M. (2008). Gender differences in usability and enjoyment of VR educational games: A study of SMILE$^{\text{TM}}$. *2008 International Conference Visualisation*, 114–119.

Adams, D. M., & Clark, D. B. (2014). Integrating self-explanation functionality into a complex game environment: Keeping gaming in motion. *Computers & Education*, *73*, 149–159.

Adams, R. B., & Kirchmaier, T. (2016). Women on boards in finance and STEM industries. *American Economic Review*, *106*(5), 277–281.

Aleksić, V., & Ivanović, M. (2017). Early adolescent gender and multiple intelligences profiles as predictors of digital gameplay preferences. *Croatian Journal of Education: Hrvatski Časopis Za Odgoj i Obrazovanje*, *19*(3), 697–727.

Aleven, V. A. W. M. M., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer‐based Cognitive Tutor. *Cognitive Science*, *26*(2), 147–179. https://doi.org/10.1207/s15516709cog2602_1

Aleven, V., McLaren, B. M., & Sewall, J. (2009a). Scaling up programming by demonstration for intelligent tutoring systems development: An open-access web site for middle school mathematics learning. *IEEE Transactions on Learning Technologies*, *2*(2), 64–78.

Aleven, V., McLaren, B. M., & Sewall, J. (2009b). Scaling up programming by demonstration for intelligent tutoring systems development: An open-access web site for middle school mathematics learning. *IEEE Transactions on Learning Technologies*, *2*(2), 64–78.

Aleven, V., McLaren, B. M., Sewall, J., Van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K. R. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers. *International Journal of Artificial Intelligence in Education*, *26*(1), 224–269.

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.

Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems. *International Journal of Artificial Intelligence in Education*, *26*(1), 205–223. https://doi.org/10.1007/s40593-015-0089-1

Amory, A. (2001). Building an educational adventure game: Theory, design, and lessons. *Journal of Interactive Learning Research*, *12*(2), 249–263.

Arroyo, I., Burleson, W., Tai, M., Muldner, K., & Woolf, B. P. (2013). Gender differences in the use and benefit of advanced learning technologies for mathematics. *Journal of Educational Psychology*, *105*(4), 957.

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*(4), 1052–1092.

Baram-Tsabari, A., & Yarden, A. (2011). Quantifying the gender gap in science interests. *International Journal of Science and Mathematics Education*, *9*(3), 523–550.

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.

Beemyn, G., Rankin, S. R., Park, P., Crawford, L., Keja, V., Chen, J., Beauchamp, T., Burke, N. B., Aizura, A. Z., & Enriquez, M. C. (2016). *Trans studies: The challenge to hetero/homo normativities*. Rutgers University Press.

Ben-Eliyahu, A., Moore, D., Dorph, R., & Schunn, C. D. (2018). Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Educational Psychology*, *53*, 87–105.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children’s interests. *Science*, *355*(6323), 389–391.

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: A systematic review. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 41–50.

Bouvier, P., Lavoué, E., Sehaba, K., & George, S. (2013). Identifying learner's engagement in learning games: A qualitative approach based on learner's traces of interaction. *5th International Conference on Computer Supported Education (CSEDU 2013)*, 339–350.

Bragg, S., Renold, E., Ringrose, J., & Jackson, C. (2018). 'More than boy, girl, male, female': Exploring young people's views on gender diversity within and beyond school contexts. *Sex Education*, *18*(4), 420–434.

Breda, T., Jouini, E., & Napp, C. (2018). Societal inequalities amplify gender gaps in math. *Science*, *359*(6381), 1219–1220.

Brinkman, B. G., Rabenstein, K. L., Rosén, L. A., & Zimmerman, T. S. (2014). Children's gender identity development: The dynamic negotiation process between conformity and authenticity. *Youth & Society*, *46*(6), 835–852.

Brockmyer, J. H., Fox, C. M., Curtiss, K. A., McBroom, E., Burkhart, K. M., & Pidruzny, J. N. (2009). The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology*, *45*(4), 624–634.

Bull, S. (2020). There Are Open Learner Models About! *IEEE Transactions on Learning Technologies*.

Burkett, I. (2012). An introduction to co-design. *Sydney: Knode*, 12.

Cameron, J. J., & Stinson, D. A. (2019). Gender (mis) measurement: Guidelines for respecting gender diversity in psychological research. *Social and Personality Psychology Compass*, *13*(11), e12506.

Chang, M., Evans, M., Kim, S., Deater-Deckard, K., & Norton, A. (2014). Educational video games and Students' game engagement. *2014 International Conference on Information*

Science & Applications (ICISA), 1–3.

Chapman, J. R., & Rich, P. J. (2018). Does educational gamification improve students' motivation? If so, which game elements work best? *Journal of Education for Business*, *93*(7), 315–322.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182.

Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243.

Chou, C., & Tsai, M.-J. (2007). Gender differences in Taiwan high school students' computer game playing. *Computers in Human Behavior*, *23*(1), 812–824.

Chung, B. G., Ehrhart, M. G., Holcombe Ehrhart, K., Hattrup, K., & Solamon, J. (2010). Stereotype threat, state anxiety, and specific self-efficacy as predictors of promotion exam performance. *Group & Organization Management*, *35*(1), 77–107.

Chung, L.-Y., & Chang, R.-C. (2017). The effect of gender on motivation and student achievement in digital game-based learning: A case study of a contented-based classroom. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(6), 2309–2327.

Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have gender gaps in math closed? Achievement, teacher perceptions, and learning behaviors across two ECLS-K cohorts. *AERA Open*, *2*(4), 2332858416673617.

Clark, D. B., Nelson, B. C., Chang, H.-Y., Martinez-Garza, M., Slack, K., & D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United

States. *Computers & Education*, *57*(3), 2178–2195.

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research*, *86*(1), 79–122.

Clement, J. (2021a). Distribution of game developers worldwide from 2014 to 2021, by gender. *Statista: Www. Statista. Com/Statistics/453634/Game-Developer-Gender-Distribution-Worldwide/Adresinden Alındı*.

Clement, J. (2021b). Percentage of teenagers who play video games in the United States as of April 2018, by gender. *Statista*.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cole, A., & Zammit, J. (2020). *Cooperative gaming: Diversity in the games industry and how to cultivate inclusion*. CrC Press. https://books.google.com/books?hl=en&lr=&id=0XjtDwAAQBAJ&oi=fnd&pg=PP1&dq=promoting+diversity+in+the+gaming+industry&ots=4h08GBiGF0&sig=z0e4Arj1npTdy4wEEcL7ZkVjxRU

Conati, C., Jaques, N., & Muir, M. (2013). Understanding attention to adaptive hints in educational games: An eye-tracking study. *International Journal of Artificial Intelligence in Education*, *23*(1), 136–161.

Connolly, T., Stansfield, M., & Boyle, L. (2009). *Games-Based Learning Advancements for Multi-Sensory Human Computer Interfaces: Techniques and Effective Practices: Techniques and Effective Practices*. IGI Global.

Cook, R. E., Nielson, M. G., Martin, C. L., & DeLay, D. (2019). Early adolescent gender development: The differential effects of felt pressure from parents, peers, and the self. *Journal of Youth and Adolescence*, *48*(10), 1912–1923.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning:

Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715.

Cowan, K., Potter, J., Olusoga, Y., Bannister, C., Bishop, J. C., Cannon, M., & Signorelli, V. (2021). Children’s Digital Play during the COVID-19 Pandemic: Insights from the Play Observatory. *Je-LKS: Journal of e-Learning and Knowledge Society*, *17*(3), 8–17.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience* (Vol. 1990). Harper & Row New York.

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math–gender stereotypes in elementary school children. *Child Development*, *82*(3), 766–779.

Dale, G., Kattner, F., Bavelier, D., & Green, C. S. (2020). Cognitive abilities of action video game and role-playing video game players: Data from a massive open online course. *Psychology of Popular Media*, *9*(3), 347.

Dele-Ajayi, O., Strachan, R., Pickard, A., & Sanderson, J. (2018). Designing for All: Exploring Gender Diversity and Engagement with Digital Educational Games by Young People. *2018 IEEE Frontiers in Education Conference (FIE)*, 1–9.

Deterding, S. (2016). Contextual autonomy support in video game play: A grounded theory. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3931–3943.

D’Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, *29*, 153–170.

Dorji, U., Panjaburee, P., & Srisawasdi, N. (2015). Gender differences in students’ learning achievements and awareness through residence energy saving game-based inquiry playing. *Journal of Computers in Education*, *2*(2), 227–243.

Doyle, R. A., & Voyer, D. (2016). Stereotype manipulation effects on math and spatial test performance: A meta-analysis. *Learning and Individual Differences*, *47*, 103–116.

Dreger, A. D., & Herndon, A. M. (2009). Progress and politics in the intersex rights movement:

Feminist theory in action. *GLQ: A Journal of Lesbian and Gay Studies*, *15*(2), 199–224.

Drey, T., Jansen, P., Fischbach, F., Frommel, J., & Rukzio, E. (2020). Towards progress assessment for adaptive hints in educational virtual reality games. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9.

Dundon, R. (2019). *Teaching social skills to children with autism using Minecraft®: A step by step guide*. Jessica Kingsley Publishers.

Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (2017). Using tutors to improve educational games: A cognitive game for policy argument. *Journal of the Learning Sciences*, *26*(2), 226–276.

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education*, *13*(3), 478–492.

Egan, S. K., & Perry, D. G. (2001). Gender identity: A multidimensional analysis with implications for psychosocial adjustment. *Developmental Psychology*, *37*(4), 451.

Ellison, G., & Swanson, A. (2023). Dynamics of the gender gap in high math achievement. *Journal of Human Resources*, *58*(5), 1679–1711.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103.

Else-Quest, N. M., Mineo, C. C., & Higgins, A. (2013). Math and science attitudes and achievement at the intersection of gender and ethnicity. *Psychology of Women Quarterly*, *37*(3), 293–309.

Everett, A., Soderman, B., deWinter, J., Kocurek, C., Huntemann, N. B., Trepanier-Jobin, G., Chien, I., Murray, S., Hutchinson, R., & Patti, L. (2017). *Gaming representation: Race, gender, and sexuality in video games*. Indiana University Press.

Farrell, D., & Moffat, D. C. (2014). Adapting cognitive walkthrough to support game based learning design. *International Journal of Game-Based Learning (IJGBL)*, *4*(3), 23–34.

Fast, A. A., & Olson, K. R. (2018). Gender development in transgender preschool children. *Child Development*, *89*(2), 620–637.

Forlizzi, J., McLaren, B. M., Ganoe, C., McLaren, P. B., Kihumba, G., & Lister, K. (2014). Decimal point: Designing and developing an educational game to teach decimals to middle school students. *8th European Conference on Games-Based Learning: ECGBL2014*, 128–135.

Galdi, S., Cadinu, M., & Tomasetto, C. (2014). The roots of stereotype threat: When automatic associations disrupt girls' math performance. *Child Development*, *85*(1), 250–263.

GameTree Team. (2019, November 5). Industry Results: Genre and Platform Preferences (Age & Gender). *GameTree Blog*. https://gametree.me/blog/global-gamer-insights-report/

Ganley, C. M., & Lubienski, S. T. (2016). Mathematics confidence, interest, and performance: Examining gender patterns and reciprocal relations. *Learning and Individual Differences*, *47*, 182–193.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, *1*(1), 20–20.

Gilbert, N. (2021). *Number of Gamers Worldwide 2021/2022: Demographics, Statistics, and Predictions*. https://financesonline.com/number-of-gamers-worldwide/

Glasgow, R., Ragan, G., Fields, W. M., Reys, R., & Wasman, D. (2000). The decimal dilemma. *Teaching Children Mathematics*, *7*(2), 89–89.

Gödöllei Lappalainen, A. F. (2017). *Game-Based Assessments of Cognitive Ability: Validity and Effects on Adverse Impact through Perceived Stereotype Threat, Test-Taking Motivation and Anxiety* [Master's Thesis]. Graduate Studies.

Goldman, A. D., & Penner, A. M. (2016). Exploring international gender differences in mathematics self-concept. *International Journal of Adolescence and Youth*, *21*(4), 403–418. https://doi.org/10.1080/02673843.2013.847850

Greenberg, B. S., Sherry, J., Lachlan, K., Lucas, K., & Holmstrom, A. (2010). Orientations to video games among gender and age groups. *Simulation & Gaming*, *41*(2), 238–259.

Gülgöz, S., Glazier, J. J., Enright, E. A., Alonso, D. J., Durwood, L. J., Fast, A. A., Lowe, R., Ji, C., Heer, J., & Martin, C. L. (2019). Similarity in transgender and cisgender children's gender development. *Proceedings of the National Academy of Sciences*, *116*(49), 24480–24485.

Hackney, E. (2017). Eliminating racism and the diversity gap in the video game industry. *J. Marshall L. Rev.*, *51*, 863.

Hamari, J., & Keronen, L. (2017). Why do people play games? A meta-analysis. *International Journal of Information Management*, *37*(3), 125–141.

Harpstead, E., Richey, J. E., Nguyen, H., & McLaren, B. M. (2019a). Exploring the Subtleties of Agency and Indirect Control in Digital Learning Games. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 121–129.

Harpstead, E., Richey, J. E., Nguyen, H., & McLaren, B. M. (2019b). Exploring the subtleties of agency and indirect control in digital learning games. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 121–129.

Hayes, A. F., & Rockwood, N. J. (2017). Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. *Behaviour Research and Therapy*, *98*, 39–57.

Higashi, R., Harpstead, E., Solyst, J., Kemper, J., Odili Uchidiuno, J., & Hammer, J. (2021). The Design of Co-Robotic Games for Computer Science Education. *Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play*, 111–116.

Hill, F., Mammarella, I. C., Devine, A., Caviola, S., Passolunghi, M. C., & Sz\Hucs, D. (2016). Maths anxiety in primary and secondary school students: Gender differences, developmental changes and anxiety specificity. *Learning and Individual Differences*, *48*, 45–53.

Hjorth, L., Richardson, I., Davies, H., & Balmford, W. (2020). Playing During COVID-19. In

Exploring Minecraft (pp. 167–182). Springer.

Holmes, J. R., To, A., Zhang, F., Park, S. E., Ali, S., Bai, Z., Kaufman, G., & Hammer, J. (2019). A good scare: Leveraging game theming and narrative to impact player experience. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.

Holstein, K., & Doroudi, S. (2022). Equity and Artificial Intelligence in education. In *The Ethics of Artificial Intelligence in Education* (pp. 151–173). Routledge.

Homer, B. D., Hayward, E. O., Frye, J., & Plass, J. L. (2012). Gender and player characteristics in video game play of preadolescents. *Computers in Human Behavior*, *28*(5), 1782–1789.

Honey, M. A., & Hilton, M. L. (2011). Learning science through computer games. *National Academies Press, Washington, DC*.

Hou, X., Nguyen, H. A., Richey, J. E., Harpstead, E., Hammer, J., & McLaren, B. M. (2021). Assessing the Effects of Open Models of Learning and Enjoyment in a Digital Learning Game. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00250-6

Hou, X., Nguyen, H. A., Richey, J. E., Harpstead, E., Hammer, J., & McLaren, B. M. (2022a). Assessing the effects of open models of learning and enjoyment in a digital learning game. *International Journal of Artificial Intelligence in Education*, *32*(1), 120–150.

Hou, X., Nguyen, H. A., Richey, J. E., Harpstead, E., Hammer, J., & McLaren, B. M. (2022b). Assessing the Effects of Open Models of Learning and Enjoyment in a Digital Learning Game. *International Journal of Artificial Intelligence in Education*, 1–31.

Hou, X., Nguyen, H. A., Richey, J. E., & McLaren, B. M. (2020a). Exploring how gender and enjoyment impact learning in a digital learning game. *International Conference on Artificial Intelligence in Education*, 255–268.

Hou, X., Nguyen, H., Richey, J. E., & McLaren, B. M. (2020b). *Exploring How Gender and*

Enjoyment Impact Learning in a Digital Learning Game.

Hsu, C.-Y., & Tsai, C.-C. (2011). Investigating the impact of integrating self-explanation into an educational game: A pilot study. *International Conference on Technologies for E-Learning and Digital Entertainment*, 250–254.

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, *28*(1), 1–35.

Huang, X., Zhang, J., & Hudson, L. (2019). Impact of math self-efficacy, math anxiety, and growth mindset on math and science career interest for middle school students: The gender moderating effect. *European Journal of Psychology of Education*, *34*(3), 621–640.

Hussein, M. H., Ow, S. H., Elaish, M. M., & Jensen, E. O. (2021). Digital game-based learning in K-12 mathematics education: A systematic literature review. *Education and Information Technologies*, 1–33.

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, *74*(2), 171.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, *107*(2), 139.

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*(5888), 494–495.

Irwin, K. C. (2001). Using everyday knowledge of decimals to enhance understanding. *Journal for Research in Mathematics Education*, *32*(4), 399–420.

Isotani, S., McLaren, B. M., & Altman, M. (2010a). Towards intelligent tutoring with erroneous examples: A taxonomy of decimal misconceptions. *International Conference on Intelligent Tutoring Systems*, 346–348.

Isotani, S., McLaren, B. M., & Altman, M. (2010b). Towards intelligent tutoring with erroneous examples: A taxonomy of decimal misconceptions. *International Conference on Intelligent Tutoring Systems*, 346–348.

Johnson, C. I., & Mayer, R. E. (2010). Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. *Computers in Human Behavior*, *26*(6), 1246–1252.

Joiner, R., Iacovides, J., Owen, M., Gavin, C., Clibbery, S., Darling, J., & Drew, B. (2011). Digital games, gender and learning in engineering: Do females benefit as much as males? *Journal of Science Education and Technology*, *20*(2), 178–185.

Joo, H., Lee, J., & Kim, D. (2020). Advancing the design of self-explanation prompts for complex problem-solving. *International Journal of Learning, Teaching and Educational Research*, *19*(11), 88–108.

Keller, L., Preckel, F., Eccles, J. S., & Brunner, M. (2022). Top-performing math students in 82 countries: An integrative data analysis of gender differences in achievement, achievement profiles, and achievement motivation. *Journal of Educational Psychology*, *114*(5), 966.

Khan, A., Ahmad, F. H., & Malik, M. M. (2017). Use of digital game based learning and gamification in secondary school science: The effect on student engagement, learning and gender difference. *Education and Information Technologies*, *22*(6), 2767–2804.

Killingsworth, S. S., Clark, D. B., & Adams, D. M. (2015). Self-explanation and explanatory feedback in games: Individual differences, gameplay, and learning. *International Journal of Education in Mathematics, Science and Technology*, *3*(3), 162–186.

Kinzie, M. B., & Joseph, D. R. (2008). Gender differences in game activity preferences of middle school children: Implications for educational game design. *Educational Technology Research and Development*, *56*(5–6), 643–663.

Klisch, Y., Miller, L. M., Wang, S., & Epstein, J. (2012). The impact of a science education game on students' learning and perception of inhalants as body pollutants. *Journal of Science*

Education and Technology, *21*(2), 295–303.

Koedinger, K. R., Baker, R. Sj., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining*, *43*, 43–56.

Law, E. L.-C. (2010). Learning efficacy of digital educational games: The role of gender and culture. *EdMedia+ Innovate Learning*, 3124–3133.

Lawson, A. P., & Mayer, R. E. (2021). Benefits of Writing an Explanation During Pauses in Multimedia Lessons. *Educational Psychology Review*, *33*(4), 1859–1885. https://doi.org/10.1007/s10648-021-09594-w

Lee, M. J., & Chiou, J. (2020). Animated hints help novices complete more levels in an educational programming game. *Journal of Computing Sciences in Colleges*, *35*(8).

Lehman, B., D'Mello, S., Strain, A., Mills, C., Gross, M., Dobbins, A., Wallace, P., Millis, K., & Graesser, A. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education*, *22*(1–2), 85–105.

Levine, S. C., & Pantoja, N. (2021). Development of children's math attitudes: Gender differences, key socializers, and intervention approaches. *Developmental Review*, *62*, 100997.

Lewis, W. (2010). Serious use of a serious game for language learning. *International Journal of Artificial Intelligence in Education*, *20*(2), 175–195.

Liben, L. S., & Bigler, R. S. (2002). *The development course of gender differentiation*. Blackwell publishing.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123.

Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement*, *70*(4), 647–671.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research*.

Lobel, A., Engels, R. C., Stone, L. L., Burk, W. J., & Granic, I. (2017). Video gaming and children's psychosocial wellbeing: A longitudinal study. *Journal of Youth and Adolescence*, *46*(4), 884–897.

Manero, B., Torrente, J., Fernandez-Vara, C., & Fernandez-Manjon, B. (2016). Investigating the impact of gaming habits, gender, and age on the effectiveness of an educational video game: An exploratory study. *IEEE Transactions on Learning Technologies*, *10*(2), 236–246.

Martin, C. L., Andrews, N. C., England, D. E., Zosuls, K., & Ruble, D. N. (2017). A dual identity approach for conceptualizing and measuring children's gender identity. *Child Development*, *88*(1), 167–182.

Mayer, R. E. (2014). *Computer games for learning: An evidence-based approach*. MIT Press.

Mayer, R. E. (2019). Computer games in education. *Annual Review of Psychology*, *70*, 531–549.

Mayer, R. E., & Johnson, C. I. (2010). Adding instructional features that promote learning in a game-like environment. *Journal of Educational Computing Research*, *42*(3), 241–265.

McLaren, B., Farzan, R., Adams, D., Mayer, R., & Forlizzi, J. (2017). Uncovering gender and problem difficulty effects in learning with an educational game. *International Conference on Artificial Intelligence in Education*, 540–543.

McLaren, B. M., Adams, D. M., Mayer, R. E., & Forlizzi, J. (2017a). A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning (IJGBL)*, *7*(1), 36–56.

McLaren, B. M., Adams, D. M., Mayer, R. E., & Forlizzi, J. (2017b). A computer-based game that promotes mathematics learning more than a conventional approach. *International Journal of Game-Based Learning (IJGBL)*, *7*(1), 36–56.

McLaren, B. M., & Nguyen, H. A. (2023). Digital learning games in artificial intelligence in education (AIED): A review. *Handbook of Artificial Intelligence in Education*, 440–484.

McLaren, B. M., Richey, J. E., Nguyen, H. A., & Mogessie, M. (2022a). Focused Self-Explanations Lead to the Best Learning Outcomes in a Digital Learning Game. *Proceedings of the 17th International Conference of the Learning Sciences*, 1229–1232.

McLaren, B. M., Richey, J. E., Nguyen, H. A., & Mogessie, M. (2022b). A Digital Learning Game for Mathematics that Leads to Better Learning Outcomes for Female Students: Further Evidence. *ECGBL 2022 16th European Conference on Game-Based Learning*.

McLaren, B. M., Richey, J. E., Nguyen, H., & Hou, X. (2022). How instructional context can impact learning with educational technology: Lessons from a study with a digital learning game. *Computers & Education*, *178*, 104366.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Meinck, S., & Brese, F. (2019). Trends in gender gaps: Using 20 years of evidence from TIMSS. *Large-Scale Assessments in Education*, *7*(1), 1–23.

Mejía-Rodríguez, A. M., Luyten, H., & Meelissen, M. R. M. (2021). Gender Differences in Mathematics Self-concept Across the World: An Exploration of Student and Parent Data of TIMSS 2015. *International Journal of Science and Mathematics Education*, *19*(6), 1229–1250. https://doi.org/10.1007/s10763-020-10100-x

Melero, J., Hern’ndez-Leo, D., & Blat, J. (2012). Considerations for the design of mini-games integrating hints for puzzle solving ICT-related concepts. *2012 IEEE 12th International Conference on Advanced Learning Technologies*, 154–158.

Milovanović, I. (2020). Math anxiety, math achievement and math motivation in high school students: Gender effects. *Croatian Journal Educational/Hrvatski Casopis Za Odgoj I*

Obrazovanje, *22*(1). https://www.researchgate.net/profile/Ilija-Milovanovic/publication/341509254_Math_Anxiety_Math_Achievement_and_Math_Motivation_in_High_School_Students_Gender_Effects/links/5ec4eaea92851c11a8779a12/Math-Anxiety-Math-Achievement-and-Math-Motivation-in-High-School-Students-Gender-Effects.pdf

Namkung, J. M., Peng, P., & Lin, X. (2019). The relation between mathematics anxiety and mathematics performance among school-aged students: A meta-analysis. *Review of Educational Research*, *89*(3), 459–496.

Nguyen, H. A., Else-Quest, N., Richey, J. E., Hammer, J., Di, S., & McLaren, B. M. (2023). Gender Differences in Learning Game Preferences: Results Using a Multi-dimensional Gender Framework. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (Vol. 13916, pp. 553–564). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_45

Nguyen, H. A., Hou, X., Richey, J. E., & McLaren, B. M. (2022). The impact of gender in learning with games: A consistent effect in a math learning game. *International Journal of Game-Based Learning (IJGBL)*, *12*(1), 1–29.

Nguyen, H., Harpstead, E., Wang, Y., & McLaren, B. M. (2018). Student Agency and Game-Based Learning: A Study Comparing Low and High Agency. *International Conference on Artificial Intelligence in Education*, 338–351.

Nguyen, H.-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314.

Nicholson, S. (2013). *Two paths to motivation through game design elements: Reward-based gamification and meaningful gamification. iConference 2013 Proceedings*.

Nicholson, S. (2012). *A User-Centered Theoretical Framework for Meaningful Gamification*. Games+Learning+Society 8.0, Madison, WI.

Nikolaenko, N. N. (2005). Sex differences and activity of the left and right brain hemispheres. *Journal of Evolutionary Biochemistry and Physiology*, *41*(6), 689–699.

NPD. (2019). *Evolution of Entertainment Study*. The NPD Group. https://igda-website.s3.us-east-2.amazonaws.com/wp-content/uploads/2019/10/16161928/NPD-2019-Evolution-of-Entertainment-Whitepaper.pdf

Ochsenfeld, F. (2016). Preferences, constraints, and the process of sex segregation in college majors: A choice analysis. *Social Science Research*, *56*, 117–132.

Ogan, A., Walker, E., Baker, R., Rodrigo, M., Mercedes, T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*, *25*(2), 229–248.

Olson, K. R., & Gülgöz, S. (2018). Early findings from the transyouth project: Gender development in transgender children. *Child Development Perspectives*, *12*(2), 93–97.

O'Neil, H. F., Chung, G. K., Kerr, D., Vendlinski, T. P., Buschang, R. E., & Mayer, R. E. (2014). Adding self-explanation prompts to an educational computer game. *Computers in Human Behavior*, *30*, 23–28.

O'Rourke, E., Ballweber, C., & Popovií, Z. (2014). Hint systems may negatively impact performance in educational games. *Proceedings of the First ACM Conference on Learning@ Scale Conference*, 51–60.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education*, *52*(1), 1–12.

Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., & Baker, R. (2020). Who's Learning? Using Demographics in EDM Research. *Journal of Educational Data Mining*, *12*(3), 1–30.

Passolunghi, M. C., Ferreira, T. I. R., & Tomasetto, C. (2014). Math–gender stereotypes and math-related beliefs in childhood and early adolescence. *Learning and Individual Differences*, *34*, 70–76.

Pekrun, R. (2005). Progress and open problems in educational emotion research. *Learning and Instruction*, *15*(5), 497–506.

Perry, D. G., Pauletti, R. E., & Cooper, P. J. (2019). Gender identity in childhood: A review of the literature. *International Journal of Behavioral Development*, *43*(4), 289–304.

Pezzullo, L. G., Wiggins, J. B., Frankosky, M. H., Min, W., Boyer, K. E., Mott, B. W., Wiebe, E. N., & Lester, J. C. (2017). “Thanks Alisha, Keep in Touch”: Gender Effects and Engagement with Virtual Learning Companions. *International Conference on Artificial Intelligence in Education*, 299–310.

Picho, K., Rodriguez, A., & Finnie, L. (2013). Exploring the moderating role of context on the mathematics performance of females under stereotype threat: A meta-analysis. *The Journal of Social Psychology*, *153*(3), 299–333.

Prensky, M. (2006). *Don’t bother me, Mom, I’m learning!: How computer and video games are preparing your kids for 21st century success and how you can help!* Paragon house St. Paul.

Punaro, L., & Reeve, R. (2012). Relationships between 9-year-olds’ math and literacy worries and academic abilities. *Child Development Research*, *2012*.

Rae, J. R., Gülgöz, S., Durwood, L., DeMeules, M., Lowe, R., Lindquist, G., & Olson, K. R. (2019). Predicting early-childhood gender transitions. *Psychological Science*, *30*(5), 669–681.

Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2013). Math anxiety, working memory, and math achievement in early elementary school. *Journal of Cognition and Development*, *14*(2), 187–202.

Read, J. C., & MacFarlane, S. (2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. *Proceedings of the 2006 Conference on Interaction Design and Children*, 81–88.

Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2019). Gender achievement gaps in US school districts. *American Educational Research Journal*, *56*(6), 2474–2508.

Reeve, J., Nix, G., & Hamm, D. (2003). Testing models of the experience of self-determination in intrinsic motivation and the conundrum of choice. *Journal of Educational Psychology*, *95*(2), 375.

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of National Assessment of Educational Progress assessments. *Journal of Educational Psychology*, *107*(3), 645.

Reis, E. (2007). Divergence or disorder?: The politics of naming intersex. *Perspectives in Biology and Medicine*, *50*(4), 535–543.

Renold, E., Bragg, S., Jackson, C., & Ringrose, J. (2017). *How gender matters to children and young people living in England*. Cardiff University.

Richey, J. E., & Nokes-Malach, T. J. (2015). Comparing four instructional techniques for promoting robust knowledge. *Educational Psychology Review*, *27*(1), 181–218.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268–302.

Rodriguez, S., Regueiro, B., Piñeiro, I., Estévez, I., & Valle, A. (2020). Gender differences in mathematics motivation: Differential effects on performance in primary education. *Frontiers in Psychology*, *10*, 3050.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2007). Designing for metacognition—Applying cognitive tutor principles to the tutoring of help seeking. *Metacognition and Learning*, *2*, 125–140.

Romrell, D. (2014). Gender and gaming: A literature review. *Annual Meeting of the AECT International Convention, Hyatt Regency Orange County, Anaheim, CA*, 170–182.

Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in Zoombinis puzzle gameplay. *Computers in Human Behavior*, *120*, 106707.

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 344–360.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31.

Shaffer, D. W., & Gee, J. P. (2006). *How computer games help children learn*. Springer.

Shaw, A. (2015). *Gaming at the edge: Sexuality and gender at the margins of gamer culture*. U of Minnesota Press.

Shute, V., Ke, F., Almond, R. G., Rahimi, S., Smith, G., & Lu, X. (2019). How to increase learning while not decreasing the fun in educational games. *Learning Science: Theory, Research, and Practice*, 327–357.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1), 127–141.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology*, *64*(2), 489–528.

Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education*, *82*, 378–392.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28.

Squire, K., & Jenkins, H. (2003). Harnessing the power of games in education. *Insight*, *3*(1), 5–33.

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of the ACM*, *64*(8), 44–49.

Stacey, K., Helme, S., & Steinle, V. (2001). Confusions between decimals, fractions and negative numbers: A consequence of the mirror as a conceptual metaphor in three different ways. *PME CONFERENCE*, *4*, 4–217.

Starr, C. R., & Simpkins, S. D. (2021). High school students' math and science gender stereotypes: Relations with their STEM outcomes and socializers' stereotypes. *Social Psychology of Education*, *24*, 273–298.

Steiner, C. M., Kickmeier-Rust, M. D., & Albert, D. (2009). Little big difference: Gender aspects and gender-based adaptation in educational games. *International Conference on Technologies for E-Learning and Digital Entertainment*, 150–161.

Stevenson, C. E., Resing, W. C., & Froma, M. N. (2009). Analogical reasoning skill acquisition with self-explanation in 7-8 year olds: Does feedback help? *Educational and Child Psychology*, *26*(3), 6.

Stryker, S. (2017). *Transgender history: The roots of today's revolution*. Hachette UK.

Takeuchi, L. M., & Vaala, S. (2014). Level up Learning: A National Survey on Teaching with Digital Games. *Joan Ganz Cooney Center at Sesame Workshop*.

Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2020). The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education*, *147*, 103781.

Tobias, S., & Fletcher, J. D. (2007). What research has to say about designing computer games for learning. *Educational Technology*, 20–29.

Tsai, F.-H. (2017). An investigation of gender differences in a game-based learning environment with different game modes. *Eurasia Journal of Mathematics, Science and Technology*

Education, *13*(7), 3209–3226.

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

VanLehn, K. (2016). Regulative loops, step loops and task loops. *International Journal of Artificial Intelligence in Education*, *26*(1), 107–112.

Vasilyeva, M., Casey, B. M., Dearing, E., & Ganley, C. M. (2009). Measurement skills in low-income elementary school students: Exploring the nature of gender differences. *Cognition and Instruction*, *27*(4), 401–428.

Veit, C. T., & Ware, J. E. (1983). Mental health inventory. *Psychological Assessment*.

Wai, J., Cacchio, M., Putallaz, M., & Makel, M. C. (2010). Sex differences in the right tail of cognitive abilities: A 30 year examination. *Intelligence*, *38*(4), 412–423.

Walsh, G. (2009). Wii can do it: Using co-design for creating an instructional game. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp. 4693–4698).

Wang, M.-T., & Degol, J. L. (2017). Gender gap in science, technology, engineering, and mathematics (STEM): Current knowledge, implications for practice, policy, and future directions. *Educational Psychology Review*, *29*(1), 119–140.

Wylie, R., & Chi, M. T. (2014). The Self-Explanation Principle in Multimedia Learning. *The Cambridge Handbook of Multimedia Learning*, 413.

Xu, Z., Wijekumar, K., Ramirez, G., Hu, X., & Irey, R. (2019). The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology*, *50*(6), 3119–3137.

Yee, N. (2017, January 19). *Beyond 50/50: Breaking Down The Percentage of Female Gamers by Genre*. https://quanticfoundry.com/2017/01/19/female-gamers-by-genre

Young, C. B., Wu, S. S., & Menon, V. (2012). The neurodevelopmental basis of math anxiety. *Psychological Science*, *23*(5), 492–501.

Zhu, S., Zhuang, Y., Lee, P., Li, J. C.-M., & Wong, P. W. (2021). Leisure and problem gaming behaviors among children and adolescents during school closures caused by COVID-19 in Hong Kong: Quantitative cross-sectional survey study. *JMIR Serious Games*, *9*(2), e26808.

Appendix

Table A.1: Example test items in test form A and their assigned level of learning transfer.

Level of transfer	**Question content**
Near	Select the largest number: 0.22, 0.31, 0.9
Near	Select the smallest number: 0.236, 0.14, 0.6
Near	Enter the next number in the sequence: 0.201, 0.401, 0.601, 0.801, ___
Near	Order the following numbers from smallest to largest: 0.7, 0, 1.0, 0.35
Near	Which list shows decimal numbers ordered from largest to smallest? ● 0.4, 0.8, 0.22, 0.61 ● 0.22, 0.4, 0.61, 0.8 ● 0.8, 0.61, 0.4, 0.22 ● 0.8, 0.4, 0.22, 0.61
Middle	Calculate the sum: 0.2 + 0.4 + 0.9
Middle	Calculate the sum: 0.387 + 0.05
Middle	Calculate the difference: 0.92 - 0.2
Middle	Calculate the difference: 0.4 - 0.004
Middle	Which of the following numbers is closest to 2.8? 2.6, 2.78, 2.81, 2.88888
Far	Is a longer decimal number always larger than a short decimal number?
Far	Is a decimal number that starts with 0 smaller than 0?
Far	Should you separately add the left and right sides, with no carrying across the decimal point?
Far	Is 786 / 987 smaller than zero, equal to zero, or greater than zero?
Far	Which of these two decimals is larger: 0.XY or 0.Y? (Note: X and Y can be 1 through 9) ● 0.XY is always larger

	● 0.Y is always larger ● Depends on what digits X and Y stand for ● Don’t know

Table A.2: Evaluation survey items used in the Fall 2017 study.

Enjoyment factor	Rating items
Lesson enjoyment	I liked doing this lesson. I would like to do more lessons like this. The material in this lesson was difficult for me. I worked hard on understanding the material in this lesson.
Attitude towards math	The lesson made me feel more like I am good at math. The lesson made me feel that math is fun.
Ease of interface use	I liked the way the material was presented on the screen. I liked the way the computer responded to my input. I could easily understand the assignment. I think the interface of the system was confusing. It was easy to enter my answer into the system.

Table A.3: Evaluation survey items used in the Spring 2018 study.

Enjoyment factor	Rating items
Enjoyment of content	I liked doing this lesson. I would like to do more lessons like this. The lesson made me feel more like I am good at math. The lesson made me feel that math is fun.
Enjoyment of interface	I liked the way the material was presented on the screen. I liked the way the computer responded to my input. I think the interface of the system was confusing. It was easy to enter my answer into the system.

Table A.4: Evaluation survey items used in the Fall 2019 study.

Enjoyment factor	Rating items
Multidimensional engagement (Ben-Eliyahu et al., 2018)	I felt frustrated or annoyed. I felt tired or sad. I felt bored. I thought about how idaes in the activity related to other things. I explained things to others. I tried out my ideas to see what would happen.
Game engagement (Brockmyer et al., 2009)	I lost track of time. If someone talked to me, I didn’t hear them. My thoughts moved fast. Playing made me feel calm. I got into the game.

Enjoyment dimension of achievement emotions (Pekrun, 2005)	I looked forward to playing Decimal Point. I enjoyed the challenge of learning the material. I enjoyed acquiring new knowledge. I enjoyed dealing with the game material. Reflecting on my progress in the game made me happy. I played more than required because I enjoyed it so much.

Table A.5: Evaluation survey items used in the Spring 2020 and Spring 2021 studies.

Enjoyment factor	Rating items
Multidimensional engagement (Ben-Eliyahu et al., 2018)	I felt frustrated or annoyed. I felt tired or sad. I felt bored. I thought about how idaes in the activity related to other things. I explained things to others. I tried out my ideas to see what would happen.
Player Experience Inventory (Abeele et al., 2020)	Playing the game was meaningful to me. The game felt relevant to me. Playing the game was valuable to me. I felt capable while playing the game. I felt I was good at playing this game. I felt a sense of mastery playing this game. The game was challenging but not too challenging. The game was not too easy and not too hard to play. The challenges in the game were at the right level of difficulty for me.
Enjoyment dimension of achievement emotions (Pekrun, 2005)	I looked forward to playing Decimal Point. I enjoyed the challenge of learning the material. I enjoyed acquiring new knowledge. I enjoyed dealing with the game material. Reflecting on my progress in the game made me happy.

Table A.6: Demographic survey with multidimensional gender questions in the Fall 2022 and Spring 2023 studies.

DEMOGRAPHIC SURVEY

Gender: ……………………

Grade:

- 5th
- 6th
- 7th
- 8th

Age:

- Less than 10
- 10
- 11
- 12
- 13
- 14

- 15
- More than 15

How would you describe your race or ethnicity? Check as many as apply to you.

- Asian or Pacific Islander
- Black or African American
- Hispanic or Latino
- Native American or Alaskan Native
- White or Caucasian
- A race / ethnicity not listed here
- I’m not sure / Prefer not to say

What I want to be.
Here is a list of jobs that people can do. Please select the option that shows how much you would want to do each of these jobs.

HOW MUCH WOULD YOU WANT TO BE A(N)	**Not At All**	**Not Much**	**Some**	**Very Much**
artist	1	2	3	4
YouTuber	1	2	3	4
professional athlete	1	2	3	4
librarian	1	2	3	4
elementary school teacher	1	2	3	4
secretary	1	2	3	4
nurse	1	2	3	4
police officer	1	2	3	4
doctor	1	2	3	4
hair stylist	1	2	3	4
construction worker	1	2	3	4
scientist	1	2	3	4
computer builder	1	2	3	4
architect	1	2	3	4
dental assistant	1	2	3	4

engineer	1	2	3	4
interior decorator	1	2	3	4
florist (arrange & sell flowers)	1	2	3	4
What I do in my free time. Here is a list of activities that people do. Please select the option that shows how often you do each of these activities.				
HOW OFTEN DO YOU	**Never**	**Rarely**	**Sometimes**	**Often or Very Often**
make jewelry	1	2	3	4
go fishing	1	2	3	4
wash clothes	1	2	3	4
fix or wash a car	1	2	3	4
take dance lessons/classes	1	2	3	4
cook dinner	1	2	3	4
shoot pool	1	2	3	4
jump rope	1	2	3	4
practice an instrument	1	2	3	4
watch sports on tv	1	2	3	4
do gymnastics	1	2	3	4
play dodgeball	1	2	3	4
play computer/video games	1	2	3	4
baby-sit	1	2	3	4
hunt	1	2	3	4
play basketball	1	2	3	4
bake cookies	1	2	3	4
draw (or design) cars/rockets	1	2	3	4

What I am like.
Here is a list of words and phrases that describe people. Please select the option that shows how much each of the words or phrases describes you.

IS THIS LIKE YOU?	**Not At All Like Me**	**Not Much Like Me**	**Somewhat Like Me**	**Very Much Like Me**
emotional (express feelings)	1	2	3	4
aggressive	1	2	3	4
talkative	1	2	3	4
adventurous	1	2	3	4
competitive	1	2	3	4
good at science	1	2	3	4
confident (sure of yourself)	1	2	3	4
enjoy physical education (gym)	1	2	3	4
logical	1	2	3	4
good at math	1	2	3	4
follow directions	1	2	3	4
has good manners	1	2	3	4
try to look good	1	2	3	4
friendly	1	2	3	4
gentle	1	2	3	4
good at social studies	1	2	3	4
neat	1	2	3	4
helpful	1	2	3	4

Table A.7: The game survey used in the Fall 2022 study.

Question 1:
Select three game genres that you enjoy the most. Then rank them from most liked to least liked.

Action
(ex: Fortnite, Splatoon)

Sports or Racing
(ex: Rocket League, FIFA)

Strategy
(ex: Age of Empires, Civilization)

Sandbox
(ex: Roblox, Minecraft)

Music & Party
(ex: Pianista, Just Dance)

Role-playing Game
(ex: Stardew Valley, Legend of Zelda)

Casual
(ex: Bejeweled, Animal Crossing)

Question 2:
Please tell us **2 or more games** you have played recently on your phone/tablet, computer, or video game console.
If you don't play computer games or have played fewer than 2, you can type "I don't play computer games" or "I only play [game name]".
Press Enter (or Return on iPad) to submit. **Your answer needs to contain at least 10 characters.**

……
……
……
……

Question 3:
Which of the following digital game ideas sound most interesting and appealing to you? Rank your preferences from most interesting to least interesting.

Treasure Hunt
You stumble upon a treasure map! Complete puzzles at ocean landmarks to solve this mystery. But beware! Your arch-nemesis will stop at nothing to get to the treasure before you do!

War at Sea
Criminal naval masterminds are planning on taking over the world. It's up to you, the world's top secret agent, to stop them! Fight the fleets and make your way to "Doom Island." Disable their secret weapon and save the world!

Helping a Sea Friend
You meet a sea creature while at the beach. Turns out, they've lost their power after a natural disaster wipes out the "core" of their underwater city, putting all the inhabitants in danger. Help your new friend save the city before it's too late!

Amusement Park
Aliens have come to Earth and want to know all about your planet and math. Lead your new friends around an amusement park and teach them all about decimals!