

The Impacts of Three Educational Technologies on Algebraic Understanding in the Context of COVID-19

Lauren E. Decker-Woodrow

Westat Inc

Craig A. Mason

University of Maine

Ji-Eun Lee

Worcester Polytechnic Institute

Jenny Yun-Chen Chan

The Education University of Hong Kong

Adam Sales

Worcester Polytechnic Institute

Allison Liu

Worcester Polytechnic Institute

Shihfen Tu

University of Maine

The current study investigated the effectiveness of three distinct educational technologies—two game-based applications (From Here to There and DragonBox 12+) and two modes of online problem sets in ASSISTments (an Immediate Feedback condition and an Active Control condition with no immediate feedback) on Grade 7 students' algebraic knowledge. More than 3,600 Grade 7 students across nine in-person and one virtual schools within the same district were randomly assigned to one of the four conditions. Students received nine 30-minute intervention sessions from September 2020 to March 2021. Hierarchical linear modeling analyses of the final analytic sample ($N = 1,850$) showed significantly higher posttest scores for students who used From Here to There and DragonBox 12+ compared to the Active Control condition. No significant difference was found for the Immediate Feedback condition. The findings have implications for understanding how game-based applications can affect algebraic understanding, even within pandemic pressures on learning.

Keywords: *algebra, classroom research, educational games, educational technology, experimental research, game-based learning, hierarchical linear modeling, mathematics education, middle schools*

ALGEBRA is a critical foundation for learning advanced mathematics (National Mathematics Advisory Panel, 2008). However, many middle and high school students struggle to understand even basic algebraic concepts (Kena et al., 2015; Kieran, 2006). For example, students struggle to achieve a robust understanding of how to create, transform, and

interpret algebraic expressions. This current study, therefore, tested the impacts of three educational technology interventions on algebraic understanding among students in Grade 7 across four conditions: (a) *From Here to There* (FH2T), (b) *DragonBox 12+* (DragonBox), (c) Immediate Feedback, and (d) Active Control. The FH2T and DragonBox



conditions represented use of game-based applications. Immediate Feedback entailed problem sets by using an online homework system, ASSISTments. For the purposes of this study, the Active Control condition mimicked traditional homework assignments while still using technology. Although this study independently investigated each of the three treatment conditions to the Active Control condition, it was hypothesized that FH2T, an interactive game developed based on theories of perceptual learning and embodied cognition, might improve students' algebraic understanding through aligning their attention to, actions on, and perception of algebraic notations with high-level mathematical concepts to a greater extent than would DragonBox or Immediate Feedback.

Algebra Learning and Educational Technologies

Algebraic principles, such as notation and transformations, can be challenging for students to learn. Many middle and high school students struggle to acquire basic algebraic principles, such as which transformations are mathematically valid (Marquis, 1988; National Math Advisory Panel, 2008) and how to convert between different algebraic representations (Koedinger & Nathan, 2004). These struggles have significant implications for students' further mathematics achievement because advanced concepts are often presented in algebraic notation.

To improve mathematical understanding and performance among middle school students, researchers and developers have designed educational technology tools to support learning in different ways. Some tools, such as FH2T and DragonBox, are designed to engage students in game-based learning, allowing students to interact with algebraic notations and solve puzzlelike problems in a playful environment. Such game-based approaches have been found to increase students' engagement and motivation as well as problem-solving and learning (Connolly et al., 2012; Foster, 2008; Gee, 2003; Ke, 2008; Samur & Evans, 2012). Other tools, such as ASSISTments, are designed to provide timely support and feedback on homework, and thus the problems resemble those in traditional mathematics textbooks. Below is an overview of the four conditions, the theories behind the design of the tools, and prior evidence of their efficacy in improving mathematical learning.

From Here to There! (FH2T). FH2T (<https://graspablemath.com/projects/fh2t>) is a dynamic research-based game that implements theories of (a) perceptual learning and (b) embodied cognition to address cognitive and affective factors that lead to low proficiency in mathematics. Perceptual learning theory suggests that reasoning and learning about mathematics are inherently perceptual (Goldstone et al., 2017; Jacob & Hochstein, 2008; Kellman et al., 2010; Kirshner & Awtry, 2004; Patsenko & Altmann, 2010), and the

visual presentation of notation affects how students reason about, process, understand, and learn mathematics (Braithwaite et al., 2016; Harrison et al., 2020; Landy & Goldstone, 2010). For example, perceptual features, such as spacing and color of algebraic notations, can direct students' attention to relevant information (e.g., highlighting the equal sign with a different font color in $4 + 7 = 13 - \underline{\quad}$ to support reasoning of equivalence; Alibali et al., 2018), and, over time, might help students develop an automatic routine for algebraic reasoning (e.g., balancing two sides of the equation when seeing an equal sign). Perceptual features also affect students' actions, which reflect and further influence their learning. Embodied cognition theories contend that students' physical experiences in the world influence their cognitive processes, including thinking and reasoning in mathematics (Abrahamson et al., 2020; Alibali & Nathan, 2012; Foglia & Wilson, 2013; Nathan et al., 2014; Shapiro, 2010; Wilson, 2002). Given that the two theories together suggest that learning is based in perception and action (Alibali & Nathan, 2012), environments within FH2T influence how students perceive and interact with instructional materials, which, in turn, is meant to inform their cognitive processes and problem-solving.

In FH2T, algebraic notations are turned into interactive objects that enforce mathematical rules through their physical movements. Students can dynamically manipulate and transform expressions by using various gestures, such as dragging and tapping, to perform operations on the screen. The system provides a fluid visualization that allows students to see how their actions transform algebraic expressions and equations. In each problem in FH2T, students are presented with two expressions: a starting expression, which is an active and transformable expression, and a target goal state, which is mathematically equivalent but perceptually different to the starting expression. Students transform the starting expression (e.g., " $9b + 27c$ ") into the target goal state (e.g., " $(3b + 9c) \cdot 3$ ") by using algebraically permissible actions and learned gestures. For example, students can drag 9 on top of 27 to factor 27 into $9 \cdot 3$, and the system transforms the starting expression into " $9b + 9 \cdot 3c$." As students encounter, discover, and practice mathematical principles in action, fluid visualizations may help direct students' attention toward structural patterns in algebraic notation and develop their perceptual-motor routines of algebraic equation-solving (Nathan et al., 2016; Nathan & Walkington, 2017). (See online supplementary material, Appendix A, for more information.)

The effectiveness of FH2T has been examined across several studies over the past decade. For instance, a preliminary study with 130 middle school students showed that, after four 30-minute intervention sessions, students in the fluid visualizations condition, which is an early version of FH2T, improved on their procedural fluency, whereas students in the manual calculations condition did not (Ottmar

et al., 2015). Similarly, a recent randomized controlled trial (RCT) with 475 middle school students showed that, after four 30-minute intervention sessions, students in the FH2T condition scored higher on mathematical equivalence compared to their peers in an online problem set condition (Chan et al., 2022). A study with an elementary version of FH2T further showed that, among 185 Grade 2 students, completing more problems in FH2T was associated with higher posttest scores (testing decomposition, operational strategies, and basic notation), and this effect was significant above and beyond students' prior knowledge, as measured by the pretest (Hulse et al., 2019). Together, these findings suggest that FH2T may be effective at improving students' algebraic understanding and mathematics performance, warranting a large-scale RCT to evaluate its effectiveness compared to other educational technologies.

DragonBox 12+ (DragonBox). DragonBox (<https://dragonbox.com/products/algebra-12>) is an educational game that introduces advanced algebraic concepts to students ages 12–17. For each problem, students are asked to isolate a box containing a dragon—equivalent to solving an equation for x . The popularity of DragonBox is reflected in the public realm, where it is lauded as one of the best educational games for teaching algebra (Kahoot!, 2019). Its design has also been praised for using several research-backed design principles, including discovery-based learning, embedded gestures, multiple representations, immediate feedback, and adaptive difficulty (Cayton-Hodges et al., 2015; Torres et al., 2016). One of DragonBox's major design principles is that students should not perceive DragonBox as a *math* game. Thus, at the start of the game, there are no numbers or variables; instead, students move cards with pictures of monsters, following the rules of algebra. Throughout the game, monster cards are gradually replaced by algebraic symbols; however, cards are never explicitly connected to mathematical properties.

Despite the popularity of DragonBox, research findings on its efficacy are mixed. In terms of supportive evidence, the DragonBox developers and the University of Washington Center for Game Science conducted an Algebra Challenge with K–12 students from 70 schools across 15 districts. They found that after 1.5 hours of playing a combined version of DragonBox 5+ and 12+, 93% of students were able to successfully answer three algebra problems (e.g., $d \cdot x + m = 8$) in a row (Liu et al., 2015). Similarly, Grade 8 students have shown significant gains in algebra problem-solving performance after using DragonBox for 8 hours across 4 weeks (Dolonen & Kluge, 2015), and its positive impacts may extend to students' confidence and attitudes toward mathematics (Siew et al., 2016). However, Long and Aleven (2014, 2017) find that students in Grades 7–8 who played DragonBox for 3.5 hours across five sessions showed no improvements in problem-solving performance

or confidence, even on problems included in the game. Further, despite the students' higher enjoyment with DragonBox, their learning gains were significantly lower than those of students using an intelligent tutor system. Dolonen and Kluge (2015) also find that Grade 8 students using DragonBox showed higher engagement but lower learning gains compared to students who practiced problems similar to those found in standard algebra textbooks.

Contrasting DragonBox and FH2T. It is appropriate to compare DragonBox to FH2T because both use a game-based interface and are grounded in intuitive perceptual-motor routines, but they also vary in important ways. First, whereas DragonBox progressively transitions from pictures of monsters to abstract algebraic notation as the game proceeds, FH2T aims to promote fluency in algebraic notation by grounding perceptual-motor routines directly in notation. Second, all of DragonBox's problems involve the same goal (i.e., isolate the dragon to solve for x); in contrast, each problem in FH2T has a unique goal state, encouraging flexibility in notational transformations. Third, DragonBox's operations are introduced as in-game rules independent from mathematics (e.g., positive- and negative-color versions of the same picture can be combined to cancel them), whereas FH2T directly connects actions and operations to mathematical principles. These similarities and differences allowed the study team to begin exploring relative advantages and disadvantages of hiding mathematical materials, having flexible goals, and explicitly grounding actions in mathematical rationales.

Problem Sets With Immediate Feedback and Hints in ASSISTments (Immediate Feedback). ASSISTments (<https://new.assistments.org/>; Heffernan & Heffernan, 2014) is an online homework system that provides feedback to students as they solve traditional textbook problems. The problem sets in ASSISTments are adapted from open-source curricula, thus resembling problems students would encounter in their textbooks and homework assignments. In ASSISTments, students are presented with problems one at a time on their screen; they can request hints if they need additional support on problem-solving and receive immediate correctness feedback on their responses. Teachers receive class reports on problem sets and students' performance so they can identify struggling students and challenging problems, and then they use these data to inform instruction in their classroom. (See online supplementary material, Appendix A, for more information.)

The design of ASSISTments is informed by research on formative assessment and timely feedback. Teachers use formative assessments to gather data from students and guide their instruction to meet students' needs (Black & Wiliam, 1998; Boston, 2002). When teachers adjust instruction based on formative assessments, students show significant improvement

on their achievement (Bergan et al., 1991; Speece et al., 2003). Further, timely feedback and support during learning are beneficial to students (Butler & Woodward, 2018; Shute, 2008). Studies suggest that receiving feedback *immediately* after giving responses or completing problem sets might be effective for improving students' procedural and conceptual knowledge (Azevedo & Bernard, 1995; Corbett & Anderson, 2001; Dihoff et al., 2003; Phye & Andre, 1989).

ASSISTments has had positive impacts on student learning. A preliminary study with 28 Grade 5 students found that they learned significantly more when provided with immediate feedback and on-demand hints in ASSISTments compared to completing traditional paper-and-pencil homework (Mendicino et al., 2009). Similarly, a study with 63 Grade 7 students conducted within ASSISTments showed that students who received immediate correctness feedback and opportunities to reattempt problems outperformed students who completed problems without feedback or reattempts (Kelly et al., 2013). A large-scale RCT with 2,850 students in Grade 7 further showed that combining ASSISTments with teacher training on adaptive teaching significantly improved students' end-of-year performance on standardized mathematics scores compared to those of a business-as-usual control group (Roschelle et al., 2016). Together, these findings suggest that offering online problem sets with hints and immediate feedback plus teacher training is effective at improving students' mathematics performance. In the current study, problem sets with immediate hints and feedback in ASSISTments were used as a high-quality, nongamified intervention treatment. Unlike the work of Roschelle et al. (2016), this study did not provide teacher training on adaptive instruction because FH2T and DragonBox only involve student components and do not include teacher training. Therefore, the current study provides insights into independent effects of the student components without teacher components in ASSISTments.

Active Control Problem Sets With Post-Assignment Feedback (Active Control). The fourth condition was a version of ASSISTments that provides post-assignment, rather than immediate, hints and feedback. In this condition, problem sets were administered in "test mode," so students did not receive any feedback or hints during problem-solving. Instead, students received a report with correctness feedback at the end of the problem set, and they could review their responses, revisit problems, and request hints. This condition served as the active control in the study, as it mimicked traditional homework assignments and allowed all students involved in the study to participate via technology.

The Current Study

The goal of the current study was to independently examine the efficacy of three widely used treatment conditions

(i.e., FH2T, DragonBox, and ASSISTments with immediate feedback) on students' algebraic understanding compared to the Active Control. The study was conducted between September 2020 and April 2021, which was during the peak of the COVID-19 pandemic in the United States. Given the restrictions of physical distancing, the school district offered students and their families a choice of classroom format (100% in person or 100% asynchronous virtual academy) for the 2020–2021 school year prior to the start of the fall semester. Random assignment to study condition occurred across formats, so the classroom format was not aligned to any one study condition. Although their choice was intended to be for the full school year, families were allowed to change their selection throughout the year. Regardless of students' classroom format, all study sessions (i.e., assessments and interventions) were administered online, and students worked individually at their own pace, using a device.

Methods

Participants

A total of 52 Grade 7 mathematics teachers and their students from 11 middle schools (10 in-person schools and one virtual academy) were recruited from a large, suburban district in the southeastern United States in the summer of 2020. Together, teachers taught 190 mathematics classes and 4,092 students. Prior to random assignment, one school declined to participate. This school included four teachers and 377 students. Students enrolled in resource settings were not included in this study. This resulted in 10 schools (nine in person and one virtual), 37 teachers, 156 classes, and 3,612 students. The study sample was 49.1% White, 14.1% Hispanic, and 27.9% Asian. Additionally, 42.5% began the school year in the Virtual Academy. Student-level socioeconomic data (e.g., free or reduced-price lunch status) were not available. (All baseline and posttest sample sizes are presented in Figure 1.)

Design

This study randomly assigned students to study conditions within classrooms, ensuring that study conditions were equivalent with respect to teacher characteristics and classroom curricula. This approach was possible because teachers were able to use all four technology-based interventions within the classroom. Students were ranked within classrooms based on their prior state mathematics assessment scores, blocked into sets of five (i.e., quintets), and then randomly assigned into either the FH2T (40%), DragonBox (20%), Immediate Feedback (20%), or Active Control (20%) condition. Because a primary goal of the RCT was designed to test the efficacy of FH2T compared to other educational technologies, a larger proportion of students was assigned to FH2T. When classroom size did not allow for a quintet to be

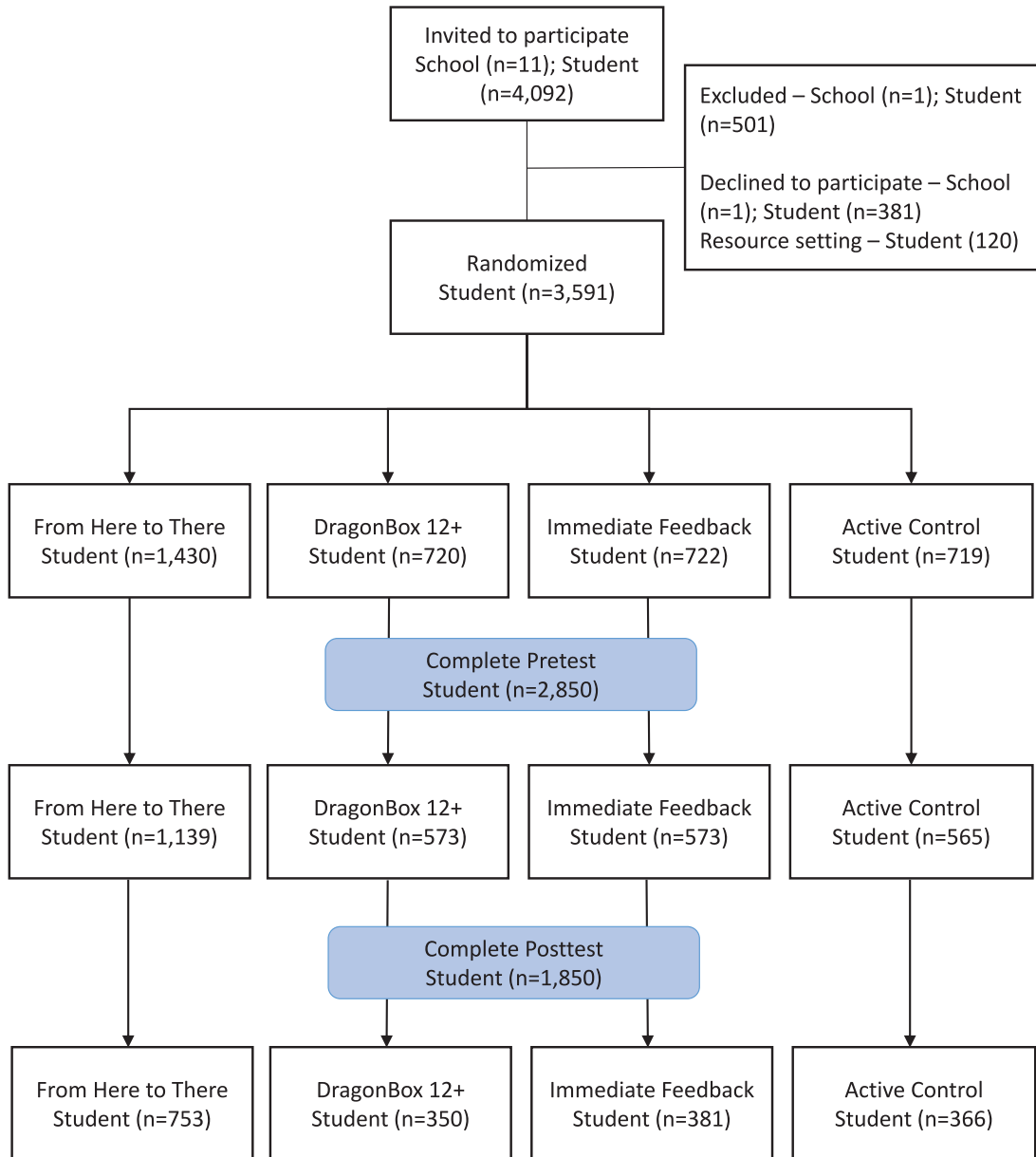


FIGURE 1. Consort figure representing attrition.

formed (e.g., 18 students, resulting in three excess students not in a complete quintet), remaining students in each classroom were placed in quintets drawn from all excess Grade 7 students within each school. Cross-school quintets were formed, as needed, to complete overall random assignment.

Attrition Analysis

One school dropped out following pretest assessment, resulting in a final pool of nine schools (eight in person and one virtual), 34 teachers, 143 classes, and 3,271 students

participating at the start of the interventions. From this pool of 3,271 students, 1,850 had pretest and posttest assessments and constituted the analytic sample for this study. These students were enrolled in 127 classes across 34 teachers. Given the larger context of the COVID-19 pandemic, the overall attrition rate was 48.8%.

Our approach to causal inference under attrition, following the guidelines of What Works Clearinghouse (WWC; 2020) and others, was to estimate treatment effects for the subset of students in our analysis sample with complete pretest and posttest data. Under this approach, missing data

TABLE 1

Student Demographic Information by Condition and Pretest Scores (N = 1,850)

	Full Sample N = 1,850 (100%)	FH2T n = 753 (40.7%)	DragonBox n = 350 (18.9%)	Immediate Feedback n = 381 (20.6%)	Active Control n = 366 (19.8%)
Gender					
Male	932 (50.4%)	393 (52.2%)	171 (48.9%)	183 (48.0%)	185 (50.5%)
Female	918 (49.6%)	360 (47.8%)	179 (51.1%)	198 (52.0%)	181 (49.5%)
Race/Ethnicity					
White	969 (52.4%)	392 (52.1%)	181 (51.7%)	210 (55.1%)	186 (50.8%)
Asian	458 (24.8%)	180 (23.9%)	82 (23.4%)	97 (25.5%)	99 (27.0%)
Hispanic	269 (14.5%)	118 (15.7%)	56 (16.0%)	44 (11.5%)	51 (13.9%)
African American	80 (4.3%)	35 (4.6%)	16 (4.6%)	14 (3.7%)	15 (4.1%)
Native American	10 (0.5%)	2 (0.3%)	3 (0.9%)	1 (0.3%)	4 (1.1%)
Pacific Islander	2 (0.1%)	0 (0.0%)	1 (0.3%)	0 (0.0%)	1 (0.3%)
Other race	62 (3.4%)	26 (3.5%)	11 (3.1%)	15 (3.9%)	10 (2.7%)
Accelerated math class	446 (24.1%)	178 (23.6%)	81 (23.1%)	94 (24.7%)	93 (25.4%)
EIP	134 (7.2%)	51 (6.8%)	29 (8.3%)	28 (7.3%)	26 (7.1%)
Gifted	309 (16.7%)	129 (17.1%)	45 (12.9%)	70 (18.4%)	65 (17.8%)
IEP	168 (9.1%)	58 (7.7%)	39 (11.1%)	35 (9.2%)	36 (9.8%)
Virtual	605 (32.7%)	249 (33.1%)	111 (31.7%)	125 (32.8%)	120 (32.8%)
Completed assignments (SD)	6.49 (2.95)	6.60 (2.86)	5.53 (3.38)	6.81 (2.68)	6.86 (2.77)
Average pretest (SD)	4.84 (2.67)	4.80 (2.70)	4.96 (2.60)	4.94 (2.70)	4.70 (2.67)
Average posttest (SD)	4.57 (2.90)	4.59 (2.96)	4.81 (2.88)	4.58 (2.90)	4.29 (2.79)

Note. EIP = Early Intervention Program; IEP = Individualized Education Program; SD = standard deviation.

methods, such as full-information maximum likelihood or multiple imputation, are not necessary because no claims are made about larger populations that include potential attritors and non-attritors.

The threat of attrition bias was assessed in two ways—first, by comparing attrition rates across conditions, and second, by checking whether non-attriting students assigned to different conditions were equivalent at baseline. As to the first strategy, Figure 1 shows attrition for all key study contrasts. Importantly, no contrast exceeds attrition percentage thresholds established by WWC. In the vernacular of WWC, this study can be characterized as a low-attrition RCT, and all key contrasts are eligible to “Meet WWC Group Design Standards Without Reservations” (WWC, 2020, p. 5). This characterization assumes that WWC would use the “Optimistic” attrition threshold, which entails assuming that sample loss is exogenous (i.e., unrelated) to intervention conditions in the study.

Anecdotal evidence and exploratory data analysis (online supplementary material, Appendix B) suggest that the relatively high attrition rate for DragonBox was largely driven by students in the Virtual Academy, due to difficulty installing the extra required software. In any event, differential attrition was not statistically significant ($\chi^2(3, N = 1,849) = 3.978, p = .264$), and differential as well as overall attrition rates for active control comparisons fell within tolerable threats of bias under optimistic assumptions (WWC, 2020).

Out of an abundance of caution, baseline equivalence of the analysis sample was examined by using student demographics and pretest scores reported in Table 1. Differences in covariate means for Active Control comparisons ranged in magnitude from below 0.001 to roughly 0.22 standard deviations. Therefore, the study met the WWC (2020) baseline equivalence requirement after statistical adjustments for covariates were made in Models 2–4, below. Baseline equivalence for Active Control comparisons were tested by using the methods of Hansen and Bowers (2008); differences were significant compared to DragonBox ($\chi^2(8) = 15.6, p = .049$), but not for the other two comparisons (Immediate Feedback: $\chi^2(8) = 6.7, p = .569$ and FH2T $\chi^2(8) = 11.6, p = .17$). There was evidence of higher pretest scores in each of the experimental conditions compared to the Active Control (Immediate Feedback: $Z = 2.13, p = 0.033$; DragonBox: $Z = 3.05, p = 0.002$; FH2T: $Z = 2.13, p = 0.033$), suggesting that the association between pretest and attrition was weaker in the Active Control and increasing the importance of pretest adjustments in Models 2–4, below.

Procedure

This research was approved by the Institutional Review Board at a university in the northeastern United States. This research involved typical educational practices and thus did not require parental consent. Instead, parents were informed

about the research and the data collected from the educational technologies through a letter and could opt their child out of this study.

Students received nine 30-minute intervention sessions across the school year, with a 2-week window to complete each session. Four intervention sessions were administered in the fall semester and five sessions in the spring semester. Before and after the intervention, students received a 40- to 45-minute assessment on algebraic knowledge, mathematics anxiety, and self-efficacy. All interventions and assessments were administered online, and students worked individually at their own pace, using a device. For students receiving instruction in person, teachers dedicated instructional periods for the study assignments in mathematics classrooms. For students receiving virtual instruction, teachers included the study assignments as a part of students' learning activities.

The mathematical content (arithmetic and algebraic equation solving) was aligned between the four conditions, and all students solved algebraic problems during their intervention sessions by using their assigned technology (including the Active Control condition). Within each condition, a countdown timer was embedded in the system to help ensure that students used each technology for a similar amount of time. Students could pause and resume the timer, so they could stop for breaks and continue when they were ready. A timer was not embedded in the assessments (i.e., pre-and posttest), so students could take as long as needed to complete the assessments.

In addition to using ASSISTments as an intervention technology (i.e., Immediate Feedback and Active Control conditions), it was the platform for the RCT in which assessments were administered, assigned intervention conditions were maintained within students over time, and fidelity data were recorded for *all* conditions (see details in Chan et al., 2022). All students, regardless of their intervention condition, logged in to their ASSISTments account and opened the same assignment link at the beginning of each study session. The assignment link then directed students to their technology intervention, and students spent 30 minutes with their assigned technology.

Measures

Algebraic Knowledge Assessment. The algebraic knowledge assessment consisted of 10 multiple-choice items from a previously validated measure of algebraic understanding (Star et al., 2015; Cronbach's $\alpha = .89$; see the 10 items on osf.io/bafdr). Within the 10 items, four focused on conceptual understanding of algebraic equation-solving (e.g., the meaning of an equal sign), three focused on procedural skills of equation-solving (e.g., solving for a variable), and three focused on flexibility of equation-solving strategies (e.g., evaluating different equation-solving strategies). These 10

items together assessed a range of students' knowledge in algebraic equation-solving, which is the ultimate goal of each of the education technologies tested. The study team created the posttest assessment by substituting numbers of similar magnitudes in the pretest questions and response options. Each item in the pretest and posttest was scored as correct (1) or incorrect (0), and the reliability of these items were KR-20 (Kuder-Richardson Formula 20; a reliability measure for binary variables) = .74 at pretest and .79 at posttest. The total score out of 10 on the posttest scores was included as the outcome, and the pretest score was included as a covariate.

Assessments were administered within the ASSISTments platform, in which students saw questions one at a time and selected a response option by using a mouse. No feedback was given to students about the correctness of their response. The pretest assessment was administered in September 2020, approximately 1 week prior to the intervention sessions. The posttest assessment was administered between the end of March and the beginning of April 2021, approximately 2 weeks following completion of the intervention.

Intervention Condition. Intervention condition was a four-level dummy-coded categorical variable indicating participation in FH2T, DragonBox, and the Immediate Feedback condition. The Active Control condition served as the referent group.

Race/Ethnicity. Students' race/ethnicity was a three-level dummy-coded categorical variable indicating whether a student identified as non-Hispanic White, Asian, or some other ethnic/racial group. Non-Hispanic White was the referent group in the analyses.

Students' Biological Sex. Students' biological sex was a dummy-coded variable indicating whether a student identified as being male. Female was the reference group (0 = female, 1 = male).

Gifted Status. Students' identification as gifted was a dummy-coded variable in the analysis (0 = not identified as gifted, 1 = identified as gifted).

Accelerated Mathematics. All students in our study were placed in one of two levels of mathematics classrooms by the district: accelerated, in which the teachers implemented more challenging course materials, or regular mathematics classes. Student enrollment in an accelerated mathematics class was a dummy-coded variable (0 = not in an accelerated mathematics class, 1 = enrolled in an accelerated mathematics class).

Early Intervention Program Status. Student participation in an early intervention program (EIP) in elementary school

when identified as needing extra support was a dummy-coded variable (0 = not enrolled in EIP, 1 = enrolled in EIP).

Individualized Education Program Status. Students' individualized education program (IEP) status was a dummy-coded variable (0 = did not receive an IEP, 1 = received an IEP).

Classroom Format. Students' enrollment in an in-person versus virtual classroom was a dummy-coded variable (0 = in-person classroom, 1 = virtual classroom).

Completed Assignments. The total number of the nine intervention sessions a student completed was a continuous variable.

Intervention Conditions. In all four conditions, students accessed their learning technology through the ASSISTments platform. Students logged in to the platform with a username and a password and then clicked the assignment link that directed them to their assigned technology.

FH2T. FH2T consists of 14 worlds that focus on different mathematical concepts, and each world contains 18 problems (a total of 252 problems). In this study, all students in the FH2T condition started at World 1: Addition and worked their way through the worlds—World 2: Multiplication, World 3: Order of Operations + and \times , World 4: Subtraction and Negative Numbers, World 5: Mixed Practice of + and $-$, World 6: Division, World 7: Order of Operations, World 8: Equation + and $-$, World 9: Inverse Operations + and $-$, World 10: Distribution, World 11: Factoring, World 12: Equation +, $-$, \times , and \div , World 13: Inverse Operations, and World 14: Final Review. All students were given 30 minutes to play FH2T for each session, after which the system would log students out of the game. The system also saves students' progress, so students could start where they left off in each subsequent session.

DragonBox. DragonBox has 10 chapters, and each chapter contains 20 problems (a total of 200 problems). The problems together tap the Grade 7 mathematics standards in Common Core (Common Core State Standards for Mathematics, 2010) and cover the following content: addition, multiplication, division, parentheses, negative signs, fraction addition, combining like terms, variables, factoring, and substitution. Similar to FH2T, DragonBox students worked through problems at their own pace and started from where they left off in each subsequent session.

Because DragonBox is a phone- or tablet-based application that cannot be opened on a web browser, the study team provided tablets on which students played DragonBox in in-person classrooms. For students in virtual classrooms, the

research team provided instructions and licenses for students to download the game on their own device. Regardless of the classroom format, students would log in to ASSISTments on their laptop or Chromebook at the beginning of each session, follow the instructions to start the timer for the session, and then play DragonBox on a tablet or personal device.

Because DragonBox is a commercial application, access to students' progress and action-level data within the game was not possible. Therefore, at the end of each intervention session, students self-reported their progress—the chapter and problem at which they stopped that day—in ASSISTments. In case students did not report their progress at the end of the session, they were also asked to report their previous progress at the beginning of each intervention session before they started playing (after the first intervention session). Given that students in the DragonBox condition had to report their progress before and after gameplay, the timer for playing DragonBox was set to 25 minutes. This way, students could complete the sessions in 30 minutes without disrupting teachers' lesson plans and instruction.

Immediate Feedback. Students in the Immediate Feedback condition solved traditional mathematics problems in ASSISTments (Heffernan & Heffernan, 2014). To ensure that the problems aligned with traditional instruction, the study team selected and adapted problems from three popular open-source middle-school mathematics curricula—Engage NY (2014), Utah Math Project (2016), and Illustrative Mathematics (2017)—that already existed in ASSISTments. The topics covered across the nine intervention sessions were (a) addition and multiplication, (b) subtraction and negative numbers, (c) division and fractions, (d) order of operations, (e) addition and subtraction in equation-solving, (f) distribution, (g) factoring, (h) multiplication and division in equation-solving, and (i) review.

For each intervention session, students received a problem set consisting of 35–45 problems. They started on the first problem of a problem set at the beginning of each session and worked through the problems in the same order. After 25 minutes, or after students had completed the problem set, the system directed students to their session report, which showed their performance on each problem. Students were directed to review their session report, revisit problems, and review the hints and correct answers within the 30-minute time frame.

Active Control. In the Active Control condition, the mathematical problems and the study procedure were identical to the Immediate Feedback condition. The only difference between the two conditions was that the hints and correctness feedback were available *during* problem-solving in the Immediate Feedback condition, whereas the hints and correctness feedback were available only *after* problem-solving (i.e., after 25 minutes or completing the problem set) in the

TABLE 2

Percentage of Students Within Each Classroom Format Who Was Male, Gifted, in an Accelerated Program, or Enrolled in EIP or Had an IEP

Classroom format	Male	Gifted	Accelerated	EIP	IEP
In person	53.4%	10.2%	12.7%	8.8%	10.6%
Virtual	44.1%	30.1%	47.6%	4.0%	6.0%
Classroom format $t(92) =$	-4.528***	3.798***	2.260***	-2.295*	-1.830

Note. *** $p < .001$, * $p < .05$.

Active Control condition. At this point, students could revisit any question, review their session report, and review the hints and correct answers within the 30-minute time frame.

Analytic Approach

Intervention effects were tested through a series of three-level hierarchical linear models, with students ($N = 1,850$), nested within classrooms ($n = 127$), nested within teachers ($n = 34$). Classroom format (in person versus virtual) was a Level 2 predictor, and all other predictors were Level 1 variables. The number of assignments completed and pretest scores were centered on their respective grand means. All other predictors were dummy-coded. The outcome measure was student performance on the posttest assessment. The primary analyses were registered on the Open Science Framework (<https://osf.io/2r5zs>).

Results

Descriptive Statistics

Means and standard deviations for all variables were previously reported in Table 1. A correlation matrix is included in the online supplementary material to this article (Appendix C, Table C1); however, three sets of results within the correlation matrix are worth mentioning. As noted previously, intervention condition was related to pretest scores ($\chi^2(3, N = 1,850) = 13.303, p = .004$). In addition, condition was related to the number of assignments a student completed ($\chi^2(3, N = 1,850) = 28.657, p < .001$), with students in the DragonBox condition completing fewer assignments ($M = 5.526, SD = 3.382$) than those in the FH2T ($M = 6.603, SD = 2.857$), Immediate Feedback ($M = 6.811, SD = 2.679$), and Active Control conditions ($M = 6.858, SD = 2.770$). Intervention condition was unrelated to all other predictor variables, including child race/ethnicity ($\chi^2(6, N = 1,850) = 11.316, p = .079$), sex ($\chi^2(3, N = 1,850) = 2.322, p = .508$), gifted status ($\chi^2(3, N = 1,850) = 4.408, p = .221$), accelerated program ($\chi^2(3, N = 1,850) = 0.248, p = .969$), EIP status ($\chi^2(3, N = 1,850) = 0.881, p = .830$), IEP status ($\chi^2(3, N = 1,850) = 6.462, p = .091$), and enrollment in a virtual classroom ($\chi^2(3, N = 1,850) = 2.540, p = .468$).

Second, regarding the composition of virtual versus in-person classrooms, 67.3% of students in in-person classrooms were White, with 6.8% Asian and 25.9% identified with some other racial/ethnic group. In contrast, 61.7% of students enrolled in a virtual classroom were Asian, with 21.7% White and 16.7% identified with some other racial/ethnic group. Reflecting this pattern, enrollment in a virtual classroom was also related to most other predictors (see Table 2). Specifically, virtual classrooms had more gifted and accelerated students, but fewer males and fewer students enrolled in EIP. In addition, students in virtual classrooms completed more assignments ($M = 7.321, SD = 2.766$) than those in in-person classrooms ($M = 6.090, SD = 2.952, t(92) = 2.489, p = .015$) and had higher pretest scores (virtual: $M = 6.729, SD = 2.622$, and in-person: $M = 3.924, SD = 2.172$, respectively, $t(92) = 6.180, p < .001$).

Primary Analyses

An initial model tested intervention effects with no other covariates. This summarized the overall intervention effects only controlling for the inherent nested design. As shown in the first panel (i.e., Model 1) in Table 3, a significant intervention effect was observed ($\chi^2(3, N = 1,850) = 17.325, p = 0.001$). All three intervention conditions had higher performance on the posttest than did the Active Control condition, with the largest effect seen for DragonBox ($\gamma = .641$, Hedge's $g = .240$), followed by FH2T ($\gamma = .370$, Hedge's $g = .138$) and the Immediate Feedback conditions ($\gamma = .325$, Hedge's $g = .122$).

Model 2 controlled for various student demographic and academic covariates existing prior to the randomization. All were significantly related to posttest performance, with the exception of IEP status. Intervention condition continued to be statistically significant ($\chi^2(3, N = 1,850) = 12.226, p = 0.007$), with effects for FH2T ($\gamma = .338$) and DragonBox ($\gamma = .555$) significantly larger than the Active Control condition. To produce Model 3, an indicator was included for enrollment in an in-person or virtual classroom and a post-randomization variable, the number of assignments completed by the student (a marker of dosage). Both were statistically significant predictors of the posttest score, with higher posttest scores observed for students in virtual

TABLE 3

Summary of HLM Models Predicting Posttest Scores ($N = 1,850$)

	Model 1		Model 2		Model 3		Model 4	
	Coef	95% CI	Coef	95% CI	Coef	95% CI	Coef	95% CI
Intercept	4.084***	[3.497, 4.671]	3.577***	[3.327, 3.827]	3.462***	[3.178, 3.745]	3.446***	[3.149, 3.743]
FH2T	0.370*	[0.068, 0.672]	0.338*	[0.050, 0.627]	0.361**	[0.092, 0.630]	0.373**	[0.117, 0.630]
Dragon	0.641***	[0.328, 0.955]	0.555***	[0.243, 0.868]	0.719***	[0.408, 1.030]	0.733***	[0.432, 1.033]
Immediate Feedback	0.325*	[0.012, 0.638]	0.240	[-0.066, 0.547]	0.254	[-0.037, 0.545]	0.268	[-0.014, 0.550]
Pretest			0.313***	[0.265, 0.362]	0.279***	[0.235, 0.323]	0.199***	[0.120, 0.279]
Asian			1.032***	[0.675, 1.390]	0.824***	[0.430, 1.219]	0.827***	[0.439, 1.216]
Other race			0.167	[-0.077, 0.411]	0.130	[-0.109, 0.370]	0.135	[-0.110, 0.379]
Male			-0.322***	[-0.465, -0.180]	-0.255**	[-0.409, -0.100]	-0.259***	[-0.411, -0.106]
Gifted			0.616***	[0.343, 0.889]	0.612***	[0.325, 0.899]	0.610***	[0.328, 0.893]
Accelerated			1.792***	[1.296, 2.288]	1.668***	[1.178, 2.158]	1.679***	[1.197, 2.161]
EIP			-0.228*	[-0.411, -0.046]	-0.183*	[-0.360, -0.006]	-0.182*	[-0.361, -0.002]
IEP			0.217	[-0.078, 0.513]	0.208	[-0.063, 0.479]	0.197	[-0.075, 0.470]
Virtual					0.415*	[0.048, 0.782]	0.429*	[0.067, 0.791]
Assign %					0.120***	[0.088, 0.153]	0.119***	[0.087, 0.151]
FH2T × Pre							0.136**	[0.044, 0.227]
Dragon × Pre							0.048	[-0.094, 0.191]
Immediate Feedback × Pre							0.058	[-0.071, 0.186]

Note. All three coefficients also remained significant at $p < 0.05$ after Holm correction (Holm, 1979).

Model 1: Condition: $\chi^2(3, N = 1,850) = 17.325, p = 0.001$

Model 2: Condition: $\chi^2(3, N = 1,850) = 12.226, p = 0.007$; Race Effect: $\chi^2(3, N = 1,850) = 38.567, p < 0.001$

Model 3: Condition: $\chi^2(3, N = 1,850) = 20.863, p < 0.001$; Race Effect: $\chi^2(3, N = 1,850) = 19.962, p < 0.001$

Model 4: Condition: $\chi^2(3, N = 1,850) = 22.911, p < 0.001$; Race Effect: $\chi^2(3, N = 1,850) = 20.769, p < 0.001$; Cond x Pretest: $\chi^2(3, N = 1,850) = 11.192, p = 0.011$

Accelerated = Student in an accelerated mathematics program; Assign = Number of assignments completed; Dragon = DragonBox; Dragon × Pre = Dragon × Pretest; EIP = Enrolled in Early Intervention Program; FH2T = From Here to There; FH2T × Pre = FH2T × Pretest; Gifted = Identified as gifted; IEP = Has an Individual Education Program; Immediate = Immediate Feedback; Immediate Feedback × Pre = Immediate Feedback × Pretest Other race = Race/ethnicity other than Asian or White/non-Hispanic; Pretest = Pretest score (centered); Virtual = Enrolled in a virtual class.

classrooms ($\gamma = .415$) and those with a higher number of completed assignments ($\gamma = .120$). In this model, coefficients on intervention conditions were estimates of “natural direct effects” (VanderWeele, 2015, p.22): Effects of intervention were the number of completed assignments held fixed (assuming no unadjusted confounding between the number of completed assignments and posttest scores). Direct intervention effects were statistically significant ($\chi^2(3, N = 1,850) = 20.863, p < 0.001$), with students in FH2T ($\gamma = .361$, Hedge’s $g = .135$) and DragonBox conditions ($\gamma = .719$, Hedge’s $g = .269$) significantly outperforming students in the Active Control condition.

As our approach to attrition was conservative (excluding students without pretest data), we also fit the models including students with posttests but with missing pretests. These analyses were conducted by imputing the global mean for missing pretest scores, with an indicator variable for missing pretest scores. This approach yielded similar results as those presented here.

Follow-Up Interaction Analyses

A series of follow-up analyses tested for interactions between intervention condition and the various predictor

variables. As reported in Model 4 (see Table 3), a significant interaction was observed between intervention condition and pretest scores ($\chi^2(3, N = 1,850) = 11.192, p = 0.011$). This reflected larger FH2T effects among students with higher versus lower pretest performance (see Figure 2). No other interactions with intervention condition were found to be statistically significant: Race/Ethnicity: Asian: $\chi^2(6, N = 1,850) = 5.212, p = 0.517$; Male: $\chi^2(3, N = 1,850) = 1.193, p = 0.755$; Gifted: $\chi^2(3, N = 1,850) = 2.785, p = 0.426$; Accelerated program: $\chi^2(3, N = 1,850) = 1.799, p = 0.615$; EIP status: $\chi^2(3, N = 1,850) = 0.196, p = 0.978$; IEP status: $\chi^2(3, N = 1,850) = 1.204, p = 0.752$; Virtual classroom: $\chi^2(3, N = 1,850) = 3.870, p = 0.276$; Number of assignments completed: $\chi^2(3, N = 1,850) = 1.757, p = 0.624$.

Discussion

In summary, after a 4.5-hour intervention, students in FH2T and DragonBox conditions were found to have significantly higher posttest scores, reflecting students’ higher algebraic understanding, compared to students in the Active Control condition. These effects remained after controlling for students’ prior knowledge and demographic variables.

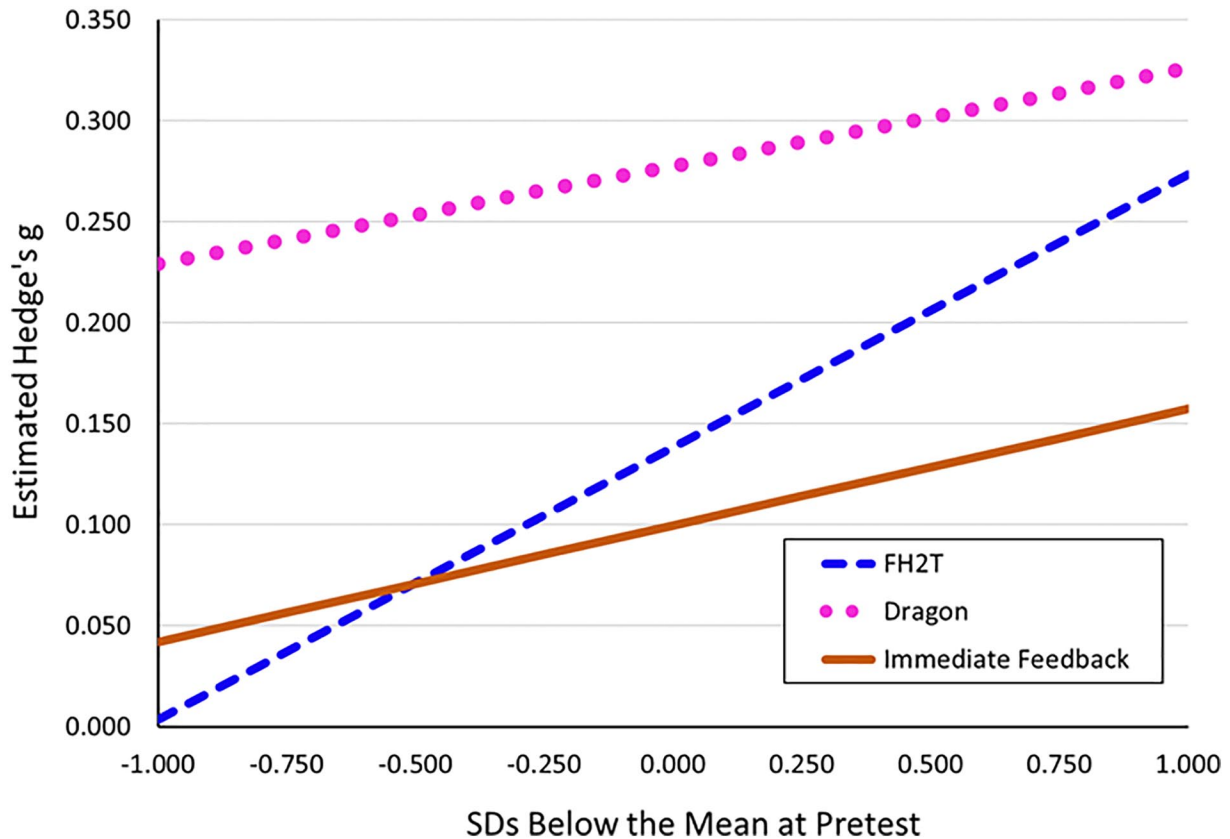


FIGURE 2. Interaction between pretest performance and intervention condition.

Although students in the Immediate Feedback condition also displayed significantly higher posttest scores compared to students in the Active Control condition, this effect was no longer significant after controlling for covariates. Together, the findings suggest that FH2T and DragonBox are effective digital games to support Grade 7 students' algebraic understanding compared to an active control condition using online problem sets with post-assignment feedback. Below, findings are discussed in detail.

Learning With Games: FH2T and DragonBox

In the current study, students in the two game-based learning interventions outperformed students in the Active Control condition on the posttest algebraic knowledge assessment. Further, the effect size for FH2T was similar to that reported in a prior study comparing the efficacy of FH2T to the Immediate Feedback condition (Hedge's $g = 0.16$; Chan et al., 2022), providing additional support for the consistent positive effects of FH2T. Although effects were found for both games, they were designed based on different theoretical perspectives. FH2T and DragonBox are grounded in intuitive perceptual-motor routines and game-based interface; however, the two games differ in important ways. For instance, FH2T introduces algebraic notations from the start,

whereas DragonBox gradually transitions students from pictures of monsters to abstract notations. Further, DragonBox asks students to isolate variables across notational contexts, whereas FH2T asks students to transform between mathematically equivalent but perceptually different expressions or equations (e.g., $9 \cdot 12 + 9 \cdot 23$ and $9 \cdot 7 \cdot 5$). It is possible that the progression from concrete pictures to abstract symbols (Fyfe et al., 2014) and the focus on isolating variables are elements within DragonBox that effectively support students' algebra performance. However, qualitative investigations of teacher and student interactions suggest that the design of DragonBox can make it difficult for teachers and students to speak precisely about concepts introduced by the game and to connect these concepts back to mathematics (Dolonen & Kluge, 2015). FH2T provides students with opportunities to explore and learn which gesture-actions with algebraic notations are appropriate and allowed across mathematical contexts (Dörfler, 2003; Landy & Goldstone, 2009; Nogueira de Lima & Tall, 2008). However, because game-based learning has other potential benefits, such as increasing motivation (Hidi & Renninger, 2006; Rotgans & Schmidt, 2011) and enhancing engagement (Mora et al., 2015), it may be the case that DragonBox and FH2T support algebraic learning through engaging students with mathematical content and fostering positive attitudes toward

mathematics. Although the current study does not tease apart these potential mechanisms, it does provide a foundation for future inquiries that could explore how various game elements differentially affect student learning from educational technology games. For example, future analyses could use the clickstream data within FH2T to explore the relations between students' in-game behaviors and their learning outcomes.

Alternatively, the positive impacts of FH2T and DragonBox may be a result of their game-based context, which may increase students' interest and engagement with mathematics and, in turn, improve their performance. Substantial research has highlighted the benefits of integrating gamification elements, such as rewards, leveling, challenges, supporting player skills, player controls, and feedback, into instruction to promote learning (Garris et al., 2002; Gee, 2003; Kallou & Mohan, 2015). Technology-based educational games have been found to increase engagement and motivation, which have also increased student problem-solving, and learning (Connolly et al., 2012; Foster, 2008; Ke, 2008; Samur & Evans, 2012). Therefore, the motivational factors related to game-based learning may be another potential mechanism through which FH2T and DragonBox may support students' algebraic understanding.

Further, anecdotes from participating teachers revealed that some students in the two problem-set conditions lacked the motivation to solve textbook problems, knowing that other students in the same classroom were playing games. Although the student-level randomization maximizes the statistical power to detect intervention effects, administering four different conditions within each classroom may inevitably highlight the differences between the interventions and ultimately contribute to the impacts of the game-based interventions compared to the problem-set conditions. Because of the COVID context, 32.7% of students received the intervention through their virtual classroom and worked on their study assignments at home, without the awareness of other intervention conditions. Additional data were also collected on students' attitudes toward mathematics as well as their engagement with the technologies. Future analyses will use these additional data to explore whether game-based interventions supported algebraic understanding through increasing students' motivations or engagements with their intervention and how the classroom context of the current study may interact with the intervention condition to influence student learning.

Interaction Between FH2T and Pretest Scores

Study effects for FH2T and DragonBox were largely unchanged when investigating interactions with multiple demographic and program variables related to students. However, results did indicate a significantly larger effect for

FH2T among students who scored higher on the pretest. This finding suggests additional benefits of FH2T for students who already possess foundational or basic knowledge of algebraic concepts. In particular, the design features of FH2T, such as introducing algebraic notations from the start and having flexible goals, may require students to have some foundational algebraic knowledge to further benefit from the game. Past research on FH2T has been mixed in finding this relation. For example, in an investigation of Grades 6–7 students, Chan et al. (2022) do not find a significant interaction between level of pretest and student understanding of mathematical equivalence. However, in a sample of Grade 2 students, a significant interaction between pretest and progress within FH2T is found (Hulse et al., 2019). Specifically, among students with lower, as opposed to higher, prior knowledge, solving more problems in FH2T is associated with higher posttest scores on an arithmetic assessment. The moderating effects of prior knowledge may vary, depending on students' grade level (i.e., second, sixth, seventh), the particular FH2T variable used (i.e., intervention assignment vs. progress within FH2T), and the outcome of interest (i.e., algebraic knowledge, mathematical equivalence, arithmetic). Future research should continue to explore the extent to which prior knowledge moderates the relationship between using FH2T and algebraic knowledge gain. Doing so will advance the understanding of when and for whom such interventions as FH2T are beneficial for learning mathematics.

Online Problem Sets

Although students in the Immediate Feedback condition also displayed significantly higher posttest scores compared to students in the Active Control condition, this effect was no longer significant after controlling for several covariates. Different from the prior findings within ASSISTments (Kelly et al., 2013; Mendicino et al., 2009), this study did not find consistent evidence that students in the Immediate Feedback condition outperformed the students in the Active Control condition with post-assignment feedback. It is important to note that the two problem-set conditions only involved the student aspects of ASSISTments (i.e., hints, feedback, performance report), whereas the prior efficacy study on ASSISTments also involved teacher training that provided support on monitoring student progress within the system and using student performance to inform instructional practices in the classroom (Roschelle et al., 2016). The current findings, therefore, extend prior research and suggest that implementing only student aspects of ASSISTments in a homework-style intervention may not consistently support students' algebra learning. Teacher training may be a critical component of the ASSISTments intervention that effectively supports students' mathematics achievement.

Study Context

It is important to note that the study was conducted within the larger context of the COVID-19 pandemic. This context affected many aspects of the current study. For instance, although significant effects were found for posttest scores of students in the FH2T and DragonBox conditions, overall, scores were lower at posttest than at pretest. This is likely due to interrupted learning and the overarching lack of typical gains seen across the country in mathematics due to the pandemic (Dorn et al., 2020). For example, Kuhfeld and Lewis (2022) find that the pandemic has had larger negative impacts on mathematics compared to reading and that those effects have not significantly diminished as of the 2021–2022 school year. Additionally, the most recent long-term trend National Assessment of Educational Progress results in mathematics achievement for 13-year-olds indicate a significant drop since the last data collection in 2012 (U.S. Department of Education, n.d.). Within this context, however, the current findings seem to suggest that FH2T and DragonBox were able to serve as supports to reduce the effects of interrupted learning for students in those conditions.

Additionally, the school district offered the students and their families options for synchronous in-person instruction in the school buildings or asynchronous virtual instruction at home (with the addition of synchronous help sessions). Regardless of the instructional format, all students in the current study received their assigned intervention online and worked individually at their own pace, using a device. Although the study procedure was not affected by the classroom format, the differences in students' classroom format inadvertently introduced systematic variations in the study contexts. Specifically, students in in-person classrooms were aware of other intervention conditions, whereas students in virtual classrooms were not. Although no significant interactions were identified between study condition and instructional format, the presence versus absence of this awareness may have influenced students' motivation to complete their intervention in unmeasured ways and, in turn, may have amplified the effects of game-based learning compared to traditional problem-solving.

Further, whether students enrolled in in-person or virtual classrooms was not random. A higher proportion of Asian students and high-performing students opted for virtual instruction, whereas a higher proportion of White students and low-performing students selected in-person instruction. These systematic differences between in-person and virtual classrooms may have been related to the high proportion of multigenerational households among Asian students, their anxiety about being bullied due to political and racial tensions, or satisfaction with virtual learning (Balingit et al., 2021). Regardless of the reasons behind the families' decision for enrolling students in in-person or virtual classrooms,

their choices had implications on the student compositions within the classrooms and school buildings. Specifically, as a higher proportion of Asian students and high-performing students opted for virtual instruction, the heterogeneity of the in-person classrooms was reduced through this process. Given the important differences between in-person and virtual classrooms, analyses controlled for the classroom format and found that intervention effects remained significant. Future work will further explore how the pandemic context affected student learning.

Limitations and Future Directions

The current study was limited in several ways. First, the attrition in the current study was nontrivial. Among the 3,612 students who participated at the start of the intervention, only 1,850 students completed pretest and posttest and thus were included in the current analyses. Although almost half of the students had dropped out of the study by the end of the intervention, the dropout rate was not related to the intervention condition, and there was minimal differential attrition between conditions. Second, as each technology condition functioned differently, time within the 30-minute sessions was used differently (all practice in FH2T, while ASSISTments included practice and review). Students in the DragonBox condition needed to stop a few minutes early to log their progress, which meant that those students did spend fewer minutes engaging in the technology. Although this step was necessary to maintain equal time within the classroom across conditions, future studies may explore differences in timing and overall effectiveness. Third, the current study focused on comparing each of the three treatment conditions to an active control condition. Future studies will examine differences across conditions to further inform the ways in which these technologies relate to the conceptual, procedural, and flexibility of algebraic thinking. Fourth, the current study provided evidence that Grade 7 students in the FH2T and DragonBox conditions did significantly better on the posttest reflecting algebraic understanding, but it did not offer insights into their areas of improvement or potential mechanisms. Preliminary follow-up analyses have revealed that FH2T and DragonBox improved on conceptual knowledge, but not on procedural knowledge or procedural flexibility (Chan et al., 2023). Future studies using the log data will explore *how* these intervention conditions affect aspects of students' conceptual knowledge, whether students' understanding was moderated by their prior knowledge or other demographic factors, and how the fidelity of implementation at the student level influenced student learning from each intervention. Addressing these questions is important, as they provide crucial information on the ways in which the interventions support algebra learning, for whom the interventions were effective, and the amount of intervention needed to result in learning gains. Finally, given the unique

context of the COVID-19 pandemic, the data set may offer insights into how student learning was affected by the pandemic and inform future practices across instructional format and during stressful emergency situations.

Conclusion

Overall, the current study contributes to the efficacy of FH2T and DragonBox as interventions for supporting algebraic performance. Both games are promising interventions that focus on training students' perceptual-motor routines in algebraic reasoning, yet they differ in several important ways. The findings have implications for future research in delineating the effective elements within each game. Further, they also have implications for educational practices, as they provide evidence that gamified learning platforms are effective ways for students to explore mathematical ideas and support their understanding.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through an Efficacy and Replication Grant (R305A180401) to Worcester Polytechnic Institute. The opinions expressed are those of the authors and do not represent views of the institute or the U.S. Department of Education. The authors would like to especially thank Kathleen, Barbara, and the teachers in the district who gave their time to participate, without whom the study would not have been possible. Also, thank you to the teams at Graspable Math, DragonBox, and ASSISTments for providing access to the technologies, as well as to Dr. Erin Ottmar for her vision of the study.

ORCID iD

Lauren E. Decker-Woodrow  <https://orcid.org/0000-0001-9404-567X>

Open Practices

The access to data files for this article can be found at <https://osf.io/r3nf2/>

Supplemental Material

Supplemental material for this article is available online.

References

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The future of embodied design for mathematics teaching and learning. *Frontiers in Education, 5*, 147.

Alibali, M. W., Crooks, N. M., & McNeil, N. M. (2018). Perceptual support promotes strategy generation: Evidence from equation solving. *British Journal of Developmental Psychology, 36*, 153–168. <https://doi.org/10.1111/bjdp.12203>

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences, 21*(2), 247–286. <https://doi.org/10.1080/10508406.2011.611446>

Azevedo, R., & Bernard, R. M. (1995, April 18–22). *The effects of computer-presented feedback on learning from computer-based instruction: A meta-analysis* [Paper presentation]. Annual Meeting of the American Educational Research Association, San Francisco, CA, United States.

Balingit, M., Natanson, H., & Chen, Y. (2021, March 4). As schools reopen, Asian American students are missing from classrooms. *Washington Post*. http://www.washingtonpost.com/education/asian-american-students-home-school-in-person-pandemic/2021/03/02/eb7056bc-7786-11eb-8115-9ad5e9c02117_story.html

Bergan, J. R., Sladeczek, I. E., Schwarz, R. D., & Smith, A. N. (1991). Effects of a measurement and planning system on kindergartners' cognitive development and educational programming. *American Educational Research Journal, 28*(3), 683–714. <https://doi.org/10.3102/00028312028003683>

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–74. <https://doi.org/10.1080/0969595980050102>

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, and Evaluation, 8*(9), 1–8.

Braithwaite, D. W., Goldstone, R. L., van der Maas, H. L. J., & Landy, D. H. (2016). Non-formal mechanisms in mathematical cognitive development: The case of arithmetic. *Cognition, 149*, 40–55. <https://doi.org/10.1016/j.cognition.2016.01.004>

Butler, A. C., & Woodward, N. R. (2018). Toward consilience in the use of task-level feedback to promote learning. In K. Federmeier (Ed.), *Psychology of Learning and Motivation* (pp. 1–38). Elsevier. <https://doi.org/10.1016/bs.plm.2018.09.001>

Cayton-Hodges, G. A., Feng, G., & Pan, X. (2015). Tablet-based math assessment: What can we learn from math apps?. *Journal of Educational Technology and Society, 18*(2), 3–20.

Chan, J. Y.-C., Closser, A. H., Ngo, V., Smith, H., Liu, A., & Ottmar, E. (2023). Examining shifts in conceptual knowledge, procedural knowledge, and procedural flexibility in the context of two game-based technologies. *Journal of Computer Assisted Learning*. <https://doi.org/10.1111/jcal.12798>

Chan, J. Y.-C., Lee, J.-E., Mason, C. A., Sawrey, K., & Ottmar, E. (2022). From Here to There! A dynamic algebraic notation system improves understanding of equivalence in middle-school students. *Journal of Educational Psychology, 114*(1), 56–71. <https://doi.org/10.1037/edu0000596>

Common Core State Standards for Mathematics. (2010). *Common Core State Standards for Mathematics* (CCSSM). National Governors Association Center for Best Practices and the Council of Chief State School Officers.

Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. *Computers and Education, 59*(2), 661–686. <https://doi.org/10.1016/J.COMPEDU.2012.03.004>

Corbett, A. T., & Anderson, J. R. (2001, March). Locus of feedback control in computer-based tutoring: Impact on learning

- rate, achievement and attitudes. In J. Jacko & A. Sears (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 245–252). Association for Computing Machinery.
- Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *Psychological Record*, 53(4), 533–548.
- Dolonen, J. A., & Kluge, A. (2015). Algebra learning through digital gaming in school. *CSSL Proceedings*, 252–259.
- Dörfler, W. (2003). Mathematics and mathematics education: Content and people, relation and difference. *Educational Studies in Mathematics 2003*, 54(2), 147–170. <https://doi.org/10.1023/B:EDUC.0000006118.25919.07>
- Dorn, E., Hancock, B., Sarakatsannis, J., & Viruleg, E. (2020). *COVID-19 and student learning in the United States: The hurt could last a lifetime*. McKinsey and Company. <http://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-and-student-learning-in-the-united-states-the-hurt-could-last-a-lifetime?cid=eml-web>
- Engage, NY. (2014). *New York State Education Department*. <https://www.engageny.org>
- Foglia, L., & Wilson, R. A. (2013). Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(3). <https://doi.org/10.1002/wcs.1226>
- Foster, A. (2008). Games and motivation to learn science: Personal identity, applicability, relevance and meaningfulness. *Journal of Interactive Learning Research*, 19(4), 597–614.
- Fyfe, E. R., McNeil, N. M., Son, J. Y., & Goldstone, R. L. (2014). Concreteness fading in mathematics and science instruction: A systematic review. *Educational Psychology Review*, 26(1), 9–25. <https://doi.org/10.1007/s10648-014-9249-3>
- Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation and Gaming*, 33(4), 441–467. <https://doi.org/10.1177/1046878102238607>
- Gee, J. (2003). What videogames have to teach us about learning and literacy. *Computers in Entertainment*, 1(1), 1–4.
- Goldstone, R. L., Marghetis, T., Weitnauer, E., Ottmar, E. R., & Landy, D. (2017). Adapting perception, action, and technology for mathematical reasoning. *Current Directions in Psychological Science*, 26(5), 434–441. <https://doi.org/10.1177/0963721417704888>
- Hansen, B. B., & Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2), 219–236. <https://doi.org/10.1214/08-STS254>
- Harrison, A., Smith, H., Hulse, T., & Ottmar, E. R. (2020). Spacing out! Manipulating spatial features in mathematical expressions affects performance. *Journal of Numerical Cognition*, 6(2), 186–203. <https://doi.org/10.5964/jnc.v6i2.243>
- Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24, 470–497. <https://doi.org/10.1007/s40593-014-0024-x>
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.
- Hulse, T., Daigle, M., Manzo, D., Braith, L., Harrison, A., & Ottmar, E. R. (2019). From here to there! Elementary: A game-based approach to developing number sense and early algebraic understanding. *Education Technology Research Development*, 67, 423–441.
- Illustrative Mathematics. (2017). *Illustrative mathematics project*. <https://www.illustrativemathematics.org>
- Jacob, M., & Hochstein, S. (2008). Set recognition as a window to perceptual and cognitive processes. *Perception and Psychophysics*, 70(7), 1165–1184. <https://doi.org/10.3758/PP.70.7.1165>
- Kahoot! (2019, May 9). *Kahoot! and DragonBox join forces to create an awesome math learning experience for all* [Press release]. <https://kahoot.com/press/2019/05/09/kahoot-dragonbox-join-forces-create-awesome-math-learning-experience-for-all/>
- Kalloo, V., & Mohan, P. (2015). A technique for mapping mathematics content to game design. *International Journal of Serious Games*, 2(4). <https://doi.org/10.17083/ijsg.v2i4.95>
- Ke, F. (2008). A case study of computer gaming for math: Engaged learning from gameplay? *Computers and Education*, 51(4), 1609–1620. <https://doi.org/10.1016/j.compedu.2008.03.003>
- Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, 2(2), 285–305. <https://doi.org/10.1111/j.1756-8765.2009.01053.x>
- Kelly, K., Heffernan, N., Heffernan, C., Goldman, S., Pellegrino, J., & Soffer-Goldstein, D. (2013). Estimating the effect of web-based homework. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *CEUR Workshop Proceedings (Vol. 1009, pp. 824–827)*. CEUR-WS. https://doi.org/10.1007/978-3-642-39112-5_122
- Kena, G., Musu-Gillette, L., Robinson, J., Wang, X., Rathbun, A., Zhang, J., Wilkinson-Flicker, S., Barmer, A., & Dunlop Velez, E. (2015). *The Condition of Education 2015* (NCES 2015-144). U.S. Department of Education, National Center for Education Statistics. Retrieved from <http://nces.ed.gov/pubsearch>
- Kieran, C. (2006). Research on the learning and teaching of algebra. In Á. Gutiérrez, & P. Boero (Eds.), *Handbook of Research on the Psychology of Mathematics Education* (pp. 11–49). Sense Publisher. <https://doi.org/10.1163/9789087901127>
- Kirshner, D., & Awtry, T. (2004). Visual salience of algebraic transformations. *Journal for Research in Mathematics Education*, 35(4), 224–257. <https://doi.org/10.2307/30034809>
- Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, 13(2), 129–164. https://doi.org/10.1207/s15327809jls1302_1
- Kuhfeld, M., & Lewis, K. (2022). *Student achievement in 2021-22: Cause for hope and continued urgency*. NWEA. <https://www.nwea.org/research/publication/student-achievement-in-2021-22-cause-for-hope-and-continued-urgency>
- Landy, D., & Goldstone, R. L. (2010). Proximity and precedence in arithmetic. *Quarterly Journal of Experimental Psychology*, 63(10), 1953–1968. <https://doi.org/10.1080/17470211003787619>

- Landy, D. H., & Goldstone, R. L. (2009). How much of symbolic manipulation is just symbol pushing? In N. A. Taatgen (Ed.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1318–1323). Cognitive Science Society. Retrieved from <http://csjarchive.cogsci.rpi.edu/Proceedings/2009/papers/253/paper253.pdf>
- Liu, Y.-E., Ballweber, C., O'Rourke, E., Butler, E., Thummaphan, P., & Popović, Z. (2015). Large-scale educational campaigns. *ACM Transactions on Computer-Human Interaction*, 22(2), 1–24. <https://doi.org/10.1145/2699760>
- Long, Y., & Alevan, V. (2017). Educational game and intelligent tutoring system: A classroom study and comparative design analysis. *ACM Transactions on Computer-Human Interaction*, 24(3). <https://doi.org/10.1145/3057889>
- Long, Y., & Alevan, V. (2014). Gamification of joint student/system control over problem selection in a linear equation tutor. In S. Trausan-Matu, K. E. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 378–387). Springer-Verlag. <https://doi.org/10.1007/978-3-319-07221-0>
- Marquis, J. (1988). Common mistakes in algebra. In A. F. Coxford, & P. Shulte (Eds.), *The Ideas of Algebra, K–12: National Council of Teachers of Mathematics yearbook* (pp. 204–205). National Council of Teachers of Mathematics.
- Mendicino, M., Razaq, L., & Heffernan, N. T. (2009). A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education*, 41(3), 331–359.
- Mora, A., Riera, D., Gonzalez, C., & Arnedo-Moreno, J. (2015). A literature review of gamification design frameworks. *2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games)*, 1–8. IEEE.
- Nathan, M. J., Ottmar, E. R., Abrahamson, D., Williams-Pierce, C., Walkington, C., & Nemirovsky, R. (2016). Embodied mathematical imagination and cognition (EMIC) working group. In M. B. Wood, E. E. Turner, M. Civil & J. A. Eli (Eds.), *Psychology of Mathematics Education—North American Chapter conference*. (pp. 1697–1697). The University of Arizona.
- Nathan, M. J., & Walkington, C. (2017). Grounded and embodied mathematical cognition: Promoting mathematical insight and proof using action and language. *Cognitive Research: Principles and Implications*, 2(1). <https://doi.org/10.1186/s41235-016-0040-5>
- Nathan, M. J., Walkington, C., Boncoddio, R., Pier, E., Williams, C. C., & Alibali, M. W. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical proof. *Learning and Instruction*, 33. <https://doi.org/10.1016/j.learninstruc.2014.07.001>
- National Mathematics Advisory Panel. (2008). *Foundations for Success: The Final Report of the National Mathematics Advisory Panel*. U.S. Department of Education.
- Nogueira de Lima, R., & Tall, D. (2008). Procedural embodiment and magic in linear equations. *Educational Studies in Mathematics*, 67(1), 3–18. <https://doi.org/10.1007/s10649-007-9086-0>
- Ottmar, E. R., Landy, D., Goldstone, R., & Weitnauer, E. (2015). Getting From Here to There! Testing the effectiveness of an interactive mathematics intervention embedding perceptual learning. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society* (pp. 1793–1798). Cognitive Science Society.
- Patsenko, E. G., & Altmann, E. M. (2010). How planful is routine behavior? A selective-attention model of performance in the Tower of Hanoi. *Journal of Experimental Psychology: General*, 139(1), 95–116. <https://doi.org/10.1037/A0018268>
- Phye, G. D., & Andre, T. (1989). Delayed retention effect: Attention, perseverance, or both? *Contemporary Educational Psychology*, 14(2), 173–185. [https://doi.org/10.1016/0361-476X\(89\)90035-0](https://doi.org/10.1016/0361-476X(89)90035-0)
- Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, 2(4), 233285841667396. <https://doi.org/10.1177/2332858416673968>
- Rotgans, J. I., & Schmidt, H. G. (2011). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction*, 21(1), 58–67. <https://doi.org/10.1016/j.learninstruc.2009.11.001>
- Samur, Y., & Evans, M. A. (2012, April 13–17). *The effects of serious games on performance and engagement: A review of the literature (2001–2011)*. In [Conference presentation]. American Educational Research Association Conference, Vancouver, British Columbia, Canada.
- Shapiro, L. (2010). *Embodied cognition*. Routledge. <https://doi.org/10.4324/9781315180380>
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Siew, N. M., Geoffrey, J., & Lee, B. N. (2016). Students' algebraic thinking and attitudes toward algebra: The effects of game-based learning using Dragonbox 12+ app. *Research Journal of Mathematics and Technology*, 5(1), 66–79.
- Speece, D. L., Molloy, D. E., & Case, L. P. (2003). Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities: Research and Practice*, 18(3), 147–156.
- Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, 40, 41–54. <https://doi.org/10.1016/j.cedpsych.2014.05.005>
- Torres, R., Toups, Z. O., Wiburg, K., Chamberlin, B., Gomez, C., & Ozer, M. A. (2016). Initial design implications for early algebra games. In A. Cox, & Z. O. Toups (Eds), *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (pp. 325–333). ACM. <https://doi.org/10.1145/2968120.2987748>
- U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress. (n. d.). *Long-term trend reading and mathematics assessments*. <https://www.nationsreportcard.gov/ltr/?age=9#overall-results-mathematics>
- Utah Math Project. (2016). *The Utah middle school math project*. <http://utahmiddleschoolmath.org>
- VanderWeele, T. J. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

What Works Clearinghouse. (2020). *What Works Clearinghouse standards handbook, Version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Available at <https://ies.ed.gov/ncee/wwc/handbook>

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9(4), 625–636. <https://doi.org/10.3758/BF03196322>

Authors

LAUREN E. DECKER-WOODROW is a principal research associate in the Education Studies practice at Westat; email: lauren-woodrow@westat.com. She is a research methodologist with interests in teacher and classroom instructional quality and supports that influence learning and growth for children.

CRAIG A. MASON is a professor of education and applied quantitative methods at the University of Maine; email: craig.mason@maine.edu. He is a research methodologist with broad interests in growth modeling as a tool for identifying factors that influence developmental trajectories in children.

JJ-EUN LEE is a postdoctoral research scientist at Worcester Polytechnic Institute; email: jlee13@wpi.edu. Her research interests focus on applying learning analytics and educational data-mining techniques to improve instructional design and student learning in online learning environments.

JENNY YUN-CHEN CHAN is an assistant professor of Early Childhood Education at the Education University of Hong Kong; email: chanjyc@eduhk.hk. Her research focuses on (a) the effects of perceptual and contextual features in the environment on learning mathematics and (b) the influences of language and executive function skills on mathematical thinking.

ADAM SALES is an assistant professor at Worcester Polytechnic Institute; email: asales@wpi.edu. His research interests focus on methods for causal inference using large, administrative data sets, primarily with applications in learning sciences and social sciences.

ALLISON LIU is a postdoctoral scholar at Worcester Polytechnic Institute; email: aliu2@wpi.edu. Her research focuses on understanding how learner characteristics and task design features influence STEM learning across the life-span, with the goal of informing instructional practices and interventions.

SHIHFEN TU is a professor of education and applied quantitative methods and director of the School of Learning and Teaching at the University of Maine; email: shihfen.tu@maine.edu. She specializes in linking and analyzing large population databases to study developmental issues in the fields of health, education, and developmental disabilities and is also interested in concept formation associated with STEM learning and using technology to teach mathematics.