

Can Games Reduce Inequity? Evaluative Framing Alone Does Not Address Racial Disparities in Assessment

Aditi Haiman, Angie Chi, Bruce M. McLaren

Abstract

Digital learning games are often proposed as a way to improve student engagement and reduce inequities by reframing learning and assessment as playful, low-stakes activities. In particular, game-based assessment has been suggested as a means of reducing the effects of evaluative pressure, which tend to disproportionately affect students from historically marginalized groups. However, it remains unclear whether simply changing how evaluative contexts are framed is sufficient to produce more equitable outcomes. In this study, we examine the effects of game-based versus traditional assessment on learning, affective experience, and racial disparities in a middle school mathematics context. Students were randomly assigned to complete assessments either as a traditional digital test or as part of a digital learning game. Both conditions contained identical mathematical content, allowing us to control for instructional differences while varying how the activity was framed and experienced. Students completed pretest, posttest, and delayed assessments, along with surveys measuring engagement, anxiety, self-efficacy, and related affective outcomes. Results showed that students in both conditions demonstrated comparable learning gains over time, indicating that assessment format did not affect learning. However, students in the traditional condition achieved higher overall scores. Contrary to expectations, there were no significant differences in affective or motivational experiences between conditions after correcting for multiple comparisons. Importantly, persistent racial disparities in both performance and affective measures were observed across conditions and were not reduced by presenting the assessment as a game. These findings suggest that in the context of this study and this experimental population, reframing assessment through game-based presentation alone is insufficient to meaningfully change student experience or address existing inequities. From a design perspective, this work highlights the need to move beyond surface-level notions of gamefulness and toward a deeper understanding of which features of game-based systems are necessary to support equitable learning and assessment experiences.

Introduction

Digital learning games are often positioned as a way to make learning more engaging, motivating, and effective (Clark et al., 2016; Hussein et al., 2022). Prior work has shown that well-designed educational games can support learning outcomes and improve motivational experiences compared to traditional instructional approaches (Mayer, 2019; McLaren, 2024). However, these benefits are not uniformly distributed across learners (Ni et al., 2024), and persisting disparities in mathematics achievement and experiences across demographic groups (Reardon, 2011) raise an open question: under what conditions do game-based learning environments meaningfully improve equity, and when do they fall short?

A central but often underexamined assumption in game-based learning research is that framing an activity as a game, rather than as a test or formal assessment, changes how learners experience that activity. Games are frequently described as “low-stakes,” intrinsically motivating, and less evaluative, in contrast to traditional tests, which are associated with anxiety, performance pressure, and stereotype threat (Gee, 2003; Mayer, 2019; Spencer et al., 1999; Steele & Aronson, 1995). However, most prior work has not isolated evaluative framing as a design variable. Instead, games differ from traditional assessments along many dimensions simultaneously, including feedback, interactivity, narrative, and reward structures. As a result, it remains unclear whether commonly observed benefits of game-based environments are driven by reduced evaluative pressure or by other aspects of their design.

This paper examines a central theoretical question: whether changing the perceived meaning of an activity from a test to a game without changing the underlying task is enough to alter learner experience and inequities. We address this question by highlighting the variable of evaluative framing in the context of assessment. We compare a traditional digital test to a game-based assessment that presents identical content in a narrative, game-like environment, while holding all other aspects constant. This design allows us to work towards addressing a key question in educational game design research: what aspects of “gamefulness” meaningfully change player experience and outcomes, and which do not?

Racial Learning Gaps in Mathematics

Persistent disparities in math achievement across racial groups are well-documented in the United States and reflect existing structural inequities such as unequal access to educational resources, differences in school quality, biased assessment practices, and systemic inequities in opportunity (Lewis & Kuhfeld, 2023; Wronowski et al., 2022). Students from groups that are underrepresented in STEM often face sociocultural and institutional barriers that impact their academic performance and engagement in mathematics learning contexts. In addition to structural inequities, psychological processes such as stereotype threat have been shown to influence performance in evaluative settings. When tasks are framed as diagnostic of ability, students from groups that are negatively stereotyped in a subject area (e.g., women in math) experience increased anxiety and reduced working memory capacity, leading to lower performance (Spencer et al., 1999; Steele & Aronson, 1995).

Importantly, racial disparities in math achievement are not limited to performance. Differences in affective experiences, such as anxiety, self-efficacy, or engagement, can be closely connected with performance gaps and can reinforce those gaps over time (Eccles & Wigfield, 2002; Pekrun, 2006). Prior work has also shown that students' responses to game-based learning environments can differ across demographic groups, particularly when design features interact with students' beliefs and motivations. For example, O'Rourke et al. (2015) found that growth mindset incentives in an educational game produced different effects across student populations (gender, age, and reduced lunch status), highlighting the importance of considering how design choices impact diverse learners. This suggests that addressing inequities in mathematics education requires focusing on improving students' affective experiences as well as learning outcomes.

Given these challenges, digital learning environments, and especially games, have been proposed as a potential way to reduce inequalities by lowering barriers to participation, increasing engagement, and providing alternative forms of assessment (DiSalvo et al., 2014; Reich et al., 2017). However, the extent to which these environments actually mitigate disparities remains an open question.

Prior Work with *Decimal Point*

The current study builds on a substantial body of prior work done with the digital learning game *Decimal Point*, a game that teaches decimal concepts to middle school students (McLaren et al., 2018; McLaren., 2024). The game has students solve problems that target specific misconceptions that are common in decimal learning. The game integrates problem-solving with interactive gameplay elements such as narrative progression and immediate feedback, and has been shown to have better learning outcomes than a comparable non-game tutor. Prior work has also found that students playing the learning game had higher engagement and better affective experiences than students using an equivalent non-game tutor (McLaren, 2024; Ni et al., 2024).

Using the existing data generated from numerous *Decimal Point* studies, Ni et al. (2024) conducted an analysis of racial and ethnic differences in learning and engagement using the game and an equivalent non-game tutoring system. Across the studies analyzed, they found that students from groups well-represented in STEM (White, Asian-American) scored higher on pretest, posttest, and delayed posttest measures than students from underrepresented groups (Black/African-American, Hispanic/Latine, Multiracial, Indigenous/Alaskan Native), even though both well and underrepresented groups had similar learning gains. In addition, students from underrepresented groups reported lower levels of engagement and higher levels of anxiety (across both the game and tutor), suggesting disparities in both performance and experience. However, students playing the game (regardless of race) reported more positive affective experiences.

These findings indicate that while digital learning games can support learning across diverse populations, they do not necessarily eliminate existing performance gaps. Additionally, these results show that differences in affective experience exist even when learning gains are

comparable, which raises concerns about whether educational technologies are designed in ways that are equally responsive to the needs of all students.

Interestingly, prior work with *Decimal Point* has also found that the game consistently helps girls achieve higher learning gains than boys, which ends up reducing or eliminating gender gaps that existed in test performance (Nguyen et al., 2022). This suggests that some aspects of game-based learning can support more equitable outcomes under specific conditions. Therefore, the body of research with *Decimal Point* demonstrates that the effects of game-based learning are complex and context-dependent, and that identifying which design features in learning games contribute to equitable outcomes remains an important research question.

Current Study: Evaluative Framing and Game-based Assessment

Prior work with *Decimal Point* has shown that there are differences in affective experiences among different racial and ethnic groups. However, broader game-based learning research suggests that game-based environments can lead to more positive experiences than traditional instructional formats (Clark et al., 2016; Hussein et al., 2022; Mayer, 2019). Together, these findings raise an important question: what aspects of game-based learning are responsible for these improved experiences, and can they meaningfully address existing disparities?

One possible answer is that games reduce the evaluative salience of traditional instruction or assessment (Gee, 2003; Shute, 2011). Unlike traditional tests that are often framed as measures of ability, games are often framed as fun activities, where failure is an inevitable part of the learning process and performance is part of an ongoing narrative. Given evidence that evaluative contexts can worsen disparities through mechanisms like stereotype threat (Steele & Aronson, 1995; Spencer et al., 1999), reframing assessment as gameplay could reduce anxiety and improve student experiences, especially for students from underrepresented groups.

However, research in game-based learning and gamification often involves multiple design features, such as feedback, interactivity, or reward structures, making it difficult to isolate the mechanisms driving observed effects (Sailer & Homner, 2020). While evaluative framing could also be manipulated through instructions alone (eg. by telling students that they are just solving problems instead of taking a test), such approaches leave the structure of the activity unchanged. In contrast, game-based assessments provide a richer reframing of the task by embedding the activity in a game with different interaction context, narrative goals, and visual design, which shape how students interpret the activity.

The present study builds on prior work by approximating the variable of evaluative framing within a controlled yet ecologically valid game-based context. We keep the instructional content of the assessments constant and minimize potential confounds in the game by removing feedback and opportunities for retrying in the game. Both the game-based and traditional assessment conditions present identical problems in the same order with the same values and answer choices. Therefore, the main difference between conditions is in how the activity is framed and experienced (through narrative, visual presentation, and interaction context), rather

than in the learning opportunities provided. Importantly, we intentionally prioritize experimental comparability over fully realized game design in the game-based assessment, and the findings should be interpreted as speaking specifically to the effects of reframing, rather than to richer forms of game-based assessment that incorporate feedback, agency, and other interactive features.

Importantly, this design does not fully isolate evaluative framing in a strict experimental sense, as the game-based condition necessarily differs in multiple aspects of presentation. Instead, it approximates evaluative framing as it is realized in practice, allowing us to test whether the kinds of framing differences that occur in real game-based systems are sufficient to alter learning and affective outcomes. In doing so, we move beyond asking whether games are broadly beneficial and instead examine whether the specific mechanism of game-based learning of reduced evaluative pressure is sufficient to produce those benefits.

To investigate this question, we conducted a between-subjects study in which students completed assessments that were either presented as a traditional digital test or embedded within a game-based environment. In the game-based condition, students were not explicitly told that they were being tested, whereas in the traditional condition, students were aware that they were taking a test.

This design allows us to address the following research questions:

1. How does assessment format affect performance and learning gains when content is held constant?
2. What are the effects of game-based assessment on students' test-taking experiences?
3. Are there differences in the experiences of students of different demographics when taking a game-based test compared to a traditional test?

Materials and Methods

Study Design

We conducted a between-subjects study to investigate the effects of game-based assessment on students' test-taking experiences and learning. Students were randomly assigned to one of two conditions: a game-based assessment condition or a traditional digital test condition. Both assessment types were delivered digitally and contained identical mathematical content, ensuring comparability across conditions. Students in the game-based test condition were not informed that their performance would be evaluated, in order to minimize evaluative salience. Students in the traditional test condition completed the same tasks but in a conventional digital test format, and were told that they would be taking a test. Researchers were present while students completed the materials in their classrooms to answer logistical questions, but did not provide assistance in answering any problems.

In addition to the tests, students also completed a game intervention that made up a majority of the study materials, as well as surveys to collect demographic information and data on test-taking experiences. The overall structure of the study was as follows: instructions → pretest → survey → intervention → posttest → survey → delayed test → survey. More details on the materials are provided below.

Materials

Students completed three isomorphic assessments to measure learning at three time points: prior to the intervention (pretest), immediately after the intervention (posttest), and approximately one week later (delayed posttest). Depending on the condition, these assessments were administered either as traditional digital tests or as game-based assessments. The game-based assessments were implemented using *Decimal Point*, which functions as a game-based test environment, while the intervention was delivered using a separate learning game, *Game Y*. The traditional tests have been validated in prior work (Nguyen et al., 2022) and cover the same mathematical content as the game-based assessments, including items designed to assess near, medium, and far transfer.

The intervention used in the study was a re-skinned version of the learning game *Decimal Point*, called *Ocean Adventure*. While *Ocean Adventure* was designed to appeal to boys, studies comparing the two games found no difference in learning gains between genders (Stec et al., 2024). Additionally, we expected that if we used *Decimal Point* for both the game-based test and intervention, students in the game-based test condition might lose interest.

Following completion of the assessments, students completed a set of self-report surveys assessing their affective and motivational experiences during learning and assessment. Measures included: a three-item affective engagement scale (e.g., “I felt frustrated or annoyed”) (Ben-Eliyahu et al., 2018); a three-item situational interest scale (e.g., “The game/tutor was exciting”) (Linnenbrink-Garcia et al., 2010); a four-item test anxiety scale (e.g., “If I did poorly on this activity, people would look down on me”) (Spencer et al., 1999); and a five-item self-efficacy scale (e.g., “I could handle this activity”) (Spencer et al. 1999). All items were rated on a Likert-type scale from 1 (strongly disagree) to 5 (strongly agree).

The 9-item TIMSS “Mathematics in school” student questionnaire was used to measure math self-concept and self-efficacy (IEA, 2014), and 5 items were adapted (language was simplified to be middle-school appropriate) from the Domain Identification Measure to assess math identification (Smith & White, 2001). The reliability of these new math survey items was assessed after the study and showed that they demonstrated good internal consistency (Cronbach’s $\alpha = 0.85$).

Game-based Assessment Design

Our game-based assessment, titled *Decimal Point* centers on an amusement park theme. To examine the impact of the delivery method and ensure any differences in student performance

are due only to the platform itself, we designed the game to perfectly mirror the content of a conventional test. Both formats share the exact same questions, presented in an identical linear order, with matching numerical values and answer options. However, while the standard condition relies on a traditional testing layout, the game-based condition transforms these identical problems into interactive tasks within a digital theme park environment.

The player's journey navigates through a map of various themed zones, driven by a storyline where students must clear levels to help a group of stranded aliens find their way home. As an illustration, a traditional multiple-choice problem that asks students to find the highest decimal value is adapted into a medieval-themed challenge (see Figure 1). In this version, the student operates a catapult to aim at and destroy the specific castle marked with the largest number. Throughout the park, a cast of alien characters appears in pop-up windows to guide the user, using visual progress markers and encouraging remarks to maintain player engagement. To preserve the integrity of the experiment, these narrative features are strictly fixed. Every student experiences the exact same character interactions and rewards at designated milestones, regardless of how accurately they answer the questions.

There were several design choices we made to ensure that the comparison between the game-based test and the traditional test was valid; we wanted to make sure that differences in performance would be due to the different format rather than one group having any extra learning opportunities. Because the standard test provides no performance feedback, we removed correctness feedback from the game to avoid students being able to change their answers based on such feedback. However, immediate feedback is an important part of gaming experiences, and removing it brought up concerns that the game-based test would not be sufficiently game-like. To ensure a game-like experience, we incorporated a milestone-based reward mechanic. Students earn digital souvenirs, specifically, new outfits for their alien companions, simply for finishing a themed area, irrespective of the correctness of their answers.

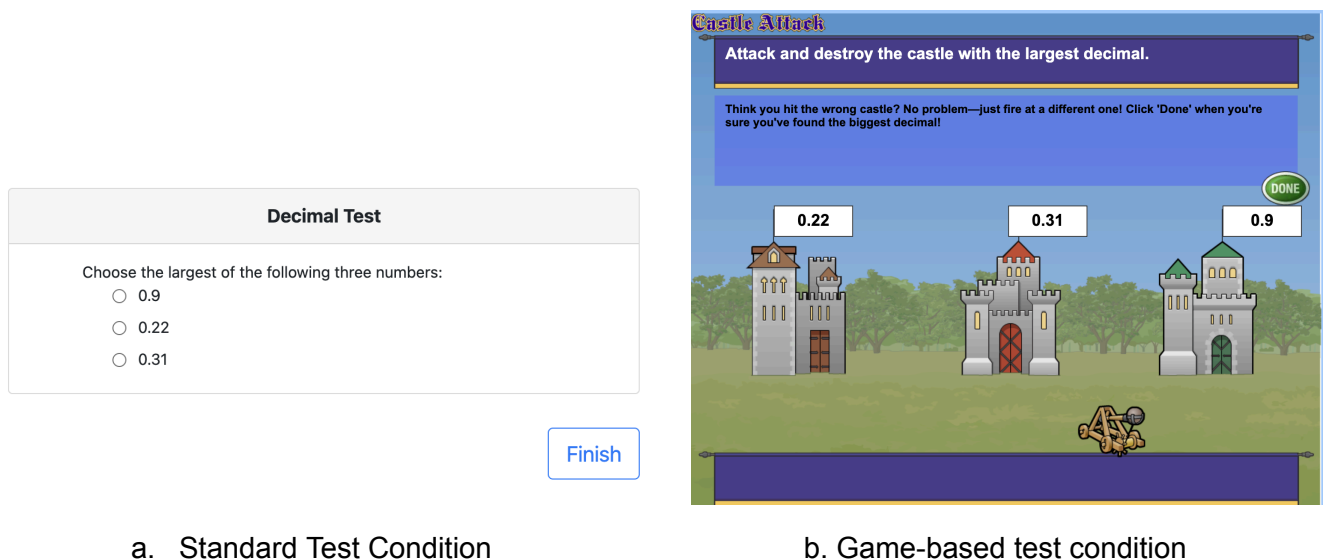


Figure 1. Comparison of standard test interface vs. game-based test interface.

Participants

The study was conducted with 196 5th and 6th-graders from four middle schools in a metropolitan area in the northeastern United States. 72 students were excluded for not finishing all the study materials (mostly due to one school being interrupted by a fire drill and assembly). Of the remaining students, 3 were Asian or Pacific Islander, 25 were Black or African American, 3 were Hispanic or Latino, 81 were White or Caucasian, and 12 were Multiracial or Biracial. Demographic data was provided by teachers, as we have observed that students of this age group often struggle to identify their own race. Additionally, because of the small sample sizes of many of the racial categories, we decided to group students as White and non-White.¹ The final sample included a total of 124 students, with 57 assigned to the game-based test condition (19 non-White, 38 White), and 67 assigned to the regular test condition (24 non-White, 43 White).

Results

Assessment Reliability and Test Item Difficulty

To examine whether the gamified 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 of 0.84 for the gamified assessment (40 items; $N = 165$) and 0.89 for the traditional assessment (40 items; $N = 195$), 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 gamified item difficulty revealed a positive relationship ($\beta = 0.52$, $t(38) = 5.23$, $p < 0.001$, $R^2 = 0.42$), indicating that items that were more difficult in the gamified assessment also tended to be more difficult in the traditional assessment. Together, these findings suggest that both assessments reliably measure the same underlying concepts and that overall differences in performance between the conditions are likely not because of any difference in difficulty or test content.

¹We acknowledge that grouping students as White and non-White is a simplification driven by small sample sizes and should not be interpreted as implying a shared or uniform experience among students categorized as non-White. This grouping may obscure important differences across racial and ethnic groups and should not be taken as positioning whiteness as a default. Future work with larger and more diverse samples is needed to examine these differences in a more nuanced way.

Effects of Condition and Race on Learning

To examine whether students learned from the materials and whether test performance differed by condition and race, we conducted a mixed ANOVA with Time (pretest, posttest, delayed test) as the within-subjects factor and Condition (regular test, game-based test) and Race (White vs. non-White) as between-subjects factors. Mauchly's test of sphericity showed no violations.

There was a significant main effect of Time, $F(2, 240) = 39.70, p < .001, \eta^2 = .036$, indicating that students improved across the three testing occasions. A significant main effect of Condition was also observed, $F(1, 120) = 17.11, p < .001, \eta^2 = .112$, with students in the traditional test condition scoring higher overall than those in the game-based condition. In addition, there was a significant main effect of Race, $F(1, 120) = 13.41, p < .001, \eta^2 = .090$, with White students outperforming non-White students across time points.

These main effects were qualified by a significant Time $\times$ Race interaction, $F(2, 240) = 3.30, p = .039, \eta^2 = .003$, suggesting that the magnitude of performance differences between racial groups varied across time. No other interactions were significant, including the Condition $\times$ Time interaction, $F(2, 240) = 0.43, p = .654, \eta^2 = .0004$, and the Condition $\times$ Race interaction, $F(1, 120) = 0.06, p = .813, \eta^2 = .0004$, indicating that performance differences between assessment conditions were consistent over time and did not differ by race. The three-way interaction between Time, Condition, and Race was also not significant, $F(2, 240) = 0.43, p = .651, \eta^2 = .0004$.

To further examine learning, we analyzed gain scores. A two-way ANOVA on pretest-to-posttest gains revealed no significant main effects of Condition, $F(1, 120) = 1.47, p = .228$, or Race, $F(1, 120) = 0.06, p = .802$, and no significant interaction between Condition and Race, $F(1, 120) = 0.75, p = .387$, indicating comparable immediate learning gains across groups.

A parallel analysis of pretest-to-delayed gains similarly revealed no significant main effect of Condition, $F(1, 120) = 0.44, p = .510$. However, there was a significant main effect of Race, $F(1, 120) = 4.45, p = .037$, with White students demonstrating larger long-term gains than non-White students. The interaction between Condition and Race was not significant, $F(1, 120) = 0.01, p = .946$.

Together, these findings indicate that although students in the traditional test condition scored higher overall, learning gains over time did not differ by assessment format.

Condition	Race	N	Pretest M (SD)	Posttest M (SD)	Delayed Posttest M (SD)
Game-based test	White	38	37.85 (15.47)	42.97 (14.04)	46.96 (15.91)
Game-based test	non-White	19	28.54 (11.58)	34.92 (15.45)	33.50 (15.94)
Traditional test	White	43	49.42 (19.73)	57.69 (17.97)	60.02 (17.14)
Traditional test	non-White	24	39.82 (19.98)	46.23 (17.58)	45.99 (16.52)

Table 1. Descriptive statistics for test scores by condition and race.

Effects of condition and race on affective measures and experiences

To examine whether affective responses differed by assessment format and race, we conducted a series of two-way ANOVAs for each survey measure with Condition (game-based vs. traditional test) and Race (White vs. non-White) as between-subjects factors. Holm corrections were applied within each family of effects (Condition, Race, and Condition $\times$ Race) to control for multiple comparisons.

Across all measures, there were no significant main effects of Condition after correction (all $p_{\text{holm}} = 1.00$), indicating that students in the game-based and traditional test conditions reported similar affective and motivational experiences.

In contrast, several significant main effects of Race were observed. White students reported higher self-concept ($p_{\text{holm}} = .022$), greater affective engagement ($p_{\text{holm}} < .001$), and higher test efficacy ($p_{\text{holm}} = .022$) compared to non-White students. Additionally, White students reported lower test anxiety ($p_{\text{holm}} = .006$). No other race effects were significant after correction.

There were no significant Condition $\times$ Race interaction effects for any of the measures after correction (all $p_{\text{holm}} \geq .36$), indicating that the relationship between race and affective outcomes did not differ by assessment format.

Discussion

In this study, we investigated whether changing the perceived meaning of the activity, without changing the underlying math content, affected students' learning outcomes and test-taking experiences. By manipulating only the presentation and evaluative framing of the assessment, we approximate the variable of evaluative framing to test whether game-based assessment can

address racial learning gaps in math. Surprisingly, considering past research about the motivational and affective benefits of game-based environments, our findings suggest that changing evaluative framing alone may not be enough to meaningfully impact learning or affective experiences.

RQ1: How do learning and test performance differ between game-based and traditional tests?

Across all time points, students in the traditional test condition scored significantly higher than those in the game-based test condition. This difference in performance was consistent across all three tests, but there was no significant difference in learning gains between conditions. Both pre to post and pre to delayed gains showed no significant effect of assessment format, which shows that students learned equally regardless of whether the assessment was presented as a test or a game.

This performance difference cannot be easily attributed to the assessment format itself. Because the pretest differed by condition, baseline differences in prior knowledge cannot be ruled out. Additionally, test item content and difficulty were held constant across conditions, making it unlikely that the observed score differences reflect the game-based test being inherently more difficult.

A more plausible explanation is a difference in test-taking motivation. Prior work on test-taking motivation suggests that students tend to exert more effort and engage more seriously when they perceive an assessment as higher stakes or more conventional in form (Penk & Richer, 2017). In this context, students taking the traditional test were told they were taking a test, while students in the game-based test were told they would be playing a game. Therefore, students in the traditional test condition may have had higher motivation to do well since a test would be viewed as higher stakes than a game. The visual differences between the test and game may also have contributed to a difference in perceived importance.

Importantly, despite these overall score differences, learning gains did not differ by condition. Students in both conditions showed significant improvement over time, and the lack of a condition x time interaction indicates that participating in the game-based assessment did not negatively impact learning. Together, these findings indicate that while the assessment format may have influenced observed performance levels, it did not prevent students from learning the content. From an educational design perspective, these results highlight the important distinction between performance on assessments and knowledge acquisition and suggest that game-based assessments may shift how students approach tasks without changing what they learn.

RQ2: What are the effects of game-based assessment on students' test-taking experiences?

Contrary to our expectations, assessment format did not significantly affect students' affective or motivational experiences during test-taking. Across a wide range of measures, including engagement, interest, self-efficacy, and test anxiety, no differences were observed between the game-based and traditional test conditions after correcting for multiple comparisons. This finding challenges the assumption in game-based learning research that embedding tasks within a game context will improve students' affective experiences. In this study, the game-based assessment changed the framing of the activity students were doing, but this did not lead to measurable changes in how students felt about the task.

One possible explanation for this lack of difference in student experiences is that students still recognized the activity as an evaluative task despite the game-like presentation, especially considering that we intentionally stripped core game mechanics to highlight the potential effects of evaluative framing. Alternatively, it may be that affective responses in assessment contexts are driven more strongly by prior experiences, domain beliefs, or broader classroom norms than by surface-level design features alone.

From a design perspective, our findings suggest that evaluative framing through aesthetic or narrative elements alone is not enough to meaningfully influence student experience, and that more substantial changes to interaction structure (eg, through more interactive game mechanics) may be required.

RQ3: Do students of different demographics experience game-based assessment differently?

While the assessment format did not affect learning or affective outcomes, consistent differences were observed between racial groups. White students performed better than non-White students across all time points and demonstrated greater long-term gains. White students also reported higher self-concept, affective engagement, and test efficacy, and lower test anxiety than non-White students.

Importantly, these differences were not moderated by assessment format. In other words, presenting the assessment to students as a game versus as a test did not reduce or exacerbate disparities in either performance or experience. This finding indicates that, in the context of race and middle school mathematics, simply reframing assessment as a game is insufficient to address inequities that may stem from broader structural, educational, or psychological factors.

Classic stereotype threat work has shown that the framing of a task can influence performance for groups that are negatively stereotyped, especially when a task is presented as diagnostic of

ability. In Steele and Aronson's (1995) work, the same math test given to African-American students produced different outcomes depending on whether it was framed as diagnostic, and Spencer et al. (1999) showed similar effects for women in mathematics. At the same time, work on low-stakes testing shows that students' perceived stakes and motivation can meaningfully affect measured performance (Penk & Richter, 2017; Wise & DeMars, 2005).

Implications for the Design of Game-Based Assessment

Our findings point to important implications for the design of game-based assessment and learning game interventions. Games are often assumed to reduce evaluative pressure by reframing tasks as playful activities. For example, work on stealth assessment argues that embedding assessment within gameplay can make evaluation less salient, allowing students to demonstrate their knowledge in low-stakes, engaging environments (Shute, 2011). Similarly, broader research on gamification and game-based learning suggests that gameful elements such as narrative, rewards, and interactivity can improve motivation and engagement (Sailer & Homner, 2020).

However, our results show that evaluative framing alone may not be enough to change student experiences, even when learning outcomes do change. This contradicts assumptions that playful framing is inherently motivating, less threatening, or more equitable.

One implication is that "gamefulness" should not be treated as a surface-level property defined by narrative, visual design, or thematic context alone. While these elements contribute to how an activity is presented, they may not be enough to fundamentally change how students interpret the activity, particularly in assessment settings where students may already hold strong expectations about evaluation. Prior work on motivation and interest shows that sustained engagement depends not only on situational features, but also on deeper perceptions of value, competence, and autonomy (Hidi & Renninger, 2006). Therefore, evaluative meaning is shaped by how an activity functions in addition to how it appears visually to students.

Our findings, therefore, suggest that designing effective game-based assessments requires attention to deeper structural features of interaction, rather than relying solely on aesthetic or narrative framing. In our study, many of the features that typically distinguish games from tests (eg., immediate feedback, opportunities for retry, flexible progression) were intentionally removed to ensure comparability across conditions. While this allowed us to better investigate the variable of evaluative framing, it also highlights that these features may play a critical role in shaping student experience.

One concrete direction for future design work is to test specific interaction mechanics rather than surface theming. For example, Gee (2003) describes how well-designed games often require players to demonstrate competence before advancing, a mastery-gated progression structure that embeds formative feedback and retry opportunities directly into gameplay rather than into narrative or visual presentation. Testing whether mechanics such as mastery-gated progression, rather than narrative or visual framing alone, can reduce evaluative pressure and support more equitable outcomes is an important next step for game-based assessment design.

Future work should also consider whether students' prior familiarity and comfort with digital games, or game literacy, shapes how game-based assessments are experienced, since unequal access to games outside of school could introduce an additional source of disparity independent of assessment content.

Conclusion

This study examined whether reframing assessment as a game, while holding content constant, affects students' learning and test-taking experiences. Across conditions, students demonstrated similar learning gains, indicating that game-based assessment did not hinder learning. However, overall performance was lower in the game-based condition, suggesting that assessment format may influence how knowledge is expressed, potentially through differences in perceived stakes or test-taking motivation. Contrary to common assumptions about the benefits of game-based environments, we found no evidence that game-based assessment improved students' affective or motivational experiences after correcting for multiple comparisons.

Importantly, persistent differences across racial groups were observed in both performance and affective measures, and these disparities were not reduced by presenting the assessment as a game. Together, these findings suggest that reframing assessment through game-based presentation alone is insufficient to meaningfully change student experience or address inequities.

From a design perspective, this work highlights the need to move beyond surface-level notions of gamefulness and toward a more precise understanding of how specific design features influence learning and experience. Rather than assuming that games are inherently more engaging or equitable, future research should identify which aspects of game-based systems are necessary to create meaningful and inclusive assessment experiences.

Several limitations of this study should be considered when interpreting the results. First, although the study aimed to manipulate evaluative framing, we did not include a direct measure of whether students perceived the game-based condition as less evaluative. This was a deliberate design choice, as explicitly asking students about evaluation could have revealed the nature of the manipulation. However, without a manipulation check, we cannot confirm whether the intended differences in evaluative framing were experienced by students as expected. Second, in order to maintain comparability between conditions, we removed key game mechanics such as correctness feedback and opportunities for retry from the game-based assessment. While this allowed us to better control for differences in learning opportunities, it also reduced the extent to which the game-based condition reflected a fully realized game experience. As a result, the study may underestimate the potential impact of game-based assessment when more interactive, gameful features are included. Third, due to small sample sizes across racial categories, we grouped students as White and non-White. This simplification does not capture the diversity of experiences across different racial and ethnic groups and may obscure important differences that deserve focused research. Additionally, we did not measure

students' prior familiarity or comfort with digital games (i.e., game literacy), which may itself vary systematically across demographic groups and could shape how the game-based assessment was experienced independent of its content. Finally, this study reflects a tradeoff between experimental control and ecological validity. While we controlled for content and key instructional features, the game-based and traditional conditions necessarily differed in multiple aspects of presentation. As such, our findings should be interpreted as reflecting the effects of evaluative framing within a realistic design context, rather than as a fully isolated effect.

Future work should explore more comprehensive manipulations of evaluative context, including designs that vary feedback, agency, and perceived stakes, as well as studies that directly measure students' interpretations of assessment. In addition, further research is needed to examine how game-based assessment interacts with broader social and cultural factors that shape students' experiences, particularly in relation to equity and inclusion.

Acknowledgements

Generative AI (ChatGPT, Claude) was used during preparation of this manuscript to assist with language editing and copyediting, such as improving sentence clarity, correcting grammar, and tightening wording.

References

- 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.
- DiSalvo, B., Guzdial, M., Bruckman, A., & McKlin, T. (2014). Saving face while geeking out: Video game testing as a justification for learning computer science. *Journal of the Learning Sciences*, 23(3), 272-315.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53(1), 109–132. <https://doi.org/10.1146/annurev.psych.53.100901.135153>
- Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in entertainment (CIE)*, 1(1), 20-20.
- Hussein, M. H., Ow, S. H., Elaish, M. M., & Jensen, E. O. (2022). Digital game-based learning in

- K-12 mathematics education: A systematic literature review. *Education and Information Technologies*, 27(2), 2859-2891.
- Lewis, K., & Kuhfeld, M. (2023). Education's Long COVID: 2022-23 Achievement Data Reveal Stalled Progress toward Pandemic Recovery. Center for School and Student Progress at NWEA.
- 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.
- Mayer, R. E. (2019). Computer Games in Education. *Annual Review of Psychology*, 70(1), 531–549. <https://doi.org/10.1146/annurev-psych-010418-102744>
- McLaren, B.M. (2024). Decimal Point: A decade of learning science findings with a digital learning game. In: Ilic, P., Casebourne, I., Wegerif, R. (eds) *Artificial Intelligence in Education: The Intersection of Technology and Pedagogy*. Intelligent Systems Reference Library, vol 261. Springer, Cham. https://doi.org/10.1007/978-3-031-71232-6_9.
- McLaren, B. M., Adams, D. M., Mayer, R. E., & Forlizzi, J. (2018). A computer-based game that promotes mathematics learning more than a conventional approach. In *Gamification in Education: Breakthroughs in Research and Practice* (pp. 415-437). IGI Global Scientific Publishing.
- 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. *International Journal of Game-Based Learning (IJGBL)* 12(1), 1–29 (2022)
- Ni, X., Nguyen, H.A., Else-Quest, N., Pagano, A., & McLaren, B.M. (2024). Investigating racial and ethnic differences in learning with a digital game and tutor for decimal numbers. In: *Proceedings of the Nineteenth European Conference on Technology Enhanced Learning (ECTEL 2024)*. Krems, Austria, September 16-20, 2024.
- Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries,

- and implications for educational research and practice. *Educational psychology review*, 18(4), 315-341.
- Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, 29(1), 55–79. <https://doi.org/10.1007/s11092-016-9248-7>
- Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, 1(1), 91-116.
- Reich, J., Ito, M., & Team, M. S. (2017). From good intentions to real outcomes. *Digital Media and Learning Research Hub*.
- Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational psychology review*, 32(1), 77-112.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning.
- Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement*, 61(6), 1040–1057. <https://doi.org/10.1177/00131640121971635>
- Stec, H., Richey, J. E., Nguyen, H., Else-Quest, N., Hammer, J., Baker, R. S., ... & McLaren, B. (2024). Using a multi-dimensional model of gender to assess learning with different game-based learning narratives.
- Steele, C., & Aronson, J. (1995). Stereotype Threat and the Intellectual Test Performance of African Americans. *Journal of Personality and Social Psychology*, 797–811. <https://doi.org/10.1037/0022-3514.69.5.797>
- 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. <https://doi.org/10.1006/jesp.1998.1373>
- International Association for the Evaluation of Educational Achievement (IEA). (2014). TIMSS and PIRLS international study center: Student questionnaire (Grade 4). TIMSS & PIRLS

International Study Center, Boston College.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, 10(1), 1–17.

https://doi.org/10.1207/s15326977ea1001_1

Wronowski, M. L., Thornton, M., Razavi-Maleki, B., Witcher, A. W., & Duarte, B. J. (2022).

Beyond tracking: The relationship of opportunity to learn and diminished math outcomes for US high school students. *Teachers College Record*, 124(6), 196-224.