Understanding Gender Effects in Game-Based Learning: The Role of Self-Explanation

J. Elizabeth Richey[1], Huy A. Nguyen[2], Mahboobeh Mehrvarz[2], Nicole Else-Quest[3], Ivon Arroyo[4], Ryan S. Baker[5], Hayden Stec[2], Jessica Hammer[2], and Bruce M. McLaren[2]

[1] University of Pittsburgh, Pittsburgh PA 15260, USA
[2] Carnegie Mellon University, Pittsburgh PA 15213, USA
[3] University of North Carolina at Chapel Hill, Chapel Hill NC 27599, USA
[4] University of Massachusetts Amherst, Amherst MA 01003, USA
[5] University of Pennsylvania, Philadelphia PA 19104 USA

Abstract. We conducted a 2x2 study comparing the digital learning game *Decimal Point* to a comparable non-game tutor with or without self-explanation prompting. We expected to replicate previous studies showing the game improved learning compared to the non-game tutor, and that self-explanation prompting would enhance learning across platforms. Additionally, prior research with *Decimal Point* suggested that self-explanation was driving gender differences in which girls learned more than boys. To better understand these effects, we manipulated the presence of self-explanation prompts and incorporated a multidimensional gender measure. We hypothesized that girls and students with stronger feminine-typed characteristics would learn more than boys and students with stronger masculine-typed characteristics in the game with self-explanation condition, but not in the game without self-explanation or in the non-game conditions. Results showed no advantage for the game over the non-game or for including self-explanation, but an analysis of hint usage indicated that students in the game conditions used (and abused) hints more than in the non-game conditions, which in turn was associated with worse learning outcomes. When we controlled for hint use, students in the game conditions learned more than students in the non-game tutor. We replicated a gender effect favoring boys and students with masculine-typed characteristics on the pretest, but there were no gender differences on the posttests. Finally, results indicated that the multidimensional framework explained variance in pretest performance better than a binary gender measure, adding further evidence that this framework may be a more effective, inclusive approach to understanding gender effects in game-based learning.

Keywords: Digital Learning Game, Gender Studies, Self-Explanation, Hints

1 Introduction

Digital learning games can promote learning through playful, engaging, and highly interactive interfaces, but depending on the features of games (e.g., narrative, design style, pacing), they may not be equally effective for all learners [1, 2]. *Decimal Point,* a digital math game designed to teach students about decimal numbers and operations, has produced consistent gender effects favoring girls [3]. In prior research seeking to understand the source of these consistent gender effects, we have found evidence that differences in girls' and boys' response patterns to self-explanation prompts seem to be at least partially responsible for the gender differences in learning outcomes [3]. Specifically, girls have demonstrated fewer errors and less gaming the system [4] on self-explanation questions in the game, and those behaviors in turn have mediated the gender effect on learning outcomes [5].

In this paper, we seek to advance our understanding of the gender effect in *Decimal Point* in two ways: first, we conducted an experiment to directly test whether self-explanation was responsible for the gender effect by varying whether students were prompted to self-explain as they played the game or solved equivalent problems in a non-game platform. Second, we sought to expand our understanding of gender differences in gameplay and learning by incorporating a multidimensional gender framework. This framework holds that gender includes multiple separate but interrelated dimensions, including aspects of identity as well as activities and interests [6]. It has the potential to reveal more nuanced gender-related characteristics that might more directly explain differences in how students play and learn from digital games. Additionally, a multidimensional gender representation promises to be a more inclusive approach to understanding gender, as it will allow us to include students of all genders in our analyses–instead of limiting analyses to students who fit within binary gender identity categories–while capturing more complex aspects of gender along a continuous spectrum. In this paper, we focus on the gender dimension of students' self-reports of gender-typed occupational interests, activities, and traits [7].

1.1 Enhancing learning through digital learning games

There is ample evidence that digital learning games can engage students and support learning [8-11]. Educational technology researchers have embraced game-based learning by building games for a variety of domains [12-14]. Games are thought to support learning through engagement [15] and flow [16], in which learners focus their full attention on game play, potentially taking focus away from negative thoughts or emotions about the instructional content. For struggling learners, game-based learning can be a particularly effective way to engage with the material [17].

Decimal Point is a digital math learning game that has produced better learning outcomes than comparable non-game instruction. In a study with 153 middle school students who either played *Decimal Point* or learned with a content-equivalent online tutor [12], *Decimal Point* students learned more than the tutored students, with relatively high effect sizes (immediate posttest: d = 0.65; delayed posttest: d = 0.59). The *Decimal Point* students also reported enjoying their experience significantly more than the tutored students, according to a post-game questionnaire. An analysis of learning outcomes by binary gender identity revealed that girls learned significantly more from the game than boys [18]. We subsequently performed experiments with different versions of the game and consistently uncovered the same finding: girls learned significantly more from the game than boys regardless of alterations [3, 19].

Some prior work has explored the role of self-explanation in game-based learning [20], and we were also interested in the impact self-explanation could have on learning with the game. As a result, the game incorporates a series of multiple-choice self-explanation questions designed to address misconceptions and promote conceptual understanding of the decimal number concepts in the game [21]. When investigating sources of the gender difference in learning outcomes, we discovered that boys' and girls' game-play behaviors and error rates especially differed on the self-explanation steps [3]. Specifically, girls tended to have lower error rates and exhibited fewer behaviors suggesting they were trying to take advantage of the affordances of the learning system to get the right answer without thinking, a behavior referred to as "gaming the system" [4]. Further analyses indicated that rates of gaming the system partially mediated the effect of gender on learning, suggesting that students' interactions with the self-explanation prompts were at least in part driving the gender effect in *Decimal Point* [5].

1.2 Gender, math, and digital learning games

Although girls and boys tend to perform equally well in math, gender differences often emerge in motivation, emotions, and perceptions around math. For example, girls report greater anxiety towards math and less self-confidence in their math abilities [22, 23].

Digital math games could provide a valuable tool for promoting more equitable engagement in math. Games may be especially effective for girls to the degree that they promote enjoyment and reduce salient cues likely to trigger stereotype threat, which occurs when being reminded of social group stereotypes impairs the performance of members of that group [24]. In the context of math, even implicit cues like labeling the nature of a task as mathematical can trigger stereotype threat for women and reduce their performance [25]. Embedding math practice within the context of a digital learning game might therefore reduce stereotype threat by de-emphasizing the mathematical nature of the task.

Digital learning games in math appear to be effective for all genders [26], despite some evidence of broader gender-based differences in game preferences [27]. In fact, other math digital learning games have uncovered gender differences in learning outcomes based on different game features [28], lending evidence that digital learning games may affect some aspects of learning for girls and boys differently.

One key to better understanding how and why digital learning games might produce gender differences involves taking a more comprehensive view of gender. All prior research investigating gender differences in digital game learning has adopted a binary view of gender, sorting learners into binary categories of boys and girls and typically excluding anyone outside the gender binary due to small numbers. However, it is likely that any effect of gender on learning reflects gender-related differences in behaviors, interests, and experiences rather than binary gender identity itself. As a result, a measure of gender that captures these nuanced, multifaceted factors may be a more powerful predictor of learning differences. It could also better illuminate which aspects of gender (e.g., interests, activities) are most predictive of differences in learning behaviors and outcomes.

We address this gap in prior research and aim to better understand gender differences in *Decimal Point* by incorporating a multidimensional gender framework and measuring gender-typed occupational interests, activities, and traits, which we will refer to as "gender-typed characteristics" [7]. We also experimentally test the evidence that self-explanation prompts are producing the gender effect observed in *Decimal Point.* In this research, we explored the following questions:

RQ1: Will the learning platform (game vs. non-game control) and the presence or absence of self-explanation questions affect learning outcomes?

RQ2: Will the learning platform (game vs. non-game control) and the presence or absence of self-explanation questions interact with gender to explain *gender-based* differences in learning outcomes?

RQ3: Will the multidimensional gender framework predict variance in learning outcomes better than a measure of binary gender identity?

For the first research question, we hypothesized that we would replicate previous studies showing a learning advantage for students in the game condition compared to the non-game tutor [12], regardless of self-explanation condition. We also hypothesized that self-explanation questions would lead to greater learning across both learning platforms, which we have not previously tested. We did not predict an interaction between self-explanation and learning platform, as we expected each to contribute an additive effect to learning.

For the second research question, we hypothesized that removing the self-explanation prompts would eliminate the gender effect in the game based on prior analyses suggesting that the self-explanation prompts were driving gender differences. Specifically, we hypothesized that girls and students with stronger feminine-typed traits would learn more than boys and students with stronger masculine-typed traits *only in the game + self-explanation condition*. We did not expect to see gender differences in the game condition without self-explanation, as this condition would be missing the learning component hypothesized to be creating gender differences. In other words, if the self-explanation step in the game was creating the gender effect, then we would no longer expect to see a gender difference in the game when self-explanation was removed. We did not expect to see a gender difference in the non-game platform, regardless of the presence or absence of self-explanation prompts, as prior research revealed no gender differences in the non-game [18].

For the third research question, if students differed in test performance by gender, we expected that the continuous measures of gender (i.e., gender-typed occupational interests, activities, and traits) would explain more variance in test performance than a binary measure of gender identity (i.e., boy or girl). This is because gender-typed characteristics reflect more nuanced aspects of gender, which are likely more closely related than binary gender identity to any motivation or emotion that might in turn predict students' learning in a math game. This prediction is also consistent with preliminary work showing that multidimensional measures of gender explain differences in game preferences better than binary gender [27].

2 Method

2.1 Participants

We conducted our research in eight elementary and middle schools in a mid-sized U.S. city and the surrounding suburban and rural areas. A total of 576 students from fifth- and sixth-grade classes participated in the study, but a technical issue resulted in data not being recorded for 90 students from one school. An additional 100 students were excluded from analyses because they did not complete all study and test materials. The remaining 386 students were assigned to one of four conditions: a game with self-explanation, a game without self-explanation, a non-game with self-explanation, or a non-game without self-explanation. Given the potential distraction of having some students play a game while others worked with a non-game tutor within a classroom, we randomly assigned students to game conditions at the classroom level and to the self-explanation conditions at the individual level. Students ranged in age from 10 to 13 years old (M = 10.85, SD = 0.65); 201 identified as male, 182 identified as female, one identified as trans or non-binary, and two preferred not to disclose their gender identity.

2.2 Materials and Procedure

Materials consisted of a pretest, posttest, and delayed posttest, as well as different versions of *Decimal Point* or the non-game tutor and a series of pre- and post-intervention questionnaires. Instructional materials were created using an open use authoring suite [29]. We presented all materials in an established online learning management system using the HTML/JavaScript framework [30].

Learning Materials. Learning materials consisted of either *Decimal Point* or a non-game tutor designed to be equivalent to *Decimal Point* in the instructional content but without the game elements or playful design. Both *Decimal Point* and the non-game tutor varied in whether they included self-explanation prompts or not.

Decimal Point is a web-based single-player game that uses an amusement park metaphor to teach middle school students about decimal numbers, as shown in Figure 1 [12]. *Decimal Point* is made up of 24 mini-games, with each mini-game focusing on a specific type of problem-solving task. In total, there are 48 decimal problems, with two problems in each of the 24 mini-games. These problems cover various tasks such as ordering decimals, placing them on a number line, completing sequences, sorting them into “buckets” based on magnitude, and adding decimals. Players must provide the correct answer for each problem in order to progress in the game. The aim is to play through all the mini-games in sequence. The game provides immediate accuracy feedback and allows students to retry problems until they find the correct solution. The game also incorporates hints, which are designed to support students’ learning progress and help them when they become stuck on a problem [31]. Students were provided with three levels of on-demand hints while solving the problems: Level 1 hints offered general reminders about relevant decimal concepts and operations; Level 2 hints provided more detailed suggestions about solution steps; and Level 3 hints provided the answer. The hints could be accessed by selecting the "Hint" button and could be navigated using the "Previous" and "Next" buttons.

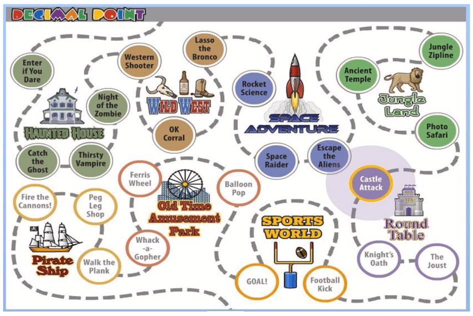
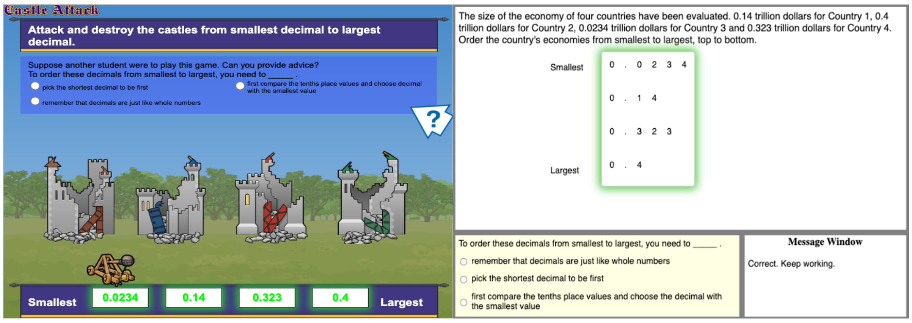


Fig. 1. The main map in *Decimal Point,* depicting the amusement park game narrative.


Castle Attack

Attack and destroy the castles from smallest decimal to largest decimal.

Suppose another student were to play this game. Can you provide advice?
To order these decimals from smallest to largest, you need to _____ .

○ pick the shortest decimal to be first
○ first compare the tenths place values and choose decimal with the smallest value
○ remember that decimals are just like whole numbers

Smallest 0.0234 0.14 0.323 0.4 Largest

The size of the economy of four countries have been evaluated. 0.14 trillion dollars for Country 1, 0.4 trillion dollars for Country 2, 0.0234 trillion dollars for Country 3 and 0.323 trillion dollars for Country 4. Order the country's economies from smallest to largest, top to bottom.

Smallest
0 . 0 2 3 4
0 . 1 4
0 . 3 2 3
0 . 4
Largest

To order these decimals from smallest to largest, you need to _____ .

○ remember that decimals are just like whole numbers
○ pick the shortest decimal to be first
○ first compare the tenths place values and choose the decimal with the smallest value

Message Window

Correct. Keep working.

Fig. 2. Self-explanation prompts at the end of the *Castle Attack* mini-game in *Decimal Point* (left) and at the end of an equivalent sorting problem in the non-game tutor (right).

After solving each pair of mini-game problems, students in the self-explanation condition were prompted to self-explain using multiple-choice questions (Figure 2), which encouraged them to think more deeply about the concepts and misconceptions targeted by the problems. Prior work has found multiple-choice self-explanation prompts to be more effective than open-ended self-explanation prompts in digital learning games, possibly because this format is less disruptive to game flow or because it introduces less cognitive load to respond [20]. In the game condition without self-explanation, the game moved on to the map without any prompting for self-explanation.

The non-game version of *Decimal Point* uses the same web-based learning management system as the game version. In the non-game version, students solve decimal problems using a conventional user interface without the game features or narrative, as shown in Figure 2 [12]. The non-game condition presents the same five types of decimal problems in the same order, with a total of 48 problems. As in the game condition, students received immediate feedback on accuracy, could access the same three levels of on-demand hints, and had to answer correctly to progress to the next problem. In the condition with self-explanation, students were prompted to self-explain decimal concepts and misconceptions with the same multiple-choice questions used in the game condition. In the non-game condition without self-explanation, the tutor moved on to the next problem set without prompting for self-explanation.

Learning Assessment Tests. We assessed students' knowledge of relevant decimal concepts using three isomorphic tests administered immediately before the learning materials, immediately after the learning materials, and approximately one week after students completed the learning materials. Tests were counterbalanced to account for any differences in difficulty. Each test contained 43 items targeting students' procedural and conceptual knowledge about decimal number operations. Some items contained multiple parts, and students could earn a total of 52 points.

Questionnaires. Questionnaires were administered immediately before students began the learning intervention (after the pretest) and immediately after they concluded the learning intervention (before the posttest). After the pretest, students responded to a series of demographic questions concerning their age, grade level, gender identity, and race. They also completed an adapted version of the Children's Occupational Interests, Activities, and Traits - Personal Measure (COAT-PM) to measure students' gender-typed characteristics [7]. This survey assesses children's interests, activities, and traits in relation to masculine- and feminine-stereotyped norms on a four-point Likert-type scale, with 18 items each in the *occupation, activity,* and *traits* subscales. Items in the *occupation* subscale targeted stereotypically gendered professions like "hairstylist" (feminine) and "construction worker" (masculine), and students reported their interests in the targeted professions on a scale from 1 (not at all) to 4 (very much). Items in the *activity* subscale targeted stereotypically gendered activities such as "making jewelry" (feminine) and "going fishing" (masculine), and students reported the frequency with which they engaged in these activities on a scale of 1 (never) to 4 (very often). Items in the *traits* subscale targeted self-perceptions of stereotypically gendered personal characteristics such as "gentle" (feminine) and "adventurous" (masculine). All subscales were averaged together to produce scales of feminine-typed characteristics ($\alpha = 0.85$) and masculine-typed characteristics ($\alpha = 0.81$).

After completing the learning intervention, students responded to a series of surveys targeting their engagement, enjoyment, and emotions, including affective and behavioral/cognitive engagement [32]; dimensions of meaning, mastery, and challenge from the Player Experience Inventory [33]; situational interest [34]; the enjoyment dimension of the Achievement Emotions Questionnaire [35]; evaluation apprehension and test anxiety [24]; and state anxiety [36]. Due to space constraints, we do not report results related to these questionnaires.

3 Results

To assess whether students learned from the intervention materials, we conducted repeated-measures analyses of variance (ANOVAs) examining changes from pretest to posttest and pretest to delayed posttest. Results indicated a large effect from pretest to posttest, $F = 97.88$, $p < .001$, $\eta^2_p = .20$, and between pretest and delayed posttest, $F = 128.33$, $p < .001$, $\eta^2_p = .25$, indicating that students generally learned from the intervention materials (see Table 1 for test performance means by condition and test).

Table 1. Average test score by gender and condition.

	N	Pretest *M (SD)*	Posttest *M (SD)*	Delayed Posttest *M (SD)*
Game conditions	203	22.32 (12.10)	25.40 (12.19)	26.12 (12.91)
Game +SE	105	21.36 (11.56)	24.61 (12.14)	25.07 (13.29)
Game -SE	98	23.35 (12.63)	26.26 (12.25)	27.26 (12.45)
Non-game conditions	183	20.83 (10.20)	23.32 (10.66)	24.34 (11.10)
Non-game +SE	96	20.93 (10.67)	23.43 (10.36)	24.67 (10.73)
Non-game -SE	87	20.72 (9.71)	23.21 (11.03)	23.99 (11.53)
+SE conditions	201	21.15 (11.12)	24.04 (11.31)	24.88 (12.11)
-SE conditions	185	22.11 (11.40)	24.82 (11.76)	25.72 (12.11)
Girls	182	20.00 (9.70)	23.71 (10.34)	24.52 (11.42)
Boys	201	23.27 (12.27)	25.28 (12.40)	26.20 (12.60)

3.1 Condition effects on learning (RQ1)

These results contradicted prior research showing a significant learning advantage for students playing *Decimal Point* compared to the non-game tutor [12]. To understand why the results might be different, we conducted a post hoc analysis of students' use of hints, as the hint feature was added in recent years after the initial research showing an advantage for the game compared to the non-game tutor.

To test whether hint requests mediated the effect of learning platform on test performance, we built mediation models with the learning platform (game or non-game tutor) as an independent variable, the number of hint requests during intervention as a mediator, and the posttest and delayed posttest scores as the dependent variables. The confidence interval of the indirect effect was estimated at the 0.05 significance level via bias-corrected non-parametric bootstrapping with 2000 iterations [37]. Based on the mediation results (Figure 3), we found that the effect of the learning platform on posttest performance was mediated by the number of hint requests. The regression coefficient between the learning platform (with the game coded as 1) and number of hint requests was positive and significant, while the coefficient between the number of hint requests and posttest score was negative and significant. In other words, students in the game tended to request more hints, which in turn was associated with worse learning outcomes. The bootstrap procedures indicated a significant indirect effect (*ab* = -0.87, 95% CI [-1.45, -0.29], $p < .001$).

Similar findings were observed in the mediation model predicting delayed posttest scores, with a significant indirect effect of the number of hint requests (*ab* = -0.97, 95% CI [-1.65, -0.41], $p < .001$; see Figure 3). On the other hand, the direct effects of the game on posttest and delayed posttest performance, without considering the mediator, were positive and significant.

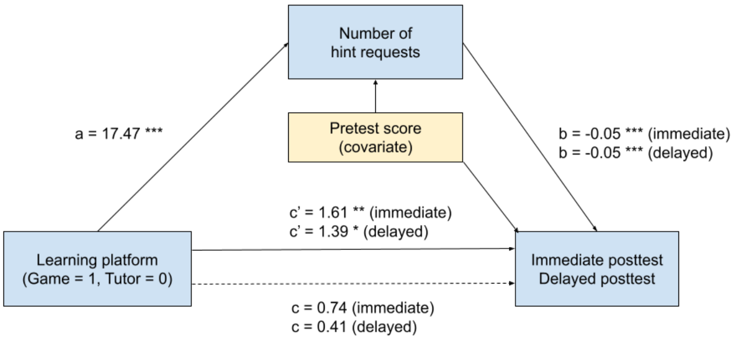


Fig. 3. The mediation pathway from learning platform to posttest and delayed posttest performance through hint usage behavior. (*) $p < .05$, (**) $p < .01$, (***) $p < .001$.

We built regression models predicting posttest and delayed posttest scores with pretest scores, number of hint requests, and learning platform (game coded as 1) to assess the unique contributions of each factor. When predicting posttest performance, number of hints requested was negatively associated with test performance (β = -0.05, p < .001); however, when controlling for hint requests and pretest (β = 0.63, p < .001), the game platform was associated with higher posttest scores (β = 1.61, p = .001). A regression model predicting delayed posttest performance with pretest, hints requests, and learning platform showed similar results: the number of hint requests was negatively associated with delayed posttest scores (β = -0.06, p < .001), but when controlling for pretest (β = 0.62, p < .001) and hint requests, the game platform was associated with higher delayed posttest scores (β = 1.39, p = .018).

In other words, while the game did lead to better learning when controlling for hint requests, students playing the game also requested more hints than those using the tutor, which in turn reduced their learning. Likely due to these conflicting trends, the total effect of the learning platform on test performance was not significant.

3.2 Binary gender differences in learning (RQ2)

We sought to understand how different dimensions of gender related to learning outcomes across instructional conditions. First, we examined differences in pretest, posttest, and delayed posttest scores based on binary gender identity; given the small number of students identifying as non-binary or trans or declining to indicate gender (N = 3), we excluded these students from analyses using binary gender identity. A one-way ANOVA revealed a significant effect of gender identity on pretest performance (F = 8.27, p = .004, η^2_p = .021), with boys receiving higher pretest scores than girls. Table 1 reports mean scores on all tests based on binary gender identity.

We also tested the interaction of binary gender with instructional condition on test scores. A series of three-way ANCOVAs examined the effects of learning platform, self-explanation condition, and binary gender on posttest and delayed posttest scores while controlling for pretest. On the posttest, results indicated no main effect of gender identity (F = 2.28, p = .13, η^2_p = .006) and no interactions between gender and learning platform (F = 0.63, p = .43, η^2_p = .002) or gender and self-explanation (F =

1.25, p = .26, η^2_p = .003). The analysis of delayed posttest also revealed no main effect of gender identity (F = 3.03, p = .08, η^2_p = .008) and no interactions between gender identity and learning platform (F = 0.05, p = .82, η^2_p < .001) or self-explanation prompts (F = 2.86, p = .09, η^2_p = .008). Our predictions that girls would learn more, but only in the game condition with self-explanation, were not supported.

3.3 Predicting learning differences with gender-typed characteristics (RQ3)

Next, we examined the correlations between binary gender identity and gender-typed characteristics. Results showed that gender identity, where "female" was coded as 1 and "male" coded as 0, was strongly, positively correlated with feminine-typed characteristics (r = 0.58, p < .001) and moderately, negatively correlated with masculine-typed characteristics (r = -0.34, p < .001). Feminine-typed characteristics were weakly, positively correlated with masculine-typed characteristics (r = 0.20, p < .001). Given the correlation coefficients, while the three gender dimensions were moderately correlated, they were not redundant.

We also analyzed test performance using the measure of gender-typed interests, activities, and traits. A regression model predicting pretest scores based on masculine-typed and feminine-typed characteristics revealed that masculine-typed characteristics were a significant, positive predictor (β = 3.64, p = .003) while feminine-typed characteristics were a significant, negative predictor (β = -4.42, p < .001).

To compare the predictive value of binary gender identity and gender-typed characteristics, we built a regression model predicting pretest scores based on binary gender identity (with "female" coded as 1 and "male" coded as 0), masculine-typed characteristics, and feminine-typed characteristics. Results showed that masculine-typed characteristics (β = 4.01, p = .008) and feminine-typed characteristics (β = -4.96, p = .004) were significant predictors, while binary gender identity was not (β = 0.75, p = .66). These results suggest the measures of gender-typed characteristics may better explain variance in pretest scores than binary gender identity.

To understand interactions between the multiple dimensions of gender and learning conditions, we built regression models predicting posttest and delayed posttest scores that included pretest score as a covariate and the following predictor variables: learning platform (game or non-game), self-explanation prompt (with or without self-explanation), masculine-typed characteristics, feminine-typed characteristics, and their interactions with the learning platform and self-explanation conditions. Results showed that none of the variables significantly predicted posttest scores or delayed posttest scores. These results are consistent with the lack of effects revealed when examining the interaction of binary gender with the learning conditions.

We also examined self-explanation errors by gender, as well self-explanation performance as a mediator between gender and learning outcomes. We considered only students who were prompted to self-explain in the game (n = 105) and non-game (n = 96) conditions. A two-way ANCOVA assessing the effects of the learning platform and gender identity on the number of self-explanation errors, with pretest score as covariate, showed a significant main effect of gender (F = 7.53, p = .007, η^2_p = .037), with girls (M = 30.35, SD = 14.35) making fewer self-explanation errors than boys (M = 35.74, SD = 13.12). The effects of learning platform (F = 0.38, p = .54, η^2_p = .002) and its interaction with gender identity (F = 0.23, p = .63, η^2_p = .001) were not significant. We also built a regression model predicting the number of self-explanation errors based on gender-typed characteristics and their interactions with the learning platform, using pretest performance as covariate. Our results showed that masculine-typed characteristics were a marginally significant, positive predictor of self-explanation errors ($\beta = 4.42$, $p = .06$), while feminine-typed characteristics were a significant, negative predictor of self-explanation errors ($\beta = -6.05$, $p = .008$). The learning platform and its interaction with gender-typed characteristics were not significant predictors. In other words, both binary gender identity and gender-typed characteristics predicted self-explanation errors in similar patterns, with girls and students with stronger feminine-typed characteristics making fewer errors.

4 Discussion and Conclusion

This research uncovered several important findings, with clear implications for future research with *Decimal Point* as well as recommendations for research on gender and digital learning games more generally. First, we failed to replicate prior research showing a learning advantage for the game compared to a non-game tutor [12]. However, our post hoc analyses uncovered a promising explanation for this result. In the time since the original study, we added hints to *Decimal Point* to help students avoid getting stuck and increase opportunities to learn from errors [31]. Based on hint use patterns in the game and non-game platforms, however, it appears that using too many hints is generally harmful to learning, which is consistent with other research on hint use [3, 38]. Students in the game tended to overuse hints *more* than students in the non-game, and this overuse of hints seemed to negate the benefits of learning with the game. As additional evidence, when we analyzed learning outcomes from the game vs. non-game and controlled for hint use, students in the game performed better than students in the non-game condition on the posttest and delayed posttest.

Second, we failed to replicate the gender effect on learning outcomes, although we replicated results showing that boys outperformed girls on the pretest. Girls have learned more than boys across many previous studies with *Decimal Point*, although this was not the case in the non-game tutor [3, 18]. We are unsure why this result failed to replicate in the game condition with self-explanation. It is possible that the 2x2 design was underpowered to detect interaction effects with gender, especially after a technical error caused data loss for some students. This unexpected result merits additional investigation, particularly given how seemingly unrelated changes to the game (e.g., the addition of hints) affected other outcomes. We replicated results that girls made fewer errors on self-explanation steps compared to boys, which contributes to evidence that the self-explanation step may be responsible for gender differences in learning with *Decimal Point.*

Third, we have found evidence that incorporating multiple dimensions of gender is a fruitful avenue for better understanding gender differences in learning. While binary gender was correlated with the measures of gender-typed characteristics (strongly with feminine-typed characteristics and moderately with the masculine-typed characteristics), results indicated that between about 40 and 65 percent of the variance in masculine- and feminine-typed occupational interests, activities, and traits was *not* explained by binary gender. These more nuanced aspects of gender, which are captured on a spectrum and target a wide range of interests and activities, are likely to more directly shape individual learning experiences that, in turn, predict students' math motivation and achievement in different learning environments. Critically, the measures of gender-typed occupational interests, activities, and traits explained differences on the pretest better than binary gender; when the masculine- and feminine-typed characteristics were included in a regression model predicting pretest performance, binary gender was no longer a significant predictor.

Results strongly support the use of multiple dimensions of gender in future research investigating gender differences in game-based learning. This is a novel contribution, as we know of no other research that has adopted such an approach to studying gender in the context of digital learning games. This approach is likely to provide a better explanation for gender-based differences in learning behaviors and outcomes compared to binary measures of gender identity, which in turn may illuminate clearer recommendations for making digital learning games more equitable and beneficial for all students. Our results also reveal the importance of examining each design choice in terms of its impact on learning behaviors and outcomes. Specifically, our results suggest that adding hints was particularly detrimental to students in the game conditions, who used (and likely abused) hints more than students in the non-game conditions. Future research should explore why students might be more likely to overuse hints in a digital learning game compared to a non-game tutor.

Acknowledgments. This work was supported by the National Science Foundation Award #DRL-2201796. The opinions expressed are those of the authors and do not represent the views of NSF. Thanks to Jimit Bhalani, John Choi, Kevin Dhou, Darlan Santana Farias, Rosta Farzan, Jodi Forlizzi, Craig Ganoe, Rick Henkel, Scott Herbst, Grace Kihumba, Kim Lister, Patrick Bruce Gonçalves McLaren, and Jon Star for important contributions to the development and early experimentation with *Decimal Point*.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arroyo, I., Burleson, W., Tai, M., Muldner, K., Woolf, B.P.: Gender differences in the use and benefit of advanced learning technologies for mathematics. J. Educ. Psychol, **105,** 4, 957 (2013). doi: 10.1037/a0032748
2. Dele-Ajayi, O., Strachan, R., Pickard, A., Sanderson, J.: Designing for All: Exploring Gender Diversity and Engagement with Digital Educational Games by Young People. In: 2018 IEEE Frontiers in Education Conference (FIE), 1–9 (2018).
3. Nguyen, H., Hou, X., Richey, J.E., McLaren, B.M.: The impact of gender in learning with games: A consistent effect in a math learning game. Int. J. Game-Based Learn. **12,** 1, 1-29 (2022). doi: 10.4018/IJGBL.309128
4. Baker, R. S., Corbett, A. T., Koedinger, K.R., Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: when students "game the system." In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 383-390 (2004). doi: 10.1145/985692.985741
5. Baker, R.S., Richey, J.E., Zhang, J., Karumbaiah, S., Andres-Bray, M., Nguyen, H. A., Andres, J.M. A.L., McLaren, B.M.: Gaming the system mediates the relationship between gender and learning outcomes in a digital learning game. (under review).

6. Hyde, J.S., Bigler, R.S., Joel, D., Tate, C.C., van Anders, S.M.: The future of sex and gender in psychology: Five challenges to the gender binary. Am. Psychol. **74**, 2, 171 (2019). doi: 10.1037/amp0000307
7. Liben, L.S., Bigler, R.S., Ruble, D.N., Martin, C.L., Powlishta, K.K.: The developmental course of gender differentiation: Conceptualizing, measuring, and evaluating constructs and pathways. Monogr. Soc. Res. Child Dev. **67**, 2, 1–183 (2002).
8. Clark, D.B., Tanner-Smith, E., Killingsworth, S.: Digital games, design, and learning: A systematic review and meta-analysis. Rev. Educ. Res **86,** 1, 79-122 (2016). doi: 10.3102/0034654315582065
9. Mayer, R.E.: Computer games in education. Annu. Rev. Psychol. **70**, 531–49 (2019). doi: 10.1146/annurev-psych-010418-102744
10. Hussein, M.H., Ow, S.H., Elaish, M.M., Jensen, E.O.: Digital game-based learning in K-12 mathematics education: A systematic literature review. Educ. Inf. Technol. **27**, 2859–2891 (2022). doi: 10.1007/s10639-021-10721-x
11. Schöbel, S., Saqr, M., Janson, A.: Two decades of game concepts in digital learning environments–A bibliometric study and research agenda. Comput. Educ. **173**, 104296 (2021). doi: 10.1016/j.compedu.2021.104296
12. McLaren, B.M., Adams, D.M., Mayer, R.E., Forlizzi, J.: A computer-based game that promotes mathematics learning more than a conventional approach. Int. J. Game-Based Learn. **7**, 1, 36-56 (2017). doi: 10.4018/IJGBL.2017010103
13. Hooshyar, D., Malva, L., Yang, Y., Pedaste, M., Wang, M., Lim, H.: An adaptive educational computer game: Effects on students' knowledge and learning attitude in computational thinking. Comput. Human Behav. **114**, 106575 (2021). doi: 10.1016/j.chb.2020.106575
14. Lester, J.C., Ha, E.Y., Lee, S.Y., Mott, B.W., Rowe, J.P., Sabourin, J.L.: Serious games get smart: Intelligent game-based learning environments. AI Magazine. **34**, 4, 31-45 (2013). doi: 10.1609/aimag.v34i4.2488
15. Gee, J.P.: What video games have to teach us about learning and literacy. Palgrave Macmillan, New York (2007).
16. Czikszentmihalyi, M.: Flow: The psychology of optimal experience. New York, Harper & Row (1990).
17. Ronimus, M., Eklund, K., Pesu, L., Lyytinen, H.: Supporting struggling readers with digital game-based learning. Educ. Technol. Res. Dev. **67**, 639-663 (2019). doi: 10.1007/s11423-019-09658-3
18. McLaren, B.M., Farzan, R., Adams, D.M., Mayer, R.E., Forlizzi, J.: Uncovering gender and problem difficulty effects in learning with an educational game. In: E. André, R. Baker, X. Hu, M.M.T. Rodrigo, and B. du Boulay (Eds.). Proceedings of the 18th International Conference on Artificial Intelligence in Education (AIED 2017). LNAI, vol. 10331, pp. 540-543. Springer, Berlin (2017). doi: 10.1007/978-3-319-61425-0_59
19. McLaren, B.M., Richey, J.E., Nguyen, H.A., Mogessie, M.: A digital learning game for mathematics that leads to better learning outcomes for female students: Further evidence. In: Proceedings of the 16th European Conference on Game Based Learning (ECGBL 2022). pp. 339-348 (2022). doi: 10.34190/ecgbl.16.1.794
20. Johnson, C.I., Mayer, R.E.: Applying the self-explanation principle to multimedia learning in a computer-based game-like environment. Comput. Human Behav. **26**, 6, 1246–1252 (2010). doi: 10.1016/j.chb.2010.03.025
21. Nokes, T.J., Hausmann, R.G., VanLehn, K., Gershman, S. Testing the instructional fit hypothesis: the case of self-explanation prompts. Instr. Sci **39**, 645-666 (2011). doi: 10.1007/s11251-010-9151-4
22. Else-Quest, N.M., Hyde, J.S., Linn, M.C.: Cross-national patterns of gender differences in mathematics: a meta-analysis. Psychol. Bull. **136**, 1, 103 (2010). doi: 10.1037/a0018053

23. Else-Quest, N.M., Mineo, C.C., Higgins, A.: Math and Science Attitudes and Achievement at the Intersection of Gender and Ethnicity. Psychol. Women Q. **37**, 3, 293–309 (2013). doi: 10.1177/0361684313480694
24. Spencer, S.J., Steele, C.M., Quinn, D.M.: Stereotype threat and women's math performance. J. Exp. Soc. Psychol, **35**, 1, 4-28 (1999). doi: 10.1006/jesp.1998.1373
25. Doyle, R.A., Voyer, D.: Stereotype manipulation effects on math and spatial test performance: A meta-analysis. Learn. Individ. Differ. **47**, 103-116 (2016). doi: 10.1016/j.lindif.2015.12.018
26. McLaren, B.M., Nguyen, H.A.: Digital learning games in Artificial Intelligence in Education (AIED): A review. In: B. du Boulay, A. Mitrovic, & K. Yacef (Eds.), Handbook of Artificial Intelligence in Education. 440-484 (2023). doi: 10.4337/9781800375413.00032
27. Nguyen, H., Else-Quest, N., Richey, J.E., Hammer, J., Di, S., McLaren, B.M.: Gender differences in learning game preferences: Results using a multi-dimensional gender framework. Proceedings of 24th International Conference on Artificial Intelligence in Education (AIED 2023). pp. 553-564 (2023). doi: 10.1007/978-3-031-36272-9_45
28. Arroyo, I., Burleson, W., Tai, M., Muldner, K., Woolf, B.P.: Gender differences in the use and benefit of advanced learning technologies for mathematics. J. Educ. Psychol. **105**, 4, 957 (2013). doi: 10.1037/a0032748
29. Aleven, V., McLaren, B.M., Sewall, J., Van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., Koedinger, K.R.: Example-tracing tutors: Intelligent tutor development for non-programmers. Int. J. Artif. Intell. Educ. **26**, 224-269 (2016). doi: 10.1007/s40593-015-00882
30. Aleven, V., Mclaren, B.M., Sewall, J., Koedinger, K.R.: A new paradigm for intelligent tutoring systems: Example-tracing tutors. Int. J. Artif. Intell. Educ **19**, 2, 105-154 (2009).
31. Lester, J.C., Spain, R.D., Rowe, J.P., Mott, B.W.: Instructional support, feedback, and coaching in game-based learning. Handbook of game-based learning, 209-237 (2020).
32. Ben-Eliyahu, A., Moore, D., Dorph, R., Schunn, C.D.: Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. Contemp. Educ. Psychol. **53**, 87–105 (2018). doi: 10.1016/j.cedpsych.2018.01.002
33. Abeele, V.V., Spiel, K., Nacke, L., Johnson, D., Gerling, K.: Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. Int. J. Hum. Comput. Stud. **135,** 102370 (2020). doi: 10.1016/j.ijhcs.2019.102370
34. Linnenbrink-Garcia, L., Durik, A.M., Conley, A.M., Barron, K.E., Tauer, J.M., Karabenick, S.A., Harackiewicz, J.M.: Measuring Situational Interest in Academic Domains. Educ. Psychol. Meas. **70**, 4, 647–671 (2010). doi: 10.1177/0013164409355699
35. Pekrun, R., Goetz, T., Frenzel, A.C., Barchfeld, P., Perry, R.P.: Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). Contemp. Educ. Psychol **36**, 1, 36-48 (2011). doi: 10.1016/j.cedpsych.2010.10.002
36. Chung, B.G., Ehrhart, M.G., Holcombe Ehrhart, K., Hattrup, K., Solamon, J.: Stereotype Threat, State Anxiety, and Specific Self-Efficacy as Predictors of Promotion Exam Performance. Group Organ. Manag. **35**, 1, 77–107 (2010). doi: 10.1177/1059601109354839
37. Hayes, A.F., Rockwood, N.J.: Regression-based statistical mediation and moderation analysis in clinical research: Observations, recommendations, and implementation. Behav. Res. Ther. **98**, 39-57 (2017). doi: 10.1016/j.brat.2016.11.001
38. Muir, M., Conati, C.: An analysis of attention to student–adaptive hints in an educational game. In: Intelligent Tutoring Systems: 11th International Conference, ITS 2012, Chania, Crete, Greece, June 14-18, pp. 112-122. Springer, Berlin Heidelberg (2012).