Gender Differences in Learning and Test-Taking Experiences with Digital Game-Based Assessment

Aditi Haiman[1(✉)], Esma Kahveci[2], Alessandro Pagano[3], Angie Chi[1], and Bruce M. McLaren[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
{ahaiman,achi2,bmclaren}@andrew.cmu.edu
[2] University of Massachusetts Dartmouth, Dartmouth, MA, USA
[3] University of Bari Aldo Moro, Bari, Italy
alessandro.pagano@uniba.it

Abstract. Mathematics assessment is not a neutral measure of students' knowledge; evaluative contexts can shape motivation, anxiety, and performance, particularly for students from groups negatively stereotyped in mathematics. Digital games offer an alternative assessment context that may reduce evaluative threat by embedding assessment in low-stakes gameplay. In this study, middle school students were randomly assigned to complete either a traditional digital test or a game-based assessment with identical mathematical content. We examined learning gains over time and students' affective experiences, including interest, anxiety, self-efficacy, and achievement emotions. Results showed no differences in learning gains between conditions, although overall scores were lower in the game-based assessment. No affective effects were statistically reliable after correcting for multiple comparisons; however, descriptive patterns suggested that girls tended to report more positive emotions and greater interest in the game-based condition. These findings indicate that assessment format may influence how students experience evaluation without altering short-term learning outcomes, highlighting the importance of considering affective factors in assessment design.

Keywords: Game-based assessment · Affective experiences · Stereotype threat · Student motivation · Assessment design

1 Introduction

Mathematics assessment is often treated as a neutral measure of students' knowledge, but research in mathematics education and educational psychology shows that assessment contexts shape how students perform and experience mathematics [2,10]. Evaluative settings can influence motivation, anxiety, and engagement, and may disproportionately affect students from groups that are negatively stereotyped in mathematics, particularly girls [17]. For example, stereotype threat research suggests that when performance is framed as diagnostic

E. G. Blanchard et al. (Eds.): AIED 2026, LNAI 16585, pp. 279–288, 2027.
https://doi.org/10.1007/978-3-032-29770-9_31

of ability, students may experience increased anxiety and cognitive load, which can impair performance [14]. Because these effects are tied to the social meaning of evaluation rather than task difficulty alone, changing how assessment is presented may influence both performance and affective experiences.

Digital games offer a potential alternative to traditional assessments by embedding evaluation within low-stakes, interactive contexts. When assessment is framed as gameplay rather than testing, students may experience reduced evaluative pressure, which can in turn influence motivation and emotional responses [15]. Prior work on game-based learning environments has shown that affective and behavioral differences can emerge across genders. In a series of studies with the digital mathematics game *Decimal Point*, girls demonstrated larger learning gains than boys despite lower initial performance. Follow-up analyses suggested that these differences were not explained by game narrative or design features alone, but were associated with differences in engagement, anxiety, and self-efficacy [11,13]. These findings raise the possibility that gender differences in learning may be partly driven by how evaluative a task feels, rather than by the instructional content itself. Although prior work has explored interventions that reduce stereotype threat through psychological framing (e.g., growth mindset messaging or self-affirmation) [5,9], these approaches typically retain conventional test formats. Other studies have examined low-stakes framing (e.g., presenting assessments as practice), but still rely on traditional test interfaces [7]. In contrast, *game-based assessment* delivers assessment through game mechanics and presentation while preserving the measurement intent and content of conventional tests [4]. By reducing evaluative salience without altering task content, game-based assessment provides a way to isolate how assessment format itself influences students' experiences.

In this study, we investigate whether using a game-based assessment as the *sole evaluative measure* affects students' learning outcomes and test-taking experiences, with a particular focus on gender differences. Middle school students were randomly assigned to complete either a traditional digital test or a content-equivalent game-based assessment presented as low-stakes gameplay. Both conditions used identical mathematical items, allowing us to isolate the effect of assessment format. Rather than directly measuring stereotype threat appraisals, which may be difficult for middle school students to articulate, we examine affective and motivational outcomes (e.g., anxiety, interest, and self-efficacy) that are commonly associated with evaluative threat.

We address the following research questions:

1. How does assessment format (game-based vs. traditional) affect performance and learning gains over time when holding content constant?
2. How does assessment format affect girls' and boys' affective experiences during assessment?

2 Methods

2.1 Study Design

We conducted a between-subjects experiment to examine the effects of assessment format on learning and test-taking experiences. Students were randomly assigned to either a traditional digital test condition or a game-based assessment condition. Both versions were delivered digitally and contained identical mathematical content. Students in the game-based condition were not informed that they were being tested in order to reduce evaluative salience, whereas students in the traditional condition were told they would be taking a test. Researchers were present during administration but did not assist with problem solving. The study followed the sequence: pretest, intervention, posttest, and delayed test (approximately one week later), with surveys administered after each assessment.

2.2 Materials: Game-Based Assessment Design and Surveys

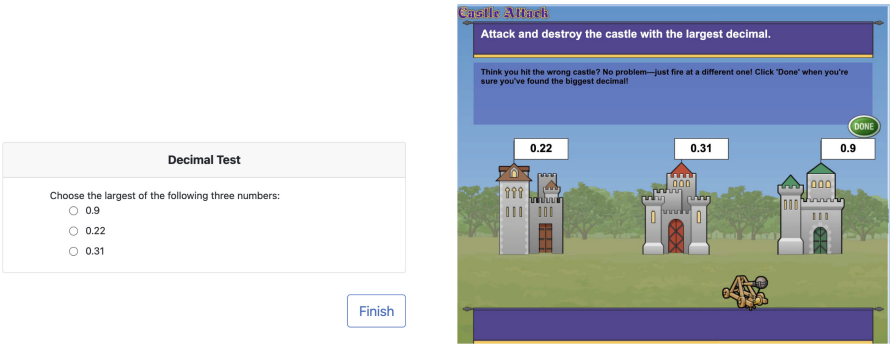
The game-based assessment presented the same items as the traditional test within an interactive game interface. Problems appeared in the same sequence with identical values and answer choices, but were embedded in a narrative-driven environment with visual progression and rewards. For example, a multiple-choice question requiring students to identify the largest decimal in the traditional-style assessment is reimagined as a medieval siege challenge (see Fig. 1). To ensure experimental consistency, the narrative elements are fixed; all students encounter the same rewards and character pop-ups at specific milestones in the game, regardless of their performance accuracy.

Students completed three isomorphic assessments (pretest, posttest, delayed), administered either as traditional tests or game-based assessments depending on condition. The assessments were identical in content, item order, and answer choices, and included items measuring near, medium, and far transfer [11]. Between assessments, students completed a different decimal learning game, *Ocean Adventure*. After each assessment, students completed surveys measuring engagement, situational interest, test anxiety, and self-efficacy using validated multi-item Likert scales [1,8,17]. The 9-item TIMSS "Mathematics in school" student questionnaire was used to measure math self-concept and self-efficacy [18], and 5 items were adapted (language was simplified to be middle-school appropriate) from the Domain Identification Measure to assess math identification [16]. The reliability of these new math survey items was assessed after the study and showed that they demonstrated good internal consistency (Cronbach's $\alpha = .85$).

2.3 Participants

Participants were 196 5th and 6th grade students from four middle schools in the northeastern United States. Seventy-two students were excluded for not completing all study materials, and four additional students who did not report gender

were excluded from analyses involving gender. The final sample included 120 students (68 girls, 52 boys), with 55 assigned to the game-based condition and 65 to the traditional test condition.



(a) Standard Test Condition **(b)** Game-based test condition

Fig. 1. Comparison of standard test interface vs. game-based test interface.

3 Results

3.1 Assessment Reliability and Item Difficulty

To examine whether the game-based and traditional assessments measured the same underlying decimal concepts, we evaluated internal consistency and compared item difficulty patterns across the two test formats. Both assessments demonstrated strong internal reliability, with Cronbach's $\alpha = .84$ for the game-based assessment and $\alpha = .89$ for the traditional assessment, indicating that each test functioned as a coherent measure of decimal knowledge. To compare item difficulty across assessments while accounting for differences in scoring structure, item scores were binarized such that any non-zero score was treated as evidence of correctness. Linear regression predicting traditional item difficulty from game-based item difficulty revealed a positive relationship ($\beta = .52$, $t(38) = 5.23$, $p < .001$, $R^2 = .42$), indicating that items that were more difficult in the game-based assessment also tended to be more difficult in the traditional assessment. Together, these analyses provide evidence of comparable item functioning across conditions, though we cannot fully rule out differences in how students interpreted items in the two formats.

3.2 Effect of Condition and Gender on Learning

To examine learning over time and differences by condition and gender, we conducted a mixed ANOVA (Analysis of Variance) with Time (pretest, posttest,

delayed) as a within-subjects factor and condition and gender as between-subjects factors. There was a significant main effect of time ($F(2, 244) = 48.8, p < .001, \eta^2_G = .04$) and condition ($F(1, 122) = 17.1, p < .001, \eta^2_G = .11$), but no interaction between time and condition. There were no main or interaction effects involving gender. Post-hoc paired-samples t-tests with Bonferroni corrections showed that scores increased from pretest to posttest ($t(123) = -7.82, p < .001, d = .70$) and from pretest to delayed test ($t(123) = -8.61, p < .001, d = .77$), with no difference between posttest and delayed test, indicating that learning gains were maintained. Across all time points, students in the game-based condition had lower overall scores ($\mu = 39.17$) than those in the traditional condition ($\mu = 51.52$). Descriptive statistics are reported in Table 1.

Table 1. Means and standard deviations on each test by condition and gender (scores scaled 0–100).

Condition	Gender	N	Pretest M(SD)	Posttest M(SD)	Delayed M(SD)
Game-based	All	55	35 (15)	41 (15)	43 (17)
	Girls	29	32 (14)	39 (14)	41 (17)
	Boys	26	39 (16)	43 (16)	45 (17)
Traditional	All	65	47 (20)	54 (19)	56 (18)
	Girls	39	44 (19)	53 (17)	54 (16)
	Boys	26	51 (22)	56 (21)	57 (21)

To examine whether learning differed by condition or gender, we analyzed pre-post and pre-delayed learning gains using factorial ANOVAs with condition (test vs. game) and gender (male, female) as between-subjects factors. For pre-post gains, there was no main effect of condition, indicating comparable immediate learning gains across assessment formats. There was a significant main effect of gender, $F(1, 116) = 3.96, p = .049$, with girls improving by 8.12 points and boys improving by 4.59 points on average. Additionally, there was no interaction effect between condition and gender ($p = .86$), suggesting that immediate learning gains for boys and girls did not depend on test format. A parallel analysis of pre-delayed gains likewise revealed no main effect of condition, a marginally significant main effect of gender, $F(1, 116) = 2.89, p = .092$, and no significant interaction effect between condition and gender.

3.3 Effects of Condition and Gender on Affective Measures and Experiences

To examine whether students' affective, motivational, and self-belief measures differed by condition and gender, we conducted a series of two-way factorial ANOVAs. Each model included main effects of condition and gender, as well as interactions between condition and gender. To control for multiple comparisons across affective outcomes, Holm-Bonferroni corrections were applied within

families of tests (condition, gender, and interaction effects). After correction, no effects remained statistically significant. A summary of results can be seen in Table 2.

Table 2. Summary of ANOVA results for affective outcomes. Holm-adjusted p-values are reported within families of tests (condition, gender, interaction).

Outcome	Effect	$F(df_1, df_2)$	p	p (Holm)	η_p^2
Math Self-Concept	Condition	F(1,114)=.04	.84	1.00	.00
	Gender	F(1,114)=3.41	.067	.337	.03
	Condition × Gender	F(1,114)=.05	.82	1.00	.00
Affective Engagement	Condition	F(1,114)=.04	.85	1.00	.00
	Gender	F(1,114)=.71	.40	1.00	.01
	Condition × Gender	F(1,114)=.83	.37	1.00	.01
Situational Interest	Condition	F(1,114)=1.54	.22	1.00	.01
	Gender	F(1,114)=.04	.85	1.00	.00
	Condition × Gender	F(1,114)=3.23	.075	.37	.03
Perceived Mastery	Condition	F(1,114)=2.60	.11	.766	.02
	Gender	F(1,114)=4.58	.035	.207	.04
	Condition × Gender	F(1,114)=1.22	.27	1.00	.01
Achievement Emotions	Condition	F(1,114)=.48	.48	1.00	.00
	Gender	F(1,114)=.001	.98	1.00	.00
	Condition × Gender	F(1,114)=6.46	.012	.087	.05
Decimal Efficacy	Condition	F(1,114)= 1.23	.27	1.00	.01
	Gender	F(1,114)= 5.20	.024	.171	.04
	Condition × Gender	F(1,114)=4.58	.034	.206	.04
Test Anxiety	Condition	F(1,114)=.36	.55	1.00	.00
	Gender	F(1,114)=.23	.63	1.00	.00
	Condition × Gender	F(1,114)=.001	.92	1.00	.00

4 Discussion

4.1 RQ1: Effects of Game-Based Versus Traditional Assessment on Performance and Learning

Across all time points, students in the game-based condition scored lower than those in the traditional test condition. One possible explanation is a difference in test-taking motivation: students in the traditional test condition were explicitly told they were taking a test, which may have increased effort compared to the game-based condition, where the activity was framed as low-stakes gameplay [12,19]. Differences in presentation and perceived stakes may therefore

have influenced how seriously students approached the assessment. Relatedly, the observed performance differences may reflect the contrast between low- and high-stakes framing rather than the game-based format itself. This is something future research can focus on, to separate the effects of framing and format.

Other factors may also have contributed to differences in scores. The additional visual elements and interactive mechanics in the game-based condition may have increased cognitive load or divided attention, potentially interfering with performance even though the underlying content was identical. Students in the game-based test may have also prioritized the gameplay rather than problem-solving (e.g. through exploration or trial-and-error). Alternative explanations related to content differences are less likely. Both conditions used identical items in the same order, and item-level analyses showed similar difficulty patterns across formats, suggesting that performance differences were not due to question difficulty. Students were also randomly assigned to conditions, which reduces the likelihood of systematic differences in prior knowledge, although the lack of a baseline equivalence test means this possibility cannot be fully ruled out.

Despite these differences in overall scores, learning gains did not differ by condition. Students in both groups improved over time, indicating that the game-based assessment did not hinder learning. Notably, lower performance in the game-based condition did not correspond to reduced learning, highlighting that assessment format can influence observed performance without necessarily reflecting differences in underlying knowledge. This difference between performance and learning suggests that assessment outcomes may be sensitive to how assessment contexts shape effort, attention, and response strategies. If the format of an assessment can meaningfully alter student behavior and resulting scores, it becomes especially important to carefully consider how assessment design may introduce systematic biases in the measurement of student knowledge.

4.2 RQ2: Effects of Assessment Format on Girls' and Boys' Learning and Affective Experiences

Across assessment formats, girls showed higher learning gains than boys, consistent with prior work using *Decimal Point* [11]. However, we did not observe an interaction between assessment format and gender, suggesting that neither format differentially benefited boys or girls in terms of learning. In contrast, affective outcomes showed suggestive gender-related patterns. Although no effects were statistically reliable after correcting for multiple comparisons, descriptive trends were consistent with our hypotheses. In particular, girls tended to report more positive achievement emotions and higher situational interest in the game-based condition, whereas boys' responses were more stable across formats or higher in the traditional test condition for measures of efficacy. These patterns suggest that the game-based, low-stakes evaluative context may have altered how girls experienced the assessment, even though these differences were not statistically robust. These trends align with sociocultural accounts of gendered participation in mathematics, which argue that mathematics is often positioned

as a masculine domain through norms emphasizing speed, individual performance, and evaluative comparison [3]. Such norms can heighten evaluative threat and reduce feelings of belonging for girls during traditional assessments [17]. From this perspective, reframing assessment as low-stakes gameplay may weaken these associations, potentially creating conditions that support more positive affective experiences for girls. At the same time, descriptive differences in perceived mastery persisted across conditions, with boys tending to report higher perceived mastery than girls. Although this pattern was not statistically reliable after correction, it is consistent with prior work suggesting that perceptions of mathematical competence are shaped by broader cultural narratives that may not be easily shifted by changes in assessment format alone. Additionally, the lack of observed differences in test anxiety may reflect the nature of the measure used, which captures broader social-evaluative concerns rather than moment-to-moment task-related anxiety.

Finally, although girls tended to report more positive affective experiences in the game-based condition, these differences did not correspond to differences in learning outcomes. This pattern is consistent with prior work suggesting that affective changes may influence engagement and persistence over longer periods of time rather than producing immediate gains [6]. Taken together, these findings suggest that game-based assessments may shape how students experience evaluation even when short-term learning outcomes remain unchanged.

5 Conclusion

This study examined whether assessment format influences students' learning and affective experiences in mathematics. While students in both conditions demonstrated similar learning gains, overall performance was lower in the game-based condition, suggesting that the assessment format may influence how knowledge is expressed without necessarily affecting learning itself. Although no affective differences were statistically reliable after correcting for multiple comparisons, descriptive patterns indicated that girls tended to report more positive emotions and greater interest in the game-based assessment, consistent with theories of evaluative threat and gendered experiences in mathematics. These findings suggest that assessment design may shape how students experience evaluation, even when immediate learning outcomes remain unchanged. However, this study is limited by attrition due to classroom disruptions, reliance on self-report measures, and a focus on a single content area and age group. Future research should examine whether such affective patterns persist over time and explore how game-based assessments can be designed to support engagement without reducing effort, as well as how assessment format interacts with broader social and cultural factors that influence students' experiences in mathematics.

Acknowledgement. Thanks to Marshall An for help with data extraction. ChatGPT was used to edit and revise this paper.

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

References

1. 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)
2. Black, P., Wiliam, D.: Assessment and classroom learning. Assessment Educ. Principles, Policy Pract. **5**(1), 7–74 (1998)
3. Boaler, J.: The development of disciplinary relationships: knowledge, practice and identity in mathematics classrooms. Learn. Math. **22**(1), 42–47 (2002)
4. Gomez, M.J., Ruipérez-Valiente, J.A., Clemente, F.J.G.: A systematic literature review of game-based assessment studies: trends and challenges. IEEE Trans. Learn. Technol. **16**(4), 500–515 (2022)
5. Good, C., Aronson, J., Inzlicht, M.: Improving adolescents' standardized test performance: an intervention to reduce the effects of stereotype threat. J. Appl. Dev. Psychol. **24**(6), 645–662 (2003)
6. Hidi, S., Renninger, K.A.: The four-phase model of interest development. Educ. Psychol. **41**(2), 111–127 (2006)
7. Johns, M., Schmader, T., Martens, A.: Knowing is half the battle: teaching stereotype threat as a means of improving women's math performance. Psychol. Sci. **16**(3), 175–179 (2005)
8. Linnenbrink-Garcia, L., et al.: Measuring situational interest in academic domains. Educ. Psychol. Measur. **70**(4), 647–671 (2010)
9. Martens, A., Johns, M., Greenberg, J., Schimel, J.: Combating stereotype threat: the effect of self-affirmation on women's intellectual performance. J. Exp. Soc. Psychol. **42**(2), 236–243 (2006)
10. McLeod, D.B.: Research on affect in mathematics education: a reconceptualization (2006)
11. Nguyen, H.A., Hou, X., Richey, J.E., McLaren, B.M.: The impact of gender in learning with games: a consistent effect in a math learning game. Inter. J. Game-Based Learn. (IJGBL) **12**(1), 1–29 (2022)
12. Penk, C., Richter, D.: Change in test-taking motivation and its relationship to test performance in low-stakes assessments. Educ. Assess. Eval. Account. **29**(1), 55–79 (2017)
13. Richey, J.E., et al.: Impact of digital learning game narratives on affective experiences and learning. In: International Conference on Computers in Education (2025)
14. Schmader, T., Johns, M., Forbes, C.: An integrated process model of stereotype threat effects on performance. Psychol. Rev. **115**(2), 336 (2008)
15. Shute, V.J.: Stealth assessment in computer-based games to support learning. Comput. Games Instruct. **55**(2), 503–524 (2011)
16. Smith, J.L., White, P.H.: Development of the domain identification measure: a tool for investigating stereotype threat effects. Educ. Psychol. Measur. **61**(6), 1040–1057 (2001)

17. 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)
18. TIMSS and PIRLS International Study Center: Student Questionnaire: Grade 4. International Association for the Evaluation of Educational Achievement (IEA), Chestnut Hill, MA (2014)
19. Wise, S.L., DeMars, C.E.: Low examinee effort in low-stakes assessment: problems and potential solutions. Educ. Assess. **10**(1), 1–17 (2005)