The Impact of Gender in Learning with Games: A Consistent Effect in a Math Learning Game

Huy A. Nguyen
Carnegie Mellon University, Pittsburgh, PA, USA

Xinying Hou
University of Michigan, Ann Arbor, MI, USA

J. Elizabeth Richey
University of Pittsburgh, Pittsburgh, PA, USA

Bruce M. McLaren
Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

There is an established gender gap in middle school math education, where female students report higher anxiety and lower engagement, which negatively impacts their performance and even long-term career choices. This work investigates the role of digital learning games in addressing this issue by studying *Decimal Point*, a math game that teaches decimal numbers and operations to 5th and 6th graders. Through data from four published studies of *Decimal Point*, spanning a period of 5 years, and involving 624 students in total, the authors identified a consistent gender difference that was replicated across all studies: male students tended to do better at pretest, while female students tended to learn more from the game. In addition, female students were more careful in answering self-explanation questions, which significantly mediated the relationship between gender and learning gains in two out of four studies. These findings show that learning games can be an effective tool for bridging the gender gap in middle school math education, which in turn contribute to the development of more personalized and inclusive learning platforms.

INTRODUCTION

Many people are highly engaged with and frequently play video and computer games. World-wide, more than 2.6 billion people play video or computer games (Gilbert, 2021) and every day more and more people are playing. For instance, the NPD Group (NPD, 2019) reports that from 2018 to 2019 there was a 6% increase in people playing computer-based games. Young people are particularly engaged in digital game play. Based on Lobel et al. (2017), children from 7-12 years old play computer-based games approximately 5 hours per week, while Homer and colleagues (2012) reported much larger numbers of weekly hours of digital play by young people.

Due to their appeal, especially to young people, digital games have the potential to be powerful tools for learning. However, researchers and educators have questioned whether all students learn equally well from digital learning games, given that there are differences in their typical game preferences

(Dindar, 2018; Phan et al., 2012) and boys tend to spend more time playing (Homer et al., 2012). Yet, digital learning games have been shown to be effective for girls -- and often more effective than for boys -- in terms of both learning and affective outcomes (Arroyo et al., 2014; Hou et al., 2020, 2022; McLaren, Farzan et al., 2017b).

Although meta-analyses reveal gender similarities in math achievement (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; Wai et al., 2010). Critically, girls tend to report less positive math affect (Ganley & Lubienski, 2016; Hill et al., 2016), which in turn predicts their STEM engagement, goals, and achievement (Deemer et al., 2014; Else-Quest et al., 2013). Given how they often engage young people, digital learning games seem to be particularly well suited to address affective experiences with math, giving them potential to serve as a useful instructional tool for girls in particular.

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, 2012, 2013; Seaborn & Fels, 2015). In its attempts to reach more 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 male and female students 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.

We have developed and experimented with a digital learning game for middle school children, *Decimal Point*, that has proven to be an excellent platform for exploring gender differences in learning with the games. Unlike many digital learning games, *Decimal Point* was carefully designed to be gender-neutral and incorporate learning science principles based on empirical evidence. Over more than eight years of development and evaluation, *Decimal Point* has been used to explore various aspects of learning games, including an initial comparison with a non-game tutor, which showed that the game leads to superior learning outcomes compared to the tutor (McLaren et al., 2017a), the effects of student agency (Nguyen et al., 2018), the use of indirect control in the game (Harpstead et al., 2019) and the balance between learning and enjoyment (Hou et al., 2022). All versions of the game have a self-explanation step (Chi et al., 1989; 1994; Wylie & Chi, 2014) that prompts students after they play each of the mini-games within *Decimal Point*. While we have identified interesting aspects of all of the various studies of the game, one finding has remained steady since our earliest experiment and is the topic of this paper: girls have generally benefited more from the game than boys. In this paper we summarize and discuss the results of four separate experiments, spanning the years 2015 to 2019, all of which resulted in at least some learning benefits that favored girls over boys. Essentially, our almost decade-long research and work with the *Decimal Point* game has helped us answer the following questions:

RQ1: Is there a difference in learning outcomes between male and female students using *Decimal Point*?
RQ2: Is there a difference in game play behavior between male and female students?
RQ3: Is there a difference in self-explanation behavior between male and female students?

Our goal in examining these questions is to call to attention a consistent trend across studies that merits additional analysis in future research of *Decimal Point*, and in digital learning games more generally. Furthermore, while these questions were raised in the context of the game *Decimal Point* specifically, they have wider implications regarding learning from digital games more generally. In this paper, we discuss gender issues with respect to learning with games, describe our findings in experimenting with *Decimal Point*, and discuss the more general lessons from our results with respect to digital learning games.

BACKGROUND

Gender and Math Achievement

While boys and girls were shown to have similar performance in standardized tests (Hyde et al., 2008; Lindberg et al., 2010), girls often hold less positive attitudes towards math (Breda et al., 2018; Hill et al., 2016; C. Huang, 2013; Lindberg et al., 2010; Reilly et al., 2015; Wai et al., 2010), although the effect size is small and varies by age. In high school, several studies have reported that female students hold lower confidence, less excitement and greater frustration toward math than male students, with small to medium effect sizes (Arroyo et al., 2013; Else-Quest et al., 2010; 2013). However, this difference isn't present in elementary school (Andre et al., 1999; Friedler & Tamir, 1990), suggesting that middle school is when math anxiety emerges among female students and therefore 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 male students, math anxiety is a predictor of lower interest in math careers for female students (Huang et al., 2019).

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 (Spencer et al., 1999). While 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 can still emerge early in childhood and persist through adulthood (Cvencek et al., 2011; Furnham et al., 2002; Nosek et al., 2002; Passolunghi et al., 2014). In turn, such perception may influence female students' performance in math and their interest in STEM careers (Adams et al., 2019; Adams & Kirchmaier, 2016; Bian et al., 2017; Frome & Eccles, 1998; Ochsenfeld, 2016). For these reasons, promoting self-efficacy, interest and achievement among female students, while at the same time reducing math anxiety and stereotype threat, remains a challenging area of research. In this work, we investigate whether digital learning games, which aim to promote both learning motivation and outcomes (Sitzmann, 2011; Vogel et al., 2006), may contribute a solution pathway.

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 digital 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. Female students tend to rank goal clarity and social interaction as more important in digital learning games than male students, while male students tend to pay more attention to challenge, progress feedback and visual appeal in digital learning games (Dele-Ajayi et al., 2018). These preferences can produce meaningful differences, with medium to large effect sizes, in learning behaviors; for example, one study found that female students reported more positive feelings and increased help-seeking behaviors when a non-player "learning companion" was present, while male students 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., 2020; Kinzie & Joseph, 2008; Law, 2010; Pezzullo et al., 2017; Steiner et al., 2009). However, recommendations for gender-based adaptations typically rely on the intuitions of game designers or preferences observed through playtesting, focus groups, or surveys about self-reported preferences and behaviors. There remains a need to empirically validate these recommendations across multiple studies and populations to better understand their interaction with gender.

Among studies examining gender differences in learning from digital learning games, female students 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 male students (Joiner et al., 2011). 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) 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, Farzan et al., 2017b; Papastergiou, 2009). While Papastergiou (2009) found no gender effect on learning, McLaren, Farzan et al. (2017b) reported that female students benefited more from the game *Decimal Point*, the subject of this paper, than male students, with medium effect sizes. This difference was then replicated by Hou et al. (2020) in a separate study of the same game. Our research reported in this paper extends these prior results by performing a more comprehensive comparison between male and female students in all published studies of *Decimal Point*, including those that reported the game's gender effect (McLaren, Farzan et al., 2017b; Hou et al., 2020) and those that did not explore or report such effects (Nguyen et al., 2018; Harpstead et al., 2019).

The Game *Decimal Point*

Decimal Point (McLaren et al., 2017a), depicted in Figures 1 and 2, 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 was developed with HTML/JavaScript and the Cognitive Tutor Authoring Tools (CTAT - Aleven et al., 2016). The game and all related materials (e.g., tests, questionnaires) have been deployed on the web-based learning

management system, TutorShop (Aleven et al., 2009), which manages the game presentation to students and logs all of their actions.

The game is composed of a series of "mini-games" within the larger amusement park map (Figure 1). Each mini-game involves one of the five types of decimal problems, as shown in Table 1. After solving each problem, students answer a multiple-choice self-explanation question to reinforce their learning; this design is based on the self-explanation principle, which has been shown to lead to deeper and more robust learning in a variety of prior studies (Chi et al., 1989, 1994; Johnson & Mayer, 2010; Mayer & Johnson, 2010; Rittle-Johnson, 2006; Wylie & Chi, 2014).

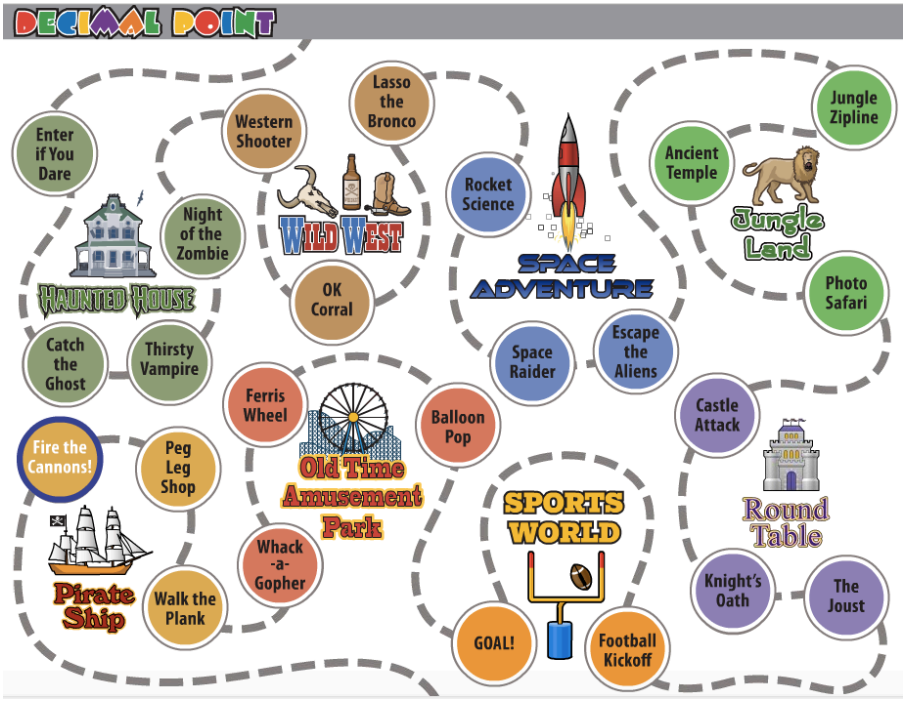
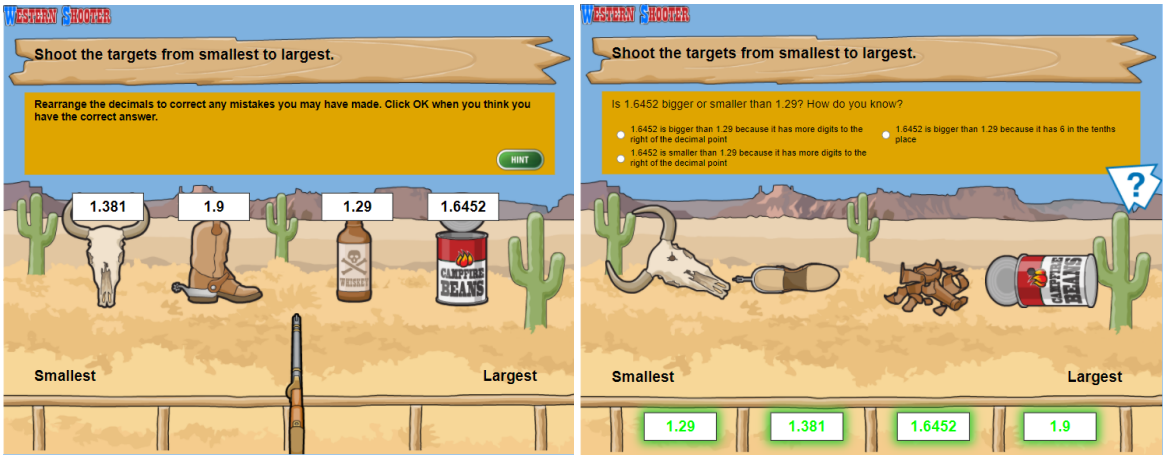


Figure 1. The main game map where students can select among 24 mini-games to play



(a) (b)

Figure 2. An example mini-game, Western Shooter, in the Sorting problem type and Wild West theme. Students first perform a sorting task (2a), then answer a multiple-choice self-explanation question about the performed task (2b).

Table 1. The list of game types and their game activities in Decimal Point.

Game type	Activity
Number Line	Locate the position of a decimal number on the number line
Addition	Add two decimal numbers by entering the carry digits and the sum
Sequence	Fill in the next two numbers in a sequence of 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 decimal numbers in ascending or descending order

As an example, in the *Sorting* mini-game, *Western Shooter* (Figure 2), students have to shoot the four objects in the correct order based on their associated number labels (i.e., smallest to largest or largest to smallest). Once all objects have been shot, students receive immediate feedback about the correctness of their sorting, and can rearrange the numbers if they are incorrectly ordered. After successfully finishing this activity, students have to answer a self-explanation question, which, in this example, is about how the number 1.6452 compares to 1.29 (Figure 2b). Students don't face any penalty for incorrect responses and can resubmit answers as many times as needed; however, they are not allowed to move forward without correctly solving all the problems in a mini-game.

The game *Decimal Point* is the result of rigorous research in learning science and game design. From the 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., 2010) and leverage the benefits of self-explanation in promoting deep, robust learning (Chi et al., 1994; Chi & Wylie, 2014). From the 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 with thirty-two 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. We also note that, in light of prior research on gender preferences in games and learning games, the amusement park was chosen to be equally appealing to both males and females. Subsequent development was carried out over a year, focusing on brainstorming the theme areas and mini-game settings that align with the overall theme and support student learning. Further details about the design process are reported in Forlizzi et al. (2014).

Decimal Point has been deployed in classroom studies over multiple years and has consistently led to significant learning in comparing before and after gameplay. In this paper, we focus on four

experiments involving a total of more than 600 student participants. The first study by McLaren et al. (2017a) showed that the game led to more learning than a non-game tutor with identical instructional content. Building on this result, subsequent studies have used the game as a platform to explore various research topics in game-based learning. In particular, Nguyen et al. (2018) 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. As a follow-up, Harpstead et al. (2019) then examined the impact of game interface elements on students' sense of agency and learning. Most recently, Hou et al. (2020) evaluated the effects of exposing students to the game's models of their learning and enjoyment. The game data collected from these studies have also been used in educational data mining research, to better understand the learning difficulties in decimal numbers (Nguyen et al., 2019), as well as the relationships between game play behaviors and learning outcomes (Hou et al., 2022; Nguyen et al., 2020; Richey et al., 2021; Wang et al., 2019)**.** However, none of these prior publications has focused on the role of gender in students' playing and learning experience. The following sections elaborate on how our analyses extend prior results of *Decimal Point* from a gender perspective.

METHODS

Experimental Procedure

Each study was conducted 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 a delayed posttest one week later.

The test items were identical across all four studies. Each test consists of 43 questions; most questions 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 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 four studies indicated no student performance difference among the three test forms at pretest, posttest, or delayed posttest (McLaren et al., 2017a; Nguyen et al., 2018; Harpstead et al., 2019; Hou et al., 2020).

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 by students immediately after game play, the students rated several statements about their enjoyment of the game elements, on a Likert scale from 1 ("strongly disagree") to 5 ("strongly agree").

Measures

To measure gender differences in learning, we partitioned the 43 test items 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 Barnett & Ceci (2002)'s taxonomy of transfer, 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 2, 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 8, under Appendix 1. Under this classification, we then measure 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 and self-explanation behavior, we consider four metrics: *game duration*, *game 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 (e.g., Figure 2a), while the latter are based on their answering of the multiple-choice self-explanation question at the end of each round (e.g., Figure 2b). 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.

While student enjoyment is also a metric of interest, the content of the evaluation survey was based on the game elements being evaluated in each study, and therefore differed across studies (for more details about the survey in each study, see the respective publications – McLaren et al., 2017a; Nguyen et al., 2018; Harpstead et al., 2019; Hou et al., 2020). As our goal in this paper is to identify consistent gender learning and gameplay patterns across studies, we will not consider these evaluation items in this paper and focus only on the learning and game play measures outlined above.

In the next section, we describe the setting of each study and the results of our analyses. To compare how male and female students differ on the above metrics, we 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.

STUDY SETTINGS

To identify consistent gender trends, we investigate our research questions in the four prior studies of *Decimal Point*. While these studies have manipulated the main game map in Figure 1, allowing students to progress through the mini-games in different ways, the learning content and gameplay mechanics of each mini-game (e.g., those in Figure 2a and 2b) were kept identical throughout. Using terminology from the area of intelligent tutoring systems (VanLehn, 2006), the four prior studies have manipulated the outer loop behavior of the game (responsible for managing and assigning all the learning tasks) while retaining the same inner loop behavior (responsible for guiding students through each step in a learning task). We briefly describe the study motivations and settings as follows.

The Spring and Fall 2015 studies were conducted by McLaren et al. (2017a) to rigorously evaluate the effectiveness of *Decimal Point* in a media comparison approach (Mayer, 2014), where the game was compared against a conventional computer tutor that had identical instructional content. In the game, students played through the 24 mini-games in the order shown in Figure 1, starting from the top left corner of the game map ("Enter If You Dare") and finishing at the bottom left corner ("Fire the Cannon"). Each mini-game consisted of two rounds, for a total of 48 rounds, with different question content each round but similar game play mechanics. The non-game tutor featured an identical problem set, with 48 rounds, but the problems were displayed in a standard tutor interface, without any fantasy settings or embellishment seen in the Game condition.

The Fall 2017 study (Nguyen et al., 2018) was motivated by whether agency – a key aspect in many computer games – is helpful to learning. While 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, in this study, Nguyen and colleagues (2018) sought to examine agency in a more meaningful context, both for learning and for playing, 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 used in the Fall and Spring 2015 studies, where students played through 48 rounds of mini-games in a fixed order. 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.

The Spring 2018 study (Harpstead et al., 2019) was conducted to further examine the effect of agency in *Decimal Point*. This study built 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 the game *Decimal Point,* the dashed line on the game map (Figure 1) 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, Harpstead and colleagues (2019) 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.

The Fall 2019 study (Hou et al., 2020) 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, students saw an open learner model that displayed their current mastery of each of the five decimal skills in Table 1; this data was computed based on their performance on the mini-game rounds completed so far. In the enjoyment-oriented version, 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 star to 5 stars) that they submitted after completing each mini-game, using an established survey format called the "fun-o-meter" (Read & MacFarlane, 2006). There was also a control condition identical to the High Agency version used in the Fall 2017 and Spring 2018 studies.

In this paper we focus on students' behaviors during the mini-games and on self-explanation prompts, which did not change across all of these experiments. Demographic information about the

participants in each study is reported in Table 2. Here the initial sample size denotes the original number of students enrolled in the study, while the final sample size indicates the number of students used for data analysis; we excluded those who did not complete all study materials or were outliers in their learning gains or delayed learning gains (more than 2.5 standard deviations away from the mean). In the Spring and Fall 2015 study, we also excluded all students in the Non-game condition, as our analysis focuses on the gender effects of the game.

Table 2. Participants and final sample in each study.

Study	**Initial sample size**	**Final sample size**	**Age M (*SD*)**
Spring and Fall 2015	213	70 (31 males, 39 females)	11.36 (0.48)
Fall 2017	197	158 (81 males, 77 females)	11.15 (0.60)
Spring 2018	287	237 (107 males, 130 females)	11.85 (0.47)
Fall 2019	196	159 (82 males, 77 females)	10.93 (0.64)
Totals	893	624 (301 males, 323 females)	11.39 (0.68)

RESULTS

Spring and Fall 2015 Studies

McLaren et al. (2017a) reported that the game led to significantly more learning and enjoyment than the conventional tutor, but did not consider any comparison between male and female students. A post-hoc analysis on this study by McLaren, Farzan et al. (2017b) showed two results related to the game's gender effect. First, both male and female students in the Game condition had higher posttest scores than their Non-game counterparts, but the effect size for female students was larger. Second, while male students in both conditions performed similarly on the delayed posttest, female students performed significantly better under the Game condition. These results were the first to indicate that the learning benefits from playing *Decimal Point* were greater for female students than for male students. Our analysis seeks to elucidate this effect by considering, among only students who played the game, whether female students learned more than male students. Additionally, McLaren, Farzan et al. (2017b) did not point to which aspect of the game may have led to the observed outcomes; in the research reported in this paper, we also examine potential gender differences in game play and self-explanation behavior, as a means of better understanding the observed gender effect on learning.

RQ1: *Is there a difference in learning outcomes between male and female students?*
Table 3 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between male and female students at each transfer level. We observed that at pretest, there were no significant differences in performance. After playing the game, female students trended toward

larger learning gains and delayed learning gains than male students at all three transfer levels; however, none of the comparisons yielded statistically significant differences.

Table 3. Comparison of test performance by gender at each transfer level.

Category	Transfer	Male M (*SD*)	Female M (*SD*)	Statistical result
Pretest score	Near	12.097 (5.896)	11.154 (4.760)	$F(1, 68) = 0.548$, $p = .461$, $\eta_p^2 = .008$
	Middle	3.484 (2.096)	3.744 (2.022)	$F(1, 68) = 0.276$, $p = .601$, $\eta_p^2 = .004$
	Far	11.542 (4.296)	10.795 (3.988)	$F(1, 68) = 0.437$, $p = .511$, $\eta_p^2 = .006$
Learning gains	Near	4.065 (4.553)	4.410 (4.381)	$F(1, 68) = 0.104$, $p = .748$, $\eta_p^2 = .002$
	Middle	0.484 (1.411)	0.846 (1.954)	$F(1, 68) = 0.753$, $p = .389$, $\eta_p^2 = .011$
	Far	1.452 (3.576)	1.949 (3.244)	$F(1, 68) = 0.370$, $p = .545$, $\eta_p^2 = .005$
Delayed learning gains	Near	4.452 (4.114)	5.179 (4.352)	$F(1, 68) = 0.507$, $p = .479$, $\eta_p^2 = .007$
	Middle	0.742 (1.879)	1.308 (1.922)	$F(1, 68) = 1.527$, $p = .221$, $\eta_p^2 = .022$
	Far	2.194 (3.331)	2.667 (3.279)	$F(1, 68) = 0.355$, $p = .554$, $\eta_p^2 = .005$

RQ2: *Is there a difference in game play behavior between male and female students?*
A one-way ANOVA showed a marginally significant difference in game duration per round in minutes, $F(1, 68) = 3.977$, $p = .050$, $\eta_p^2 = 0.055$, between male ($M = 1.355$, $SD = 0.513$) and female students ($M = 1.629$, $SD = 0.613$), with male students spending less time playing the game. There were no significant differences in the number of game errors per round, $F(1, 68) = 0.001$, $p = .978$, $\eta_p^2 < 0.001$, between male ($M = 2.922$, $SD = 2.669$) and female students ($M = 2.936$, $SD = 1.641$)

RQ3: *Is there a difference in self-explanation behavior between male and female students?*
A one-way ANOVA showed no significant differences in self-explanation duration per round in minutes, $F(1, 68) = 1.046$, $p = .310$, $\eta_p^2 = 0.015$, between male ($M = 0.319$, $SD = 0.092$) and female students ($M = 0.340$, $SD = 0.079$). However, there was a significant difference in the number of self-explanation errors per round, $F(1, 68) = 5.045$, $p = .028$, $\eta_p^2 = 0.069$, where male students ($M = 0.813$, $SD = 0.302$) made more errors than female students ($M = 0.639$, $SD = 0.337$).

Fall 2017 Study

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 by Nguyen et al. (2018) 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 that of students in the Low Agency condition. Here we compare how male and female students played and learned from the game in this study, which has not been previously reported.

RQ1: *Is there a difference in learning outcomes between male and female students?*
Table 4 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between male and female students at each transfer level. We observed that male students trended towards outperforming female students at all three transfer levels at pretest, especially at the near transfer level, where the difference was significant. However, after playing the game, female students achieved significantly higher learning gains and delayed learning gains at the far transfer level. At the same time, we found that male students trended toward higher learning gains and delayed learning gains than female students at the middle transfer level, where the difference in delayed learning gains was marginally significant.

Table 4. Comparison of test performance by gender at each transfer level, where shaded rows highlight key differences.

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

(†) $p < .1$; (*) $p < .05$

RQ2: *Is there a difference in game play behavior between male and female students?*
A one-way ANOVA showed a significant gender difference in game duration per round in minutes, $F(1, 156) = 11.727$, $p = .001$, $\eta_p^2 = 0.086$, where male students ($M = 0.786$, $SD = 0.475$) spent less time playing the game than female students ($M = 1.131$, $SD = 0.766$). There was also a significant gender difference in

number of game errors per round, $F(1, 156) = 7.16$, $p = .008$, $\eta_p^2 = 0.044$, where male students ($M = 1.784$, $SD = 1.382$) had fewer errors than female students ($M = 2.538$, $SD = 2.101$).

RQ3: *Is there a difference in self-explanation behavior between male and female students?*
A one-way ANOVA showed a marginally significant gender difference in self-explanation duration per round in minutes, $F(1, 156) = 3.072$, $p = .082$, $\eta_p^2 = 0.019$, between male ($M = 0.421$, $SD = 0.140$) and female students (M = 0.458, $SD = 0.120$), with female students trending toward longer self-explanation times. Additionally, there was a significant difference in number of self-explanation errors, $F(1, 156) = 5.735$, $p = .018$, $\eta_p^2 = 0.035$, where male students ($M = 0.661$, $SD = 0.458$) made significantly more errors than female students ($M = 0.505$, $SD = 0.354$).

Spring 2018 Study

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). Here we compare how male and female students played and learned from the game in this study, which has not been previously reported.

RQ1: *Is there a difference in learning outcomes between male and female students?*
Table 5 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between male and female students at each transfer level. We observed that male students trended towards outperforming female students at all three transfer levels at pretest, especially at the near transfer level, where the difference was significant. However, after playing the game, female students trended toward higher learning gains and delayed learning gains at all transfer levels, with female students performing significantly better than male students on the near- and middle-level items of the posttest.

Table 5. Comparison of test performance by gender at each transfer level, where shaded rows highlight key differences.

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 game play behavior between male and female students?*
A one-way ANOVA showed no significant differences in game duration per round in minutes, $F(1, 235) = 1.064$, $p = .303$, $\eta_p^2 = 0.007$, between male ($M = 0.507$, $SD = 0.365$) and female students ($M = 0.558$, $SD = 0.390$). Similarly, there were no significant differences in number of game errors per round, $F(1, 235) = 0.235$, $p = .628$, $\eta_p^2 = 0.001$, between male ($M = 1.391$, $SD = 1.357$) and female students ($M = 1.481$, $SD = 1.469$).

RQ3: *Is there a difference in self-explanation behavior between male and female students?*
A one-way ANOVA showed no significant gender differences in self-explanation duration per round in minutes, $F(1, 235) = 0.636$, $p = .426$, $\eta_p^2 = 0.003$, between male ($M = 0.383$, $SD = 0.113$) and female students ($M = 0.394$, $SD = 0.100$). However, there was a significant gender difference in self-explanation errors per round, $F(1, 235) = 11.391$, $p = .001$, $\eta_p^2 = 0.046$, where male students ($M = 0.692$, $SD = 0.518$) made significantly more errors than female students ($M = 0.495$, $SD = 0.381$).

Fall 2019 Study

Results from this 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., 2020). The authors also investigated gender differences in learning and reported that female students had higher learning gains than male students at the near and mid transfer levels, but did not examine gender differences in game play, which we analyzed and report here.

RQ1: *Is there a difference in learning outcomes between male and female students?*
Hou et al. (2020) have reported the results of this research question, which we include here for completeness. Table 6 shows the results of one-way ANOVAs comparing pretest scores, learning gains and delayed learning gains between male and female students at each transfer level. Male students performed marginally better than female students on the near transfer level of the pretest, but female students demonstrated significantly larger learning gains on the near- and middle-level items on the immediate test and near-level items on the delayed posttest.

Table 6. Comparison of test performance by gender at each transfer level, where shaded rows highlight key differences.

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 game play behavior between male and female students?*
A one-way ANOVA showed no significant gender difference in game duration per round in minutes, $F(1, 157) = 1.215$, $p = .272$, $\eta_p^2 = 0.019$, between male ($M = 1.009$, $SD = 0.834$) and female students ($M = 1.142$, $SD = 0.678$). There were no significant differences in average game errors per round, $F(1, 157) = 0.148$, $p = 0.701$, $\eta_p^2 = 0.001$, between male ($M = 2.367$, $SD = 2.580$) and female students ($M = 2.233$, $SD = 1.688$).

RQ3: *Is there a difference in self-explanation behavior between male and female students?*
A one-way ANOVA showed a significant gender difference in self-explanation duration per round in minutes, $F(1, 157) = 14.355$, $p < .001$, $\eta_p^2 = 0.084$, where male students ($M = 0.369$, $SD = 0.109$) spent less time on self-explanation questions than female students ($M = 0.449$, $SD = 0.153$). There was also a significant gender difference in self-explanation errors per round, $F(1, 157) = 8.204$, $p = .005$, $\eta_p^2 = 0.05$, with male students ($M = 0.868$, $SD = 0.397$) making more errors than female students ($M = 0.681$, $SD = 0.428$).

Result Summary and Post-hoc Analyses

In general, analyses across the four studies demonstrated consistent trends revealing gender differences in performance, time spent, and error rates. Table 7 summarizes all of the gender comparisons in the previous four studies. For RQ1 -- whether male and female students had different learning outcomes -- we observed that, across all four studies and three levels of transfer learning, male students tended to perform better than female students at pretest, but female students often had higher learning gains and delayed learning gains. This pattern is especially consistent at the near transfer level, where we also see the most

frequent occurrences of significant gender differences. For RQ2 -- whether male and female students had different game play behaviors -- our analyses showed that female students spent consistently more time than male students on game play across studies. Female students also mostly had higher game errors, but often not significantly so. For RQ3 -- whether male and female students had different self-explanation behaviors -- we saw that male students had either lower or similar self-explanation durations, compared to female students. However, there is a notable difference in the average number of self-explanation errors: male students made significantly more self-explanation errors than female students in every study of *Decimal Point*.

Table 7. Summary of learning, game play and self-explanation comparisons by gender across studies, where shaded rows highlight key differences. The value in each cell indicates which gender had higher outcomes in the corresponding category (M for male and F for female).

Category	**SF15 (*n* = 70)**	**F17 (*n* = 158)**	**S18 (*n* = 237)**	**F19 (*n* = 159)**
Learning				
Pretest - Near transfer	M	M (*)	M (*)	M (†)
Pretest - Middle transfer	F	M	M	M (*)
Pretest - Far transfer	M	M (†)	M	F
Learning gains - Near transfer	F	F	F (*)	F (*)
Learning gains - Middle transfer	F	M	F (*)	F (*)
Learning gains - Far transfer	F	F (*)	F	M
Delayed learning gains - Near transfer	F	F	F (*)	F (*)
Delayed learning gains - Middle transfer	F	M (†)	F	F
Delayed learning gains - Far transfer	F	F (*)	F	F
Game play				
Game duration	F (†)	F (*)	F	F
Game errors	F	F (*)	F	M
Self-explanation				
Self-explanation duration	F	F (†)	M	F (*)
Self-explanation error	M (*)	M (*)	M (*)	M (*)

(†) $p < .1$; (*) $p < .05$.

The fact that male students spent lower or similar amounts of time on the self-explanation activities but made significantly more errors than female students could indicate that they may have been more careless

in answering self-explanation questions, or "gamed" the questions (i.e., quickly selected the options until they got the correct answer), which in turn could have contributed to their lower learning gains. To test this hypothesis, we computed a new metric called *self-explanation error rate*, which is the total number of self-explanation errors divided by the total time spent on self-explanation activities, across the student's entire playthrough. A higher metric value indicates that the student made errors at a faster rate, which could be considered an indication of greater carelessness or gaming. We then constructed two mediation models with gender as an independent variable (where male is coded as 0 and female as 1), self-explanation error rate as a mediator, and near transfer learning gain / delayed learning gain as the dependent variable. Here we only consider learning gains at the near transfer level because this level led to the most consistent gender differences, as previously described. 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). For significant mediation effects, we reported the effect size via the absolute ratio of the indirect to the total effect, i.e., the mediation ratio, which indicates the proportion of total effect which is mediated (Preacher & Kelley, 2011). We report the result of this analysis on each of the four *Decimal Point* studies below.

Spring and Fall 2015 Studies

Our mediation models indicated no significant mediation effect of self-explanation error rate in the relationship between gender and near transfer learning gain (ab = 0.222, 95% CI [-0,107, 1.148], p = .348) or near transfer delayed learning gain (ab = 0.137, 95% CI [-0.153, 0.969], p = .525). In each model, the total effect, without accounting for the mediator, was likewise not significant: c = -0.346, p = .748 for the near transfer learning gain model, and c = -0.728, p = .479 for the near transfer delayed learning gain model.

Fall 2017 Studies

Our mediation models indicated that the effect of gender on near transfer learning gain was mediated by error rate (Figure 3). The regression coefficient between gender and error rate was significant, as was the regression coefficient between error rate and near transfer learning gain. The bootstrap procedures also indicated a significant indirect effect (ab = 0.473, 95% CI [0.151, 0.942], p = .013), with a mediation ratio of |0.473 / -0.810| = 58.4%. Similarly, the effect of gender on near transfer delayed learning gain was also mediated by self-explanation error rate, with a significant regression coefficient between self-explanation error rate and near transfer delayed learning gain. Results of the bootstrapping procedures showed a significant indirect effect (ab = 0.437, 95% CI [0.153, 0.890], p = .013), with a mediation ratio of |0.437 / -0.958| = 45.62%.

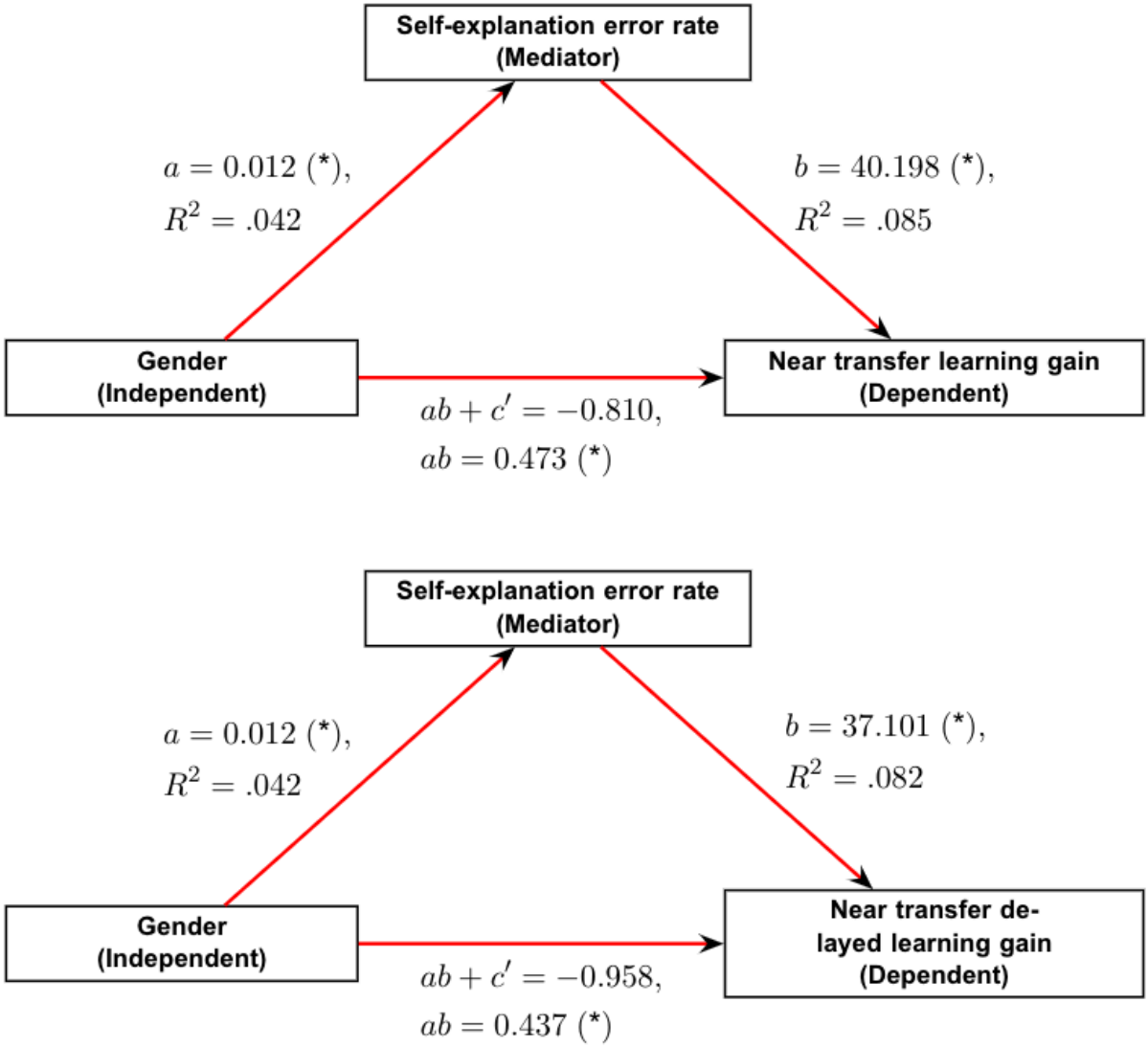


Figure 3. Diagram of the mediation model for near transfer learning gain (top) and delayed learning gain (bottom) in the Fall 2017 study. (*) indicates significance at the 0.05 level.

Spring 2018 Study

Our mediation models indicated that the effect of gender on near transfer learning gain was mediated by error rate (Figure 4). The regression coefficient between gender and error rate was significant, as was the regression coefficient between error rate and near transfer learning gain. The bootstrap procedures also indicated a significant indirect effect (*ab* = 0.384, 95% CI [0.133, 0.732], $p < .001$), with a mediation ratio of |0.384 / -1.251| = 30.7%. Similarly, the effect of gender on near transfer delayed learning gain was also mediated by self-explanation error rate, with a significant regression coefficient between self-explanation error rate and near transfer delayed learning gain. Results of the bootstrapping procedures showed a significant indirect effect (*ab* = 0.432, 95% CI [0.145, 0.814], $p < .001$), with a mediation ratio of |0.432 / -1.017| = 42.5%.

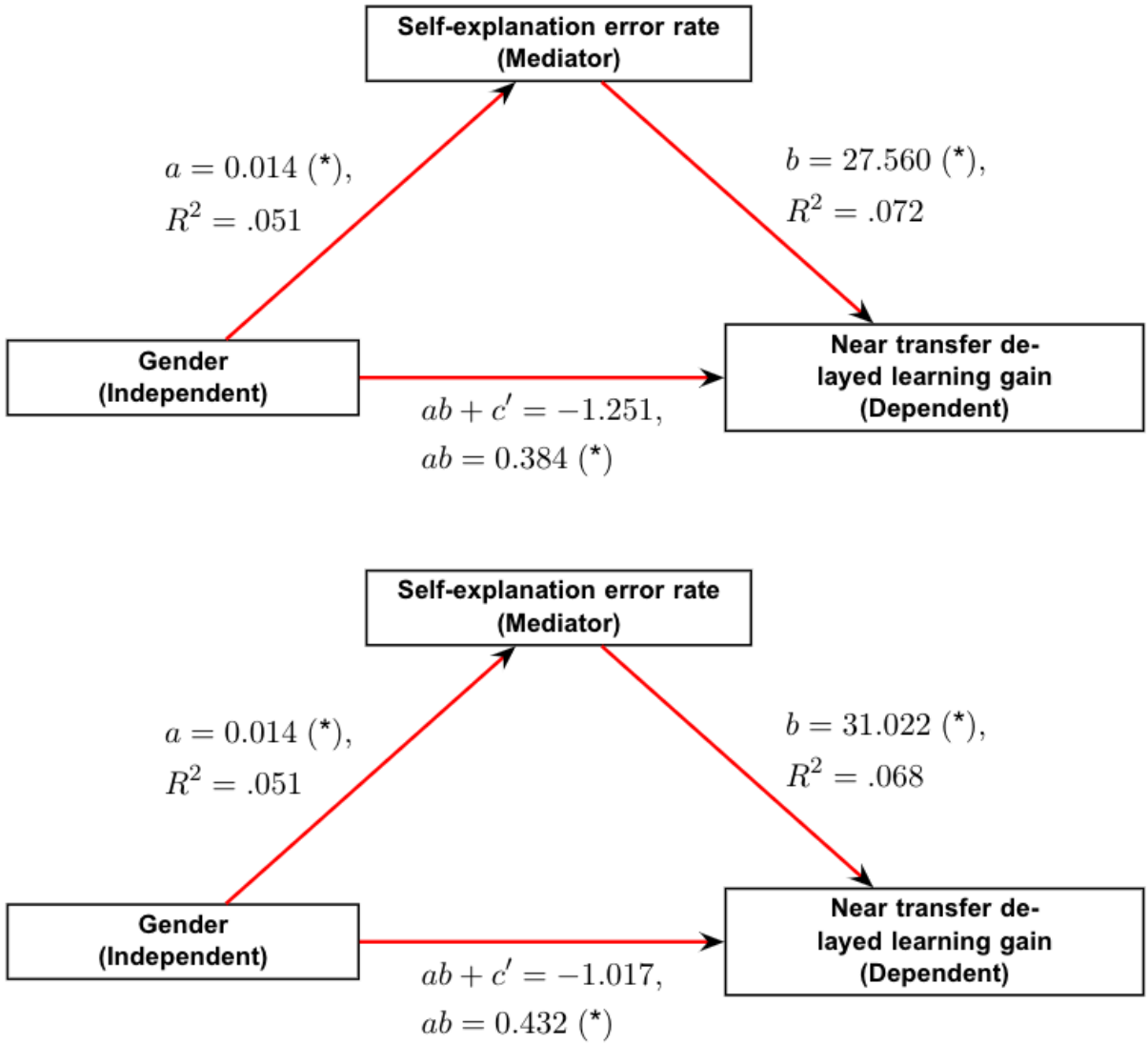


Figure 4. Diagram of the mediation model for near transfer learning gain (top) and delayed learning gain (bottom) in the Spring 2018 study. (*) indicates significance at the 0.05 level.

Fall 2019 Study

Our mediation models indicated no significant mediation effect of self-explanation error rate in the relationship between gender and near transfer learning gain (*ab* = -0.068, 95% CI [-0,495, 0.282], *p* = .725) or near transfer delayed learning gain (*ab* = -0.205, 95% CI [-0.640, 0.136], *p* = .289). However, in each model, the total effect, without accounting for the mediator, was significant: *c* = 1.166, *p* = .035 for the near transfer learning gain model, and *c* = 1.030, *p* = .047 for the near transfer delayed learning gain model.

DISCUSSION

In summary, our classroom studies with the digital learning game *Decimal Point* over a period of four years have identified consistent gender differences in students' learning outcomes and self-explanation

behaviors. First, female students under-performed compared to male students on the pretests but out-performed male students in learning gains and delayed learning gains. This result did not reach significance every year, but consistently emerged as a strong trend, especially at the near transfer level, which is closest to the game's learning content. Second, female students made fewer errors than male students on self-explanation questions, though not during the problem-solving portion of gameplay. This difference was significant in all four studies. In addition, the self-explanation error rate (total number of self-explanation errors divided by total self-explanation duration) was a significant mediator of the relationship between gender and learning gains at the near transfer level in the Fall 2017 and Spring 2018 study. These findings are striking, given that the game's amusement park theme and learning activities were designed to be gender-neutral, rather than to align with a specific gender's preferences (Anonymous authors, 2014). While the effect sizes of our gender comparisons are small, the consistent trend is noteworthy and could point to an important game design feature that may be leveraged in future work to further support female students' learning and bridge the gender gap in math education. We elaborate on these implications below.

Our observation of male students having higher pretest scores is consistent with prior literature demonstrating male students' tendency to do better at math than female students in late elementary and early middle school (Robinson & Lubienski, 2011). However, the fact that female students had consistently higher learning gains and delayed learning gains is an important pattern. Notably, this pattern was not due to the ceiling effect, as both male and female students' 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 female students catch up with their male counterparts in math performance after playing. In turn, our work contributes to the body of research showing that digital learning games can lead to gender differences in learning outcomes that favor female students (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), we also set out to explore why *Decimal Point* was more beneficial for female students.

Our first conjecture was that female students learned more because they approached the self-explanation questions more carefully and deliberately. In contrast, male students had significantly higher error rates, which might be due to their carelessness or gaming of the questions (i.e., they may have selected all of the multiple-choice options rapidly until arriving at the correct answer). Given that 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 partially supported by our post-hoc analysis, which reveals a significant mediation effect of the self-explanation error rate in two out of four studies of *Decimal Point*. Although some have speculated that young girls' more rapid development of verbal learning strategies might give them an advantage over boys when learning from self-explanation (Nikolaenko, 2005; Stevenson et al., 2009), much of the prior literature examining self-explanation interventions did not report on gender differences (Bisra et al., 2018; Durkin, 2011; Rittle-Johnson, 2006). One prior study testing this idea with 7- to 9-year-olds found significant gender differences in learning through self-explanation, where female students performed better than male students 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, our

results 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 (Johnson & Mayer, 2010; Mayer & Johnson, 2010).

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 female students (Doyle & Voyer, 2016; Nguyen & Ryan, 2008; Picho et al., 2013; Spencer et al., 1999). By reducing female students' stereotype threat-triggered anxiety, games may free up more working memory space for learning about mathematical concepts and, as a result, allow female students to catch up to male students on the posttest despite typically receiving lower scores on the pretest (Gödöllei Lappalainen, 2017; Sitzmann, 2011). If the game affords female students 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 female students were more thoughtful on self-explanation questions, spending more time on them and making fewer errors. Future research may test this hypothesis by measuring students' anxiety as a means of assessing the impact of stereotype threat. If stereotype threat was reduced for female students in *Decimal Point* compared to a non-game version, we would expect female students to report higher anxiety than male students in the non-game version but similar or lower levels of anxiety in the game. Measures of anxiety would also allow us to examine whether male students felt anxious about the self-explanation questions, given that they tended to have lower language skills than female students at the middle school level (Park et al., 2007; Stevenson et al., 2009). This anxiety, if present, would help explain the higher error rates in self-explanation questions observed among male students in our studies.

An opposite trend was observed in the problem-solving activities in the game (e.g., Figure 2a), where female students tended to spend more time and make more errors than male students, although not significantly so. This difference can be attributed to female students' 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 shown 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). From our studies, we indeed observed that female students were able to acquire higher learning gains after game play. When examining the role of the problem-solving activities in inducing this effect, we 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 2a, students would get to interact with different objects representative of the Wild West 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 female students – could tackle them more comfortably than they did in the pretest.

Taken together, our findings suggested several mechanisms through which learning games can bridge the gender gap in middle-school math education. First, females demonstrated better learning with self-explanation than males, which could potentially lead to their higher learning gains. Second, the informal game context could reduce the stereotype threat that female students face while studying math. Third, the immersive game themes and narratives could offset the negative emotions that students may experience during the learning process. Most notably, while these mechanisms appear to have stronger effects on females than males, they have the potential to benefit both genders alike. In other words, promoting female students' math learning does not need to be at the expense of male students. Rather

than catering the game to a specific population's preferences, learning game researchers could employ inclusive mechanisms that both support every student and narrow the existing gap in learning outcomes. Through this work, we propose three such mechanisms – self-explanation, informal context and immersion – which merit additional validation and extension at a larger scale in other learning domains.

LIMITATIONS AND FUTURE WORK

One issue with our research -- and, in fact, with much of research that investigates the impact of educational technology on gender -- is that the conventional binary classification of gender (male versus female) does not account for the spectrum of variance in gendered behavior (Hyde et al., 2019). Some research in gender studies has moved towards a multi-dimensional gender framework that also incorporates gender identity, typicality, occupational interests, activities and traits (Egan & Perry, 2001; Liben & Bigler, 2002; Martin et al., 2017). Collecting these attributes via pre-intervention surveys would allow us to build a more holistic and individualized profile for each student, thereby allowing deeper studies of which gender dimensions and game features best predict learning outcome, how they interact, and how they are mediated by different cognitive processes. The above attributes are also critical in implementing real-time adaptivity within the game, which is a prominent area at the intersection of AI and education, and has been explored in a previous study of *Decimal Point* (Hou et al., 2022). We could then derive principles for how to design digital learning games for all students, with the intention of ultimately generalizing our findings across different learning games.

In addition, future studies of *Decimal Point* would benefit from having a unified method of collecting engagement and affect data, in order to more deeply explore and compare each gender's affective and cognitive processes, self-reported feelings, step-by-step actions, and learning outcomes. To this end, we could build machine-learned detectors for engaged concentration, delight, boredom, and behavioral measures of disengagement (e.g., gaming the system, careless errors, behavior not aimed at completing the learning task; Baker & Ocumpaugh, 2015; Baker et al., 2010; D'Mello, 2013; Shute et al., 2015). We will also assess, at a more fine-grained level, self-reported engagement (Ben-Eliyahu et al., 2018) and situational interest (Linnenbrink-Garcia et al., 2010). If female students found the game more engaging than male students, we would expect that female students would experience greater engaged concentration, delight, and interest, in addition to less boredom and disengagement, compared to male students. We also expect these measures to partially mediate the relation between gender and learning gains. More generally, identifying whether each gender experiences a different affective state, and how it influences their learning experience, is an important step towards building an effective learning game for bridging the gender gap in mathematics motivation and performance.

Finally, while our work has identified the gender differences in learning and self-explanation across four *Decimal Point* studies, we found that not all patterns were consistent across studies. The Fall and Spring 2015 studies, in particular, suffered from a small sample size (n = 70) and did not yield any significant results. Having more replication studies in the future with larger sample sizes would help reinforce the game's gender effect and provide more insights into which game elements are conducive to this effect. We envision that *Decimal Point* will help foster positive math affect among girls, which can offset the gender stereotypes that impact students' learning trajectories (Cvencek et al., 2011; Nosek et al., 2002) and, in the long term, broaden STEM participation (Bian et al., 2017; Doyle & Voyer, 2016; Passolunghi et al., 2014).

CONCLUSION

In this work, we investigated the differences between male and female students in playing and learning from the digital learning game *Decimal Point,* which was developed using a rigorous design process based on learning science principles. Through analysis of data from four previous classroom studies with over 600 students, we identified a trend of female students having lower pretest scores but higher learning gains after game play, especially at the near transfer learning level. This is a highly consistent and important finding which can be attributed to several factors, including the students' self-explanation performance, the game's fantasy setting, and its effect in reducing math anxiety. In turn, our results underline the potential of digital learning games in bridging the gender gap in math education, while raising crucial questions about which game elements are most conducive to the gender effect, and which dimensions of gender have the most impact in this context. Addressing these questions in future studies is an important step towards promoting inclusive and effective games for education.

ACKNOWLEDGMENT

Thanks to Jodi Forlizzi, Jon Star, Michael Mogessie Ashenafi, Scott Herbst, Craig Ganoe, Darlan Santana Farias, Rick Henkel, Patrick Bruce Gonçalves McLaren, Grace Kihumba, Kim Lister, Kevin Dhou, John Choi, and Jimit Bhalani, all of whom made important contributions to the design, development, and early experimentation with the *Decimal Point* game. Special thanks to Jon Star who provided input on the hint and error messages used in this study. Special thanks also to Rosta Farzan who first had the idea to do gender analyses of *Decimal Point* use.

Conflict of Interest

The authors of this publication declare there is no conflict of interest.

Funding Agency

This work was supported by the National Science Foundation Award #DRL-1661121. The opinions expressed are those of the authors and do not represent the views of NSF.

REFERENCES

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

Adams, R. B., Barber, B. M., & Odean, T. (2019). The Math Gender Gap and Women's Career Outcomes. *Available at SSRN 2933241*.

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., McLaren, B. M., & Sewall, J. (2009). 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.

Andre, T., Whigham, M., Hendrickson, A., & Chambers, S. (1999). Competency beliefs, positive affect, and gender stereotypes of elementary students and their parents about science versus other school subjects. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *36*(6), 719–747.

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.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, *24*(4), 387–426.

Baker, Rsj., & Ocumpaugh, J. (2015). *Interaction-based affect detection in educational software*. New York: Oxford University Press.

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.

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.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). *Inducing self-explanation: A meta-analysis*. Springer.

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.

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

Bull, S. (2020). There are open learner models about!. *IEEE Transactions on Learning Technologies*, *13*(2), 425-448. https://doi.org/10.1109/TLT.2020.2978473.

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.

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, 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.

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.

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

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.

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.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, *29*(3), 241–250.

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

d Baker, R. S., Gowda, S. M., & Corbett, A. T. (2011). Towards predicting future transfer of learning. *International Conference on Artificial Intelligence in Education*, 23–30.

d Baker, R. S., Mitrović, A., & Mathews, M. (2010). Detecting gaming the system in constraint-based tutors. *International Conference on User Modeling, Adaptation, and Personalization*, 267–278.

Deemer, E. D., Thoman, D. B., Chase, J. P., & Smith, J. L. (2014). Feeling the threat: Stereotype threat as a contextual barrier to women's science career choice intentions. *Journal of Career Development*, *41*(2), 141–158.

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. https://doi.org/10.1145/2858036.2858395.

Dindar, M. (2018). An empirical study on gender, video game play, academic success and complex problem solving skills. *Computers & Education*, *125*, 39–52.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, *105*(4), 1082.

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.

Durkin, K. (2011). The Self-Explanation Effect when Learning Mathematics: A Meta-Analysis. *Society for Research on Educational Effectiveness*.

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

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.

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.

Friedler, Y., & Tamir, P. (1990). Sex differences in science education in Israel: An analysis of 15 years of

research. *Research in Science & Technological Education*, *8*(1), 21-34.

Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology*, *74*(2), 435.

Furnham, A., Reeves, E., & Budhani, S. (2002). Parents think their sons are brighter than their daughters: Sex differences in parental self-estimations and estimations of their children's multiple intelligences. *The Journal of Genetic Psychology*, *163*(1), 24–39.

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.

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.

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.

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. (2019). 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.

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.

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.

Hou, X., Nguyen, H. A., Richey, J. E., Harpstead, E., Hammer, J., & McLaren, B. M. (2022). 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., & McLaren, B. M. (2020). Exploring how gender and enjoyment impact learning in a digital learning game. In *International Conference on Artificial Intelligence in Education* (pp. 255-268). Springer, Cham.

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.

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

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

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

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.

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.

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.

Kostyuk, V., Almeda, M. V., & Baker, R. S. (2018). Correlating affect and behavior in reasoning mind with state test achievement. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 26–30.

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

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.

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

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

Nguyen, H., Hou, X., Stamper, J., & McLaren, B. M. (2020). Moving beyond Test Scores: Analyzing the Effectiveness of a Digital Learning Game through Learning Analytics. *Proceedings of the 13th International Conference on Educational Data Mining*, 487–495.

Nguyen, H., Wang, Y., Stamper, J., & McLaren, B. M. (2019). Using Knowledge Component Modeling to Increase Domain Understanding in a Digital Learning Game. *Proceedings of the 12th International Conference on Educational Data Mining*, 139–148.

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

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

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.

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math= male, me= female, therefore math≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44.

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.

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.

Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. *Journal of Learning Analytics*, *1*(1), 107–128.

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.

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.

Phan, M. H., Jardina, J. R., Hoyle, S., & Chaparro, B. S. (2012). Examining the role of gender in video game usage, preference, and behavior. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1496–1500.

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.

Preacher, K. J., & Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, *16*(2), 93.

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.

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

Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R. S., & McLaren, B. M. (2021). Gaming and Confrustion Explain Learning Advantages for a Math

Digital Learning Game. *International Conference on Artificial Intelligence in Education*, 342–355.

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.

Rodrigo, M. M. T., Baker, R. S., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., Lim, S. A. L., Pascua, S. A., Sugay, J. O., & Tabanao, E. S. (2009). Affective and behavioral predictors of novice programmer achievement. *Proceedings of the 14th Annual ACM SIGCSE Conference on Innovation and Technology in Computer Science Education*, 156–160.

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

San Pedro, M. O. Z., d Baker, R. S., & Rodrigo, M. M. T. (2014). Carelessness and affect in an intelligent tutoring system for mathematics. *International Journal of Artificial Intelligence in Education*, *24*(2), 189–210.

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

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

Shute, V. J., D'Mello, S., Baker, R., Cho, K., Bosch, N., Ocumpaugh, J., Ventura, M., & Almeda, V. (2015). Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game. *Computers & Education*, *86*, 224–235.

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.

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.

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.

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.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research*, *34*(3), 229–243.

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.

Wang, Y., Nguyen, H., Harpstead, E., Stamper, J., & McLaren, B. M. (2019). How Does Order of Gameplay Impact Learning and Enjoyment in a Digital Learning Game? *International Conference on Artificial Intelligence in Education*, 518–531.

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

APPENDIX 1

Table 8**.** 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