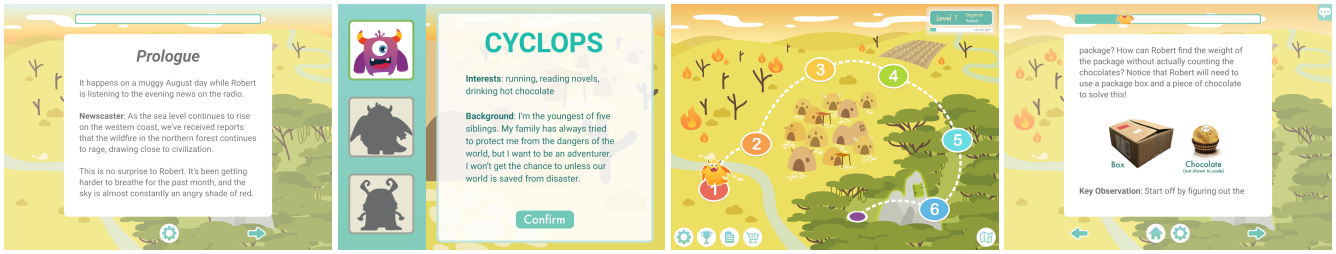
Supporting Children's Math Learning with Feedback-Augmented Narrative Technology

Sherry Ruan[1], Jiayu He[1], Rui Ying[1,2], Jonathan Burkle[1], Dunia Hakim[1], Anna Wang[1,3], Yufeng Yin[1,2,4], Lily Zhou[1], Qianyao Xu[1,2], Abdallah AbuHashem[1], Griffin Dietz[1], Elizabeth L. Murnane[1], Emma Brunskill[1], James A. Landay[1]

[1]Stanford University,[2]Tsinghua University,[3]University of Pennsylvania,[4]University of Southern California

{ssruan, jiayuhe, jtburkle, dunia, lilyzhou, aabuhash, gdietz44, emurnane, ebrun, landay}@stanford.edu



(a) Story interface (b) Monster selection interface (c) Map interface (d) Learning task interface

Figure 1: Our platform provides an interactive, narrative-based learning experience. (a) The user is a protagonist in a multi-chapter story. (b) A customizable monster chatbot provides help and personalized feedback during the adventure. (c) Accompanied by the selected monster, the user embarks on a heroic journey to save the world from natural disasters. (d) To foster genuine interest in learning, the narrative is interspersed with interactive problems that tie the real world into the virtual experience.

ABSTRACT

A key challenge in education is effectively engaging children in learning activities. We investigated how a narrative story impacts engagement and learning, as well as how feedback can provide further benefits. To do so, we created an interactive, tablet-based learning platform with a multi-step math task designed using Common Core State Standards. Subjects completed a pretest and then were assigned to a condition, either one of three variations of the system (narratives, narratives with hints, and narratives with a tutoring chatbot using wizard-of-oz techniques) or a control system that has children complete the same learning task without narratives nor feedback, before the subjects completed a post test. 72 children in U.S. grades 3–5 participated. Our results showed that embedding learning activities into narratives boosted children's engagement as evaluated by coding video responses and surveys, and the integration of a tutoring chatbot improved learning outcomes on the assessment. These results provide evidence that a narrative-based tutoring system with chatbot-mediated help may support effective learning experiences for children.

Author Keywords

Narrative-based learning; intelligent tutoring system; chatbot.

IDC '20 Interaction Design and Children, June 21–24, 2020, London, UK

ACM ISBN 978-1-4503-7981-6/20/06.

DOI: 10.1145/3392063.3394400

CCS Concepts

•**Applied computing** → **Education;** •**Social and professional topics** → *K-12 education;* Children; •**Human-centered computing** → Empirical studies in HCI;

INTRODUCTION

Engagement is known to be a key factor in student performance and achievement. One study found that elementary school students rated as highly engaged by teachers were twice as likely to do well on performance indices, and middle school students rated as highly engaged were 75% more likely to do well [43]. Unfortunately many students are not engaged in traditional classroom activities. One promising way to engage children in learning activities is by creating experiences similar to the media they are already motivated to use (e.g., books, films, television shows, and video games), including to tap into the inherent enjoyment of *narrative* [24]. Some popular novels, such as *The Number Devil* [22], do leverage narrative to immerse readers while simultaneously teaching them about a topic (e.g., math; see Whitin [75] for review). However, as these books are non-interactive, they do not adapt to a reader's abilities or circumstances and do not offer feedback that could enhance learning outcomes. On the other hand, computationally-augmented modes of learning have the potential to deliver personalized education in a cost-effective and scalable manner. However, many existing models, such as Massive Open Online Courses (MOOCs) and other styles of online education, still use standard curricula and instructional methods (e.g., traditional exercises, lectures, and texts) and are primarily successful with students who would already be motivated in conventional education environments [16, 18].

Computer-based educational games incorporate more informal approaches and show promise in engaging a wide range of learners, though there is still need for more evidence-based research to delineate the effect of different design features with respect to various learners and learning content [11, 49, 77]. Regarding educational games' use of narrative specifically, theories suggest that a narrative context can benefit learning in many ways such as contextualizing learning, providing background for sense-making, promoting intrinsic motivation, and improving concept retention [24, 72]. However, empirical studies have provided mixed results regarding the effectiveness of narrative in digital learning systems [2, 51, 76], necessitating research to reconcile the discrepancy. Meanwhile, advancements in design and AI have sparked interest in their application to learning technologies, especially to provide real-time feedback using natural language dialogue [35, 64]. Recent work suggests the educational benefits of such technologies, including chatbots [25]; yet observed effect sizes are modest, and more work is necessary to better understand how much and in what ways such technologies enhance learning.

In this paper, we examine the learning and engagement effects of technologies that support instructional features — narratives, hint feedback, and chatbot interactions — to investigate their merit in teaching math content to children. Specifically, we designed and built an interactive, narrative-based learning platform that contains a multi-step math learning task based on Common Core State Standards [7]. We also added a hint extension we designed to resemble features commonly found in current educational applications [1, 30, 32] as well as a tutoring chatbot instantiated using wizard-of-oz techniques. These value-added conditions were compared against a control system that asked users to do the same learning task but without any narratives or feedback. We ran a between-subjects study across these four conditions with 72 children in U.S. grades 3–5. Results indicate that embedding a learning task into narratives can make learning significantly more engaging and also improve children's learning outcomes; these findings suggest design insights for building future child-centric, intelligent learning systems that are both fun and effective.

RELATED WORK

We review three classes of related work: narrative-based learning, feedback-augmented learning, and educational games.

Narrative-Based Learning

There is a strong theoretical foundation to support the idea that including narrative in educational instruction can enhance learning. Narrative can be considered as incorporating a clear plotline that directs a learner's action toward a goal in the story [2, 76]. A narrative framing leverages the intrinsic appeal of stories and their ability to help us make sense of the world, including to develop mental models about abstract concepts [24, 66]. Similarly, narratives provide an intuitive and relatable context for subject matter [45, 50, 51], which also improves self-regulated learning [61]. Narrative can further captivate learners because it increases immersion [44, 50, 51].

While literary theory establishes the benefits of narrative, existing work to develop or test narrative-based learning technologies faces a few common limitations. Specifically, prior evaluations have typically relied on subjective measures or have been preliminary in scope, largely due to the difficulty in developing and deploying such systems [51]. In fact, some efforts focus only on developing the design specification for narrative learning environments rather than the actual implementation and evaluation of a functional tool [10].

Another challenge is that studies involving empirical work often produce inconclusive results regarding the efficacy of narrative-based learning technologies, and their targeted populations and learning topics vary widely. A meta-analysis conducted by Wouters et al. [76] suggests that compared to a conventional learning method, serious games with a narrative are (non-significantly) less effective than serious games without a narrative. For example, Adams et al. [2] found that a strong narrative theme does not necessarily promote learning, observing that students in a narrative condition performed only slightly better. This work also indicates that narrative content can distract students from learning content and misdirect cognitive capacity towards processing the story, confirming that superfluous narrative details impede learners' recall and problem-solving ability [24]. McQuiggan et al. [51] found that, despite higher levels of student reported presence, a narrative condition showed the lowest learning gains compared to conditions involving minimal-narrative or only PowerPoint.

Other studies more strongly demonstrate the efficacy of narrative-based digital learning tools. For example, elementary school children who used a version of an arithmetic game with a narrative fantasy theme performed better in a post-test [19]. Similarly, greater learning and transfer was shown by 3–4th graders who used the version of an educational computer program that incorporated a fantasy context [60]. Overall, despite the benefits of narrative promised by theory, efforts to implement and evaluate narrative-based learning technologies are only just emerging, motivating further research to explore best practices for adding narrative themes to learning technologies.

Feedback-Augmented Learning

There is a general consensus that providing formative feedback that goes beyond verification of answers to explain and modify a student's thinking can foster learning [27, 34, 68]. In digital learning tools, feedback can be delivered through pedagogical agents, such as characters programmed to guide learners [36]. With advances in AI and language technologies, there is increasing interest in agents that integrate adaptive dialogue [35]. Theories suggest that conversational agents can be seen as educational partners and offer opportunities for social interaction, which is known to aid learning [28, 42, 48, 53, 54]. Researchers who emphasize learning as a social endeavor have highlighted the importance of rapport between tutor and student, which helps maintain a student's sense of approval and autonomy when receiving feedback [46, 47, 59, 69, 79]. Of particular relevance are studies that use virtual agents to build rapport and that enhance learning efforts and performance in mathematics [40]. Further, because natural language dialogue is highly engaging and allows learners to actively construct knowledge when formulating responses, researchers postulate that dialogue-based systems can promote greater engagement, learning gains, and retention [35, 38].

Preliminary results of studies that have implemented educational chatbots suggest they may be beneficial in supporting learning (e.g., math skills and engagement) outside of traditional classroom settings [25, 64]. However, given concerns about their cost-effectiveness [15, 67], more research is needed to evaluate conversational agents' efficacy, including with respect to the learning environment in which they are embedded, the learning domain, and learners' and agents' characteristics.

Educational Games

While we categorize our tool as a narrative-centered learning technology, many of our design decisions were inspired by research in educational games, given prior work that indicates game-like approaches could offer educational efficacy [28, 49]. Decades of research from the learning sciences generally agrees on a few key conditions that promote learning: engagement with learning materials, meaningful experiences related to our lives, social interactions, and clear learning goals [28]. Considering educational games with respect to these ideas, the play experience can be seen as active and engaging, and game elements like narratives and interactive characters present learning opportunities.

Empirically, Mayer [49] summarized studies that compared learning outcomes from playing a game versus conventional media, concluding that game-play leads to better learning outcomes than conventional media and is especially promising for topics such as science and second-language learning. However, three out of five of these studies that involved mathematics learning found only a small effect size. Similarly, Young et al.'s meta-review of over 300 video games and academic achievement related articles fails to find strong support for video games' academic value in math and science [77].

Drawing together these ideas about the power of narrative scaffolding, the importance of feedback, and the need for more design and experimental research aimed at applying and rigorously testing the impacts of such approaches, we were motivated to develop our learning platform, described next.

THE LEARNING PLATFORM

Our system has three main components: narratives, learning activities, and feedback support. The goal is to leverage these rich interactions, as opposed to plain drill and practice, to help children build understanding of new concepts.

Fantasy-Based Narratives

The platform uses a fantasy-based narrative in which a child assumes the role of the main character. Accompanied by a monster of his or her choice (selected upon setup), the child and the monster embark on a heroic journey to save the world from various natural disasters (see Figure 1).

The narrative was created using common narrative archetypes, as well as the overarching narrative structure of a fairy tale [63]: the child takes on the role of the protagonist, the monster is the deuteragonist, and the natural disasters function as the primary source of conflict within the narrative arc. By obtaining a magical element — a special gem at the heart of each landscape — the monster and child are able to undo the environmental devastation. The plot is advanced by the child's interactions with learning activities within the story; completing these activities allows the child to progress further into the narrative and eventually reach the magical gem. For example, in Chapter 1 (used in the experimental condition of our study), the child and monster need to convince a boat captain that it is safe to let them cross a river with a box of chocolates, which requires the child to calculate the weight of the box, as illustrated in Figure 1d. Additionally, in one variation of the system as shown in Figure 2d, the monster assumes the role of the chatbot during these learning activities, allowing the child to receive help and advice from a perceived peer.

We also personalize the narrative in several ways, given personalized instruction and fantasy play can boost engagement and learning [6, 73]. First, to allow the child to role-play as the main character, the system requests and inserts his or her name into the story, which is told in the third person. The introduction also presents a monster customization page from which the child can give the monster a name and select its appearance and interests (see Figure 1b). Finally, the interface background (e.g., the sky – see Figure 1) adapts to reflect the time of day and weather at the child's location.

To craft the story and ensure it would be compelling and understandable for children of the targeted age range (grades 3–5), we undertook rounds of design and testing with two educational practitioners, a narratologist, and pilot users ($N = 14$), using their feedback and ratings to refine the narrative's text and visuals, including the monster's look and feel.

Math Learning Activity

The narratives embed a learning task that requires users to apply math concepts to complete a relatable challenge. We designed the task as a multi-step math problem based off Common Core Mathematics Standards [7]. The present study's educational goals were centered around learning the concept of volume, its knowledge components (measurement and multiplication), and the concept of fractions.

For the task, learners were given a piece of chocolate and a cardboard box and asked to estimate the weight of the box if it were full of chocolates weighing $\frac{1}{2}$ oz each. The problem was broken down into six smaller steps. Learners first need to answer how many chocolates can fit along each edge of the box, which requires grade 1 level math knowledge in measurement and data (1.MD.A.2). They were then asked how many chocolates the box can hold given their measurement, which requires understanding volume (5.MD.C.3, 5.MD.C.4, 5.MD.C.5) and the application of multi-digit arithmetic (4.NBT.B.5). Finally, to calculate how much the box of chocolates weighs, learners must apply fractions as numerical magnitudes (3.NF.A) and multiply a fraction by a whole number (4.NF.B.4). Table 1 summarizes these steps and the corresponding knowledge components. To proceed at each step, learners must provide a correct response. The task was designed to be challenging for learners at or below grade 5, who typically lack mastery in these concepts, to let us examine the effectiveness of the feedback support. The activity does not require prior knowledge of volume, but it does not provide teaching support for students that lack prerequisite multiplication and addition skills. These skills are assessed in the pre and post test.

Table 1: Components of each step of the math learning task evaluated with users. Number of hints indicates the maximum number of hints students could obtain in conditions C and D through the hint system or the chatbot, respectively.

Steps	*Step 1-3*	*Step 4*	*Step 5*	*Step 6*
Problems to Solve	How many chocolates fit along the height/length/width of the box?	Given the measurement of height, length, and width, how many chocolates are in the box?	If each chocolate weighs half an ounce, how much does the box of chocolates weigh?	Given that the boat can hold at most 320 ounces, can the user safely board the boat?
Knowledge Components	Measurement & Data	Measurement & Data Operations & Algebraic Thinking	Number & Operations	Number & Operations in Base Ten
Quizzing Concepts	NA	Volume & Multiplication	Fractions	NA
Grade Levels	Grade 1	Grade 4 & Grade 5	Grade 3 & Grade 4	Grade 2
Common Core Standards	1.MD.A.2	5.MD.C.3, 5.MD.C.4, 5.MD.C.5, 4.NBT.B.5	3.NF.A.1, 4.NF.B.4	2.NBT.A.4
Number of Hints	4 hints	8 hints	8 hints	2 hints

Feedback Support

We designed and implemented two methods to provide learners with clues when stuck. The first is a hint system, shown in Figure 2c, with each preset hint delivering an increasing amount of instruction. This is similar to hint implementations in prominent learning platforms [1, 30, 32]. The second method is a chatbot, shown in Figure 2d, presenting the user's chosen monster as an adventure partner. Users can type or speak to the chatbot, which responds with personalized clues using wizard-of-oz. To facilitate comparison between the hint system and chatbot, both provide learners with the same set of clues, which were designed by the educational practitioners who helped refine the narrative. Table 1 summarizes the number of hints available at each step of the task.

The hint system and chatbot differ in three main ways. First, the chatbot is designed to be polite [37, 74] and friendly to build rapport with users [46, 47, 59, 69, 79]. It also uses a conversational style that matches the adventure monster's personality, and it phrases hints with familiar natural language. Next, the chatbot can engage learners in conversations not directly related to the task to promote social interaction [28, 42, 48, 53, 54] (see wizard's protocol below). Lastly, while the hint system delivers clues in a preset sequence, the norm in most systems, the chatbot can assess a user's understanding and deliver hints in an order to optimize assistance [41].

Variations

We designed, implemented, and tested three variants of our narrative-centered learning platform (the task, plus: narrative alone, narrative with hints, narrative with chatbot) to compare these feature combinations against a control (the task alone).

A: Task

To ensure a fair comparison, the control system (Figure 2a) offered the same learning task as the narrative variants and the same aesthetic design, color scheme, and textual instructions.

B: Task and Narratives

As shown in Figure 2b, System B embedded the learning task into the narrative and introduced personalization with the customizable monster partner, who served as the deuteragonist in the story. The narratives included elements such as illustrations and maps to make them visually appealing.

C: Task, Narratives, and Hints

System C included all the elements of system B plus an expandable window containing step-by-step hints, as shown in Figure 2c. Learners could tap "I Need a Hint" to see the first hint for a given step and "I Need Another Hint" for additional hints. Table 1 indicates the number of possible hints for each task step. As mentioned, the hint system was designed to reflect current practice in the field [1]. Similar to Khan Academy, as users ask for more hints for a certain step, the hints become increasingly elaborate, with the last hint revealing the final answer. To prevent users from abusing the hint system (i.e., continuously pressing on "I Need Another Hint" to reach the answer), this button disables for 10 seconds after clicking.

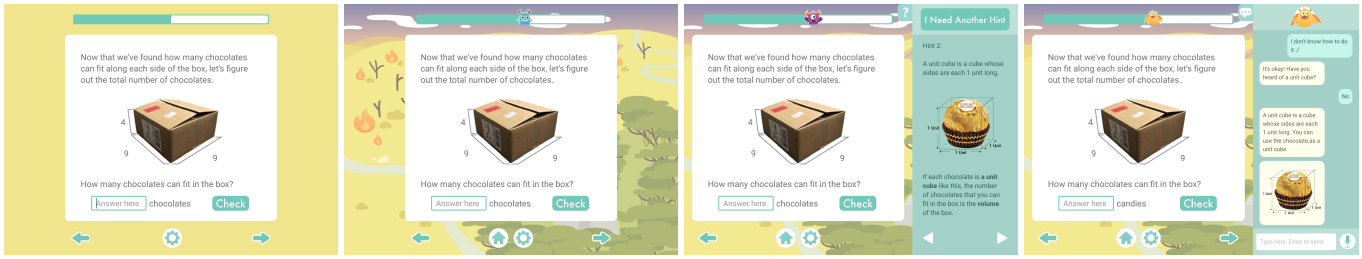
D: Task, Narratives, and Chatbots

System D also included all the elements of system B, plus the chatbot system, as shown in Figure 2d. The chatbot provided the same scope of assistance as the hint system but responded in a friendly, conversational style. The child could tap on the chat icon to expand the chat window and then type or use a speech-to-text button to send messages to the chatbot.

Wizard Protocol. The chatbot was implemented via wizard-of-oz techniques. The same person always played the wizard to ensure consistency in behavior and conversation style. The wizard assumed a friendly, encouraging, and patient personality throughout all interactions. Figure 3 shows the UI for the wizard, who was allowed to perform four types of actions:

- Small talk: To engage users, the wizard could use non-task-related casual conversations such as sending greetings or jokes (e.g., "I would tell you a joke about dragons, but I'm afraid they tend to drag-on for too long.") Prior work indicates such dialogue can enhance user engagement [78] and has found that messaging improves learning outcomes when presented in a conversational style [19, 53, 54]. The rationale is that learners may work harder when they feel they are working together with a social partner [48]; and children may develop parasocial relationships with on-screen characters, which can bring educational benefits [28].

- Encouraging users: If a user remained at the same step for an extended period (one minute), the wizard would provide cheerful encouragement (e.g., "You're doing awesome!



(a) Control system: task only (b) Task + Narrative (c) Task + Narrative + Hints (d) Task + Narrative + Chatbot

Figure 2: Four variants of our learning platform: (a) a control system that delivers the learning task with no narrative, hint, or chatbot features; (b) the same learning task embedded in narratives; (c) the same learning task and narratives together with a hint system; (d) the same learning task and narratives together with a tutoring chatbot implemented using wizard-of-oz techniques.

You're almost there; just keep trying!") The aim was to both add a personal touch to the interaction as well as cultivate a growth mindset [21] by praising effort and motivating perseverance. Along with motivational messages, the wizard would sometimes include parasocial displays (e.g., animated images of the chatbot monster looking joyous) to help bring users' attention to the app and increase engagement [28].

- Providing hints: The wizard provided the user with hints specific to each step. Again, while the instructional content was the same as system C's hint system, the wizard worded it more conversationally. Unlike in system C where learners must progress in a specific order through hints (which provide increasing scaffolding and ultimately reveal the answer), the wizard provided the hint she believed would be most appropriate at that time. For example, if it seemed the user could benefit from a reminder to first apply the concept of volume, the wizard could hint, "Think of volume as the number of chocolates that you can fit in the box. Does this ring a bell?" Further, the wizard was not allowed to give the user the final answer when asked for it. Instead, the wizard responded with encouragement to keep trying, such as "Alas, I am but a simple Cyclops. Let's think this through together. What do you think we should do next?"
- Checking understanding: To provide support, the wizard sometimes asked probing questions such as "Have you heard of a unit cube?" to offer implicit hints while gauging understanding of concepts underlying the task.

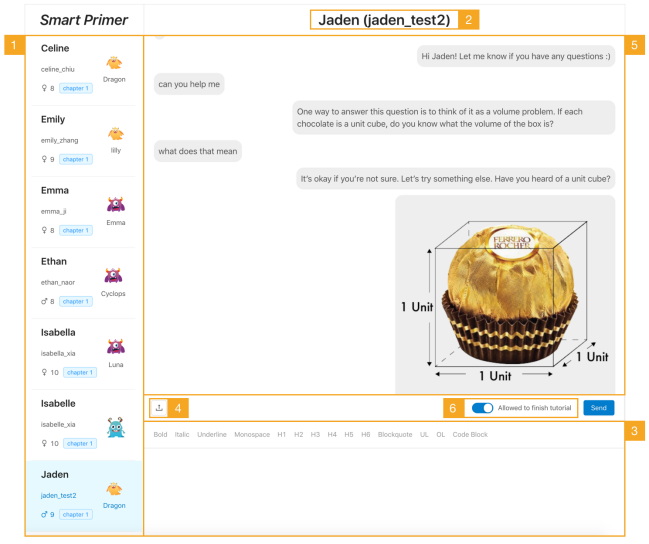


Figure 3: The wizard's dashboard UI: (1) The wizard selects a user to send messages. (2) The selected user. (3) The wizard inputs messages into the dialog box and adds styles. (4) The wizard can send uploaded pictures to the user. (5) The conversation between the user and the wizard. (6) The wizard gives new users a short tutorial on how to interact with the chatbot.

Implementation

We developed our platform using the Android framework and TypeScript [52]. Specifically, we implemented the user interface as an Android tablet application. We built the wizard's interface using React [23] and deployed it on Netlify. To serve the static images required by the wizard, we instantiated an Amazon Simple Storage Service (S3) instance. Additionally, we built the backend service using Hasura GraphQL Engine [26] and deployed the service onto Heroku with a Postgres [62] database. To provide children with responsive and configurable feedback, we used GraphQL's subscription features.

EVALUATION

We performed a between-subjects lab study to evaluate the three variants of the learning platform compared to the control. Despite mixed evidence [2, 19, 51, 76], we hypothesize that conditions involving narrative will yield better engagement scores than the non-narrative control, given the consensus that a narrative context promotes motivation and enhances experience [24, 44, 50, 51, 66, 72]. However, we suspect that narrative without feedback, an essential part of educational instruction [27, 34, 68], will be limited in the learning gains it can achieve. We therefore hypothesize that integrating the motivating, relatable properties of narrative with the instructional value of feedback systems will lead to improved engagement scores as well as yield higher learning gains. Altogether, we therefore hypothesize that, compared to the control:

H1. Narrative will improve engagement.
H2. Narrative will not improve learning.
H3. Narrative and feedback will improve engagement.
H4. Narrative and feedback will improve learning.

Study Procedure

Our evaluation included a lab study and follow-up assessment.

In-Lab Study

The in-lab user studies were held in university labs. Participants took a pre-study survey that included a math anxiety scale survey [13] followed by a pre-study math quiz, after which they were randomly assigned to one of the four system variants. The child used the learning system alone in a quiet room, as shown in Figure 4a, and a camera was set up to record the entire duration of the child's interaction. The tablet interface was also recorded for later review. An observer sat in the same room as the child to take notes and answer any questions the child might have about the interface (but not the learning task). The child could opt out at any time; we marked the study as "unfinished" in this case. Regardless of whether participants solved the task, they took post-study math and engagement assessments, as described in the next section.

Follow-Up Assessment

One month after participation in the study, we asked each child to perform the same learning task administered in the lab study, except virtually via a website we built. Numerical values in the task were also altered to avoid answer recall. The child was asked to answer these follow-up questions at home without help from external resources including their parents.

Measures and Instruments

We were interested in the effects of the various learning experiences on two main metrics: learning and engagement.

Assessing Learning Effects

To assess learning outcomes, participants took a math assessment (quiz) immediately before and after their interaction with our platform. The quiz contained three questions for each knowledge component: *fractions*, *multiplication*, and *volume*, for a total of nine questions that were a mix of fill-in-the-blank and multiple-choice. All questions were drawn from Khan Academy [1] and Singapore Math Books [31], with selection guided by the Common Core State Standards [7] and examined by educational experts to ensure appropriateness. To limit learning effects from repeated measures [33], we utilized two versions of the quiz (*a* and *b*), where the only difference was that numerical values in the problems were changed and we counterbalanced administration such that half of participants took assessment *a* pre-study and assessment *b* post-study, with the other half taking them in the opposite order. The assessment items were chosen to involve the same skills as the learning task in our designed system, but were not isomorphic. This was done to assess learner's ability to do related tasks after completing the learning activity, relative to their performance on similar tasks during the pre-assessment.

Following guidelines on evaluating computer-based learning tools [49], we also measured delayed retention (the extent to which essential information is remembered) in the follow-up activity after one month and transfer (ability to apply gained knowledge and skills to solve new problems).

We piloted the math quiz with nine children in grades 3–6. Average scores for those in grades 3, 4, 5, and 6 were 4.0, 6.0, 7.0, and 8.8 out of 9, respectively. In particular, $\frac{3}{4}$ of 6th graders achieved perfect scores — they had already mastered the knowledge components relevant to our study. This motivated us to scope our study to children in grades 3–5, who exhibited knowledge gaps in the concepts our system covered.

Assessing Engagement

After the post-study quiz, participants completed a survey with a short form of the User Engagement Scale, which measures engagement in digital domains [56]. For children in conditions B, C, and D (i.e., the narrative conditions), their post-study survey also included the Narrative Engagement Scale (NES) [12], which assesses engagement across five dimensions: cognitive accessibility, empathy, involvement, perspective taking, and realism. After completing the lab study, participants who used system A (task only, no narrative) read the narrative (which is 1130 words long) and then completed the NES. To enable us to calibrate participants' reactions to our story relative to popular children's novels of variable lengths, these participants also read and rated two additional award-winning stories: *Frindle* [17] and *The Number Devil* [22]. We provided participants with the first chapter of *Frindle* (850 words). *The Number Devil* is a novel focusing on mathematical concepts; we provided participants with the first chapter (2580 words). All surveys to measure engagement were conducted as interactive interviews between the observer and the child to ensure questions were well explained and understood.

To obtain an objective measure of engagement, we coded all except two participants' videos (who asked not to be recorded) using a common protocol for evaluating educational software [8, 20, 65]. We segmented every video into 20 second segments, and two research team members who did not know to which condition the child was assigned independently coded each segment as one of seven possible emotions: *engagement*, *boredom*, *confusion*, *curiosity*, *happiness*, *frustration*, or *neutral*. This was inspired by the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) [57]. Inter-rater reliability (Cohen's kappa) was $\kappa = 0.55$, which is close to the default 0.6 cut-off. The slightly lower agreement is not unexpected, given that kappa is known to be lower when observing affect as opposed to behavior and that coders watched video rather than children in-situ. Our agreement falls within the broader acceptable range of 0.4–0.8 for this type of analysis [57].

Participants

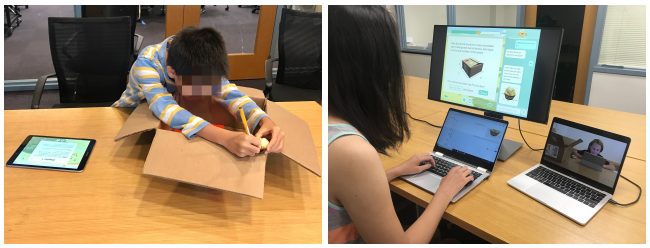
We recruited $N = 72$ children through NextDoor, university mailing lists, and word of mouth. They attended 43 different schools. 57% were Asian, 22% White, 14% reported two or more races, and 7% did not disclose. 32 children were in 3rd grade, 24 in 4th grade, and 16 in 5th grade. Half of participants (36) were boys and the other half girls. We randomly divided these 72 children into four groups such that the grades and genders were equally distributed: i.e., for each group, 8 participants were 3rd graders, 6 were 4th graders, and 4 were 5th graders, with 9 boys and 9 girls in each group.

Apparatus

Our platform was deployed to a Samsung Galaxy Tab S3 with an external keyboard. Experimental items included a card-

board box, Ferrero Rocher chocolate, pencil, and scratch paper (shown in Figure 4a), which were placed next to the child to enable use at any time during the study. These items were provided to mimic an informal learning environment where learners can use available objects to help solve problems.

The setup for the wizard can be seen in Figure 4b. The wizard's dashboard as shown in Figure 3 was opened on the wizard's own laptop. The Galaxy screen was live streamed to the large monitor in Figure 4b via Android Debug Bridge [71]. The wizard observed the child in real-time through a Zoom conference call between the second laptop in Figure 4b and a phone camera located in the same room as the child.



(a) Setup for the child (b) Setup for the wizard

Figure 4: The setup for the child (a) and the wizard (b).

RESULTS

In this section, we present results of the study and interpret findings with respect to our previously presented hypotheses. Results from math and engagement assessments across the four conditions are summarized in Table 2. We use the standard mean (standard deviation) notation.

Impacts on Learning

First, we examine the immediate and sustained learning effects of the platform variants.

Math Performance Before and After Using the System

Math improvement was computed by subtracting pre-study quiz scores from post-study quiz scores. Quizzes contained 9 questions, each worth 11.1 points for a total of 100 points per quiz. Figure 5a and Table 2 summarize participants' total and subject-wise improvements. Conditions exerted a significant effect on improvement in volume concepts ($F_{3,68} = 3.239, p < .05$); but we found no difference in fraction improvement ($p = .265$), multiplication improvement ($p = .951$), or total improvement ($p = .209$). Post-hoc multiple comparisons corrected by Holm's sequential Bonferroni procedure showed that improvement on volume between A (task-only control) and D (task+narrative+chatbot) was statistically significant ($W = 93.5, p < .05$) with an effect size of 0.88 but not between A and B (task+narrative) ($W = 100.5, p = .080$) or A and C (task+narrative+hints) ($W = 134, p = .159$).

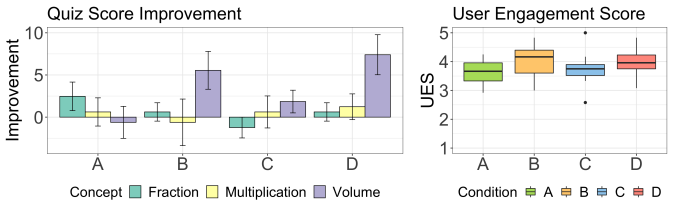
Thus only the condition with narratives and a chatbot had a positive effect on children's learning outcomes, indicating that simply supplementing a narrative with any feedback is not enough. Rather, that feedback's delivery makes a difference for learning. Further, we did not find a statistically significant different across conditions on total time spent ($F_{3,68} = .709, p = .173$) or task time spent ($F_{3,68} = 2.338, p = .081$), indicating that embedding the math task into a narrative did not make the learning activity significantly longer.

Examining the Learning Task Step-by-Step

Our learning task consisted of six steps, as shown in Table 1. Participants could opt out at any step. Table 2 presents the results of children who solved the task (finished all six steps) versus those who did not. Among solvers and non-solvers, there was no significant difference in user engagement ($t_{70} = -.565, p = .574$) as determined by an unpaired two-samples t-test, nor on improvements for fractions ($W = 675.5, p = .095$), multiplication ($W = 529, p = .5365$), volume ($W = 462.5, p = .145$), and total ($W = 493.5, p = .306$), as determined by Mann Whitney tests. Not surprisingly, a Mann Whitney test showed that children who solved the task were older than those who could not solve it ($W = 343.5, p < .01$). Further, we were interested in children who independently solved the task in conditions A and B (i.e., did not receive any feedback assistance). We found no differences for these children in user engagement ($p = .404$) nor differences in improvement for fractions ($p = .254$), multiplication ($p = .693$), volume ($p = .199$), or total ($p = .925$).

In addition, solvers and non-solvers spent an equal amount of time using the learning system ($W = 706.5, p = .120$) but a different amount of time on the learning task ($W = 756, p < .05$). The average amount of time children spent on each step of the learning task is plotted in Figure 6a, which shows that different steps required different amounts of time and that this trend was similar across different conditions. Moreover, the trend was consistent with the difficulty of underlying knowledge components and suggested grade levels, as shown in Table 1.

To track when participants got stuck, unstuck, or opted out, we designed the system so that users had to solve the current step to reach the next step. Figure 6b presents the percentage of children who successfully solved each step. As we can see, children found the fourth step most difficult and the fifth step the next most difficult. This was also consistent with the grade levels we designed around (see Table 1). The trend when hints were not available (conditions A & B) was similar: about one third of children eventually solved the task. The trend when hints were available (conditions C & D) was also similar: almost all the children solved the task with feedback offered through either the hint system or the chatbot.

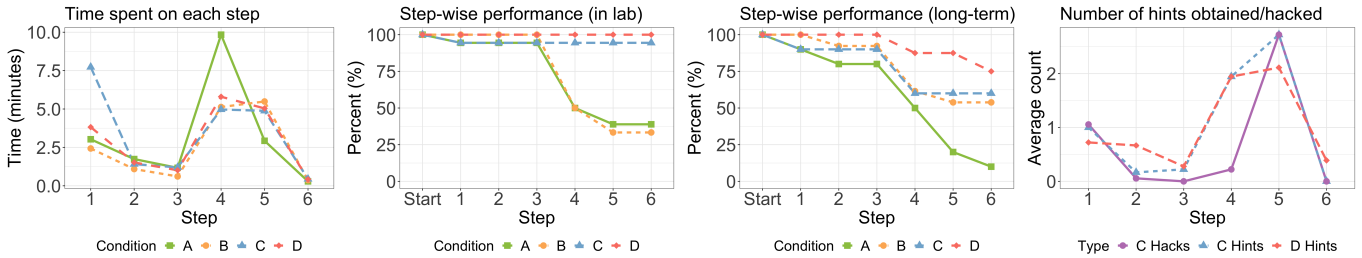


(a) Improvement plot (b) UES plot

Figure 5: (a) Bar plot of improvement (points) on each math concept after using the learning system. (b) Box plot of user engagement scores. Error bars represent +/-1 standard error.

Table 2: A = Task, B = Task + Narrative, C = Task + Narrative + Hints, D = Task + Narrative + Chatbot. Solved / Total represents number of children who solved the task out of total number of children in that condition. Total time and task time in minutes. Total improvements graded out of 100 points, and subject-level improvements graded out of 33.3 points. Dagger sign indicates statistically significant difference.

Condition	Solved / Total	Age	Math Anxiety	Total Time	Task Time	User Engagement	Total Improvement	Multiplication Improvement	Volume Improvement	Fraction Improvement
System A	7 / 18	8.7 (1.0)	15.7 (3.6)	17.9 (13.4)	17.9 (13.4)	3.62 (0.41)$^{\dagger 12}$	2.47 (11.14)	0.62 (7.09)	-0.62 (8.05)$^{\dagger 3}$	2.47 (7.18)
System B	6 / 18	8.7 (0.9)	14.2 (4.9)	20.4 (7.0)	13.0 (6.8)	4.02 (0.52)$^{\dagger 1}$	5.55 (13.33)	-0.62 (11.72)	5.55 (9.52)	0.62 (4.62)
System C	17 / 18	8.7 (0.8)	15.3 (4.1)	27.1 (17.9)	22.5 (16.5)	3.72 (0.48)	1.23 (9.24)	0.62 (8.05)	1.85 (5.71)	-1.23 (5.23)
System D	18 / 18	8.8 (0.8)	14.4 (4.2)	24.4 (11.6)	17.5 (12.1)	4.01 (0.45)$^{\dagger 2}$	9.25 (14.37)	1.23 (6.47)	7.40 (10.07)$^{\dagger 3}$	0.62 (4.62)
Grade 3	17 / 32	7.9 (0.5)$^{\dagger 4}$	15.0 (4.2)	28.9 (11.9)$^{\dagger 5}$	23.8 (12.5)$^{\dagger 6}$	3.85 (0.47)	3.12 (13.00)	0.00 (7.97)	1.73 (8.96)	1.39 (6.14)
Grade 4	15 / 24	8.9 (0.3)$^{\dagger 4}$	14.8 (3.9)	21.1 (14.1)$^{\dagger 5}$	16.6 (12.9)$^{\dagger 6}$	3.80 (0.61)	6.01 (13.08)	0.00 (9.26)	5.55 (9.26)	0.46 (6.11)
Grade 5	16 / 16	9.9 (0.2)$^{\dagger 4}$	14.9 (4.9)	11.7 (4.9)$^{\dagger 5}$	7.4 (3.2)$^{\dagger 6}$	3.90 (0.30)	5.55 (9.93)	2.08 (8.33)	4.16 (7.98)	-0.69 (2.78)
Unsolved	0 / 24	8.3 (0.8)$^{\dagger 7}$	15.0 (4.0)	25.3 (14.1)	21.2 (13.9)$^{\dagger 8}$	3.80 (0.53)	2.78 (13.19)	-0.46 (9.53)	0.93 (9.21)	2.31 (6.53)
Solved	48 / 48	8.9 (0.8)$^{\dagger 7}$	14.9 (4.3)	21.0 (12.8)	16.0 (12.1)$^{\dagger 8}$	3.87 (0.47)	5.55 (11.90)	0.93 (7.88)	4.86 (8.54)	-0.23 (4.85)



(a) Step time (b) Steps solved (c) Steps solved one month later (d) Step hints

Figure 6: (a) Average amount of time children in each condition spent on each step. (b) Percentage of children in each condition who solved each step of the learning task. Children who solved step 6 were regarded as having successfully solved the learning task. (c) Percentage of children in each condition who solved each step of the virtual learning task presented to them one month later. (d) Average number of hints / hacks children in conditions C & D obtained.

Narratives and Chatbots Improve Retention

To test retention, we sent out a follow-up test to all participants one month after their lab study. 41 of 72 participants completed the test. Their performance is depicted in Figure 6c. We can see that children in the narrative conditions (B, C, & D) retained what they learned during the lab study better than children who did the task without narratives (A), we imagine because the context provided by the narrative helps cement concepts. Additionally, children who interacted with a chatbot had an even higher retention: 75% of condition D participants successfully solved the same learning task one month later, demonstrating the sustained effect of chatbots as well.

Engagement, Emotions, and Overall Experience

Engagement with the System

According to responses to the User Engagement Scale (UES) [56] (see Figure 5b), which assessed engagement with the system in terms of aesthetic appeal, feelings of focus, perceived usability, and rewarding experience, a statistically significant difference can be observed across the four conditions as determined by one-way ANOVA ($F_{3,68} = 3.409, p < .05$). A Shapiro-Wilk test on the ANOVA residuals show that residuals were normally distributed, indicating the appropriateness of our models ($W = 0.991, p = .872$). According to a Shapiro-Wilk normality test, the normality was assumed for the user engagement scores in conditions A, B, and D but not C. We therefore performed an unpaired two-samples Wilcoxon test comparing UES between A and C and an unpaired two-samples t-test comparing UES between A and B as well as A and D. We corrected the post-hoc multiple comparison results with Holm's sequential Bonferroni procedure [29]. Results showed a statistically significant difference in engagement between A and B ($t_{34} = -2.568, p < .05$) with an effect size of 0.86 and also between A and D ($t_{34} = -2.702, p < .05$) with an effect size of 0.90 but not between A and C ($W = 147.5, p = .657$). Overall, embedding a learning task into narratives thus improved engagement as we hypothesized. However, when we supplemented the narratives with hints, engagement dropped, contradictory to H3. Yet combining narratives with a chatbot increased engagement, again supporting our prior finding that when delivering feedback, the format matters.

Engagement with the Narrative

Next we aimed to understand how engaging children found our narratives. On a 1-7 scale, ratings of our narrative, *Frindle*, and *The Number Devil* were 4.50 ($SD = 1.18$), 5.04 ($SD = 0.65$), and 4.40 ($SD = 1.38$), respectively. We did not find a statistically significant difference between participants' ratings of the three stories as determined by a one-way repeated measures ANOVA ($F_{2,22} = 2.544, p = .101$).

Impact on Affective States

Table 3 shows the proportion of time users spent in each affective state throughout the study, according to our coding of the video data. The most frequently observed emotions in all conditions are engagement followed by confusion, which is consistent with prior work [20, 65]. A one-way ANOVA showed that the proportion of the engagement state was statistically different across conditions ($F_{3,66} = 2.944, p < .05$). Post-hoc pair-wise tests adjusted by Holm's Bonferroni Sequential Procedure [29] showed there was a significant difference between the proportion of engagement between children in condition A and D ($p < .05$). Engagement analysis derived from the video coding is therefore consistent with user engagement based on the UES, with both showing a statistically significant difference of engagement between conditions A and D. Condition also exerted a significant effect on the confusion state ($F_{3,66} = 2.950, p < .05$), and the difference between A and D was again statistically significant as examined by post-hoc pair-wise tests adjusted by Holm's Bonferroni Sequential Procedure ($p < .05$). Altogether, these findings indicate that the inclusion of a narrative along with a chatbot promote more engagement and less confusion during the learning experience.

Affective State	A	B	C	D
engagement	76.1 (19.9)	83.1 (12.5)	85.4 (13.7)	**90.0 (7.3)**
boredom	**4.8 (8.2)**	4.6 (6.2)	3.4 (4.8)	1.0 (2.1)
confusion	**13.7 (12.5)**	8.8 (6.2)	9.2 (8.7)	5.0 (5.2)
curiosity	0.3 (1.1)	0.3 (0.5)	0.58 (1.1)	**1.1 (2.1)**
happiness	1.3 (3.4)	1.2 (1.9)	0.7 (1.5)	**2.7 (2.9)**
frustration	**3.0 (4.9)**	1.9 (3.2)	0.7 (1.4)	0.2 (0.6)
neutral	**0.9 (2.5)**	0.2 (0.7)	0.0 (0.0)	0.0 (0.0)

Table 3: Percentage (%) of time participants spent in each affective state. Largest percent for each state in bold.

Math Anxiety

Because participants' math anxiety scores did not follow a normal distribution ($W = 0.940, p < .01$), we conducted Kendall rank correlation tests. Results showed that children who were less anxious about math had higher engagement ($r_\tau = -2.197, p < .05$). No correlations were observed between math anxiety and improvement, suggesting initial anxiety affected engagement levels but not learning outcomes.

Unpacking Interactions

Hint System Versus Chatbot

Conditions C and D both offered assistance on the learning task, but children receiving feedback from the chatbot exhibited more engagement and learning. Examining the difference between these two feedback modes, we classified every click made when using the hint system and manually labeled every message sent by the wizard and users in condition D.

Figure 6d presents the number of hints users obtained or hacked at each step. That is, we define *hint hacking* as any clicks on the "I Need Another Hint" button during the 10s enforced waiting period between hints. The 18 participants in condition C had in total 165 clicks on the hint button, with 56 (33.9%) of them hint hacking, indicating that about one third of the time, children kept clicking on the hint button without reading the content in the hints carefully.

Table 4: Classification of messages sent by the wizard averaged across the 18 participants in condition D.

	Small Talk	*Encouraging*	*Checking Understanding*	*Providing Hints*
Messages	11.7 (3.6)	6.4 (5.7)	6.2 (6.7)	16.1 (16.2)
Percentage	29.0%	16.0%	15.3%	39.8%

The trend of hints obtained in condition C and D are similar, showing children got a similar amount of help in the two conditions. However, the distribution of hints they received at each step was quite different, as illustrated in Figure 7. As shown in Table 1, every step offered a different number of hints. Children in condition C needed to go through the hints in an linear order, as this is the standard design in modern hint systems [1]. However, children in condition D could get hints in any order. In particular, since the wizard had the ability to check a child's understanding, it sent the most appropriate hint for the child during conversations. Figure 7 shows the average number of hints a child received per step and confirms that children in condition D received hints in a non-linear order.

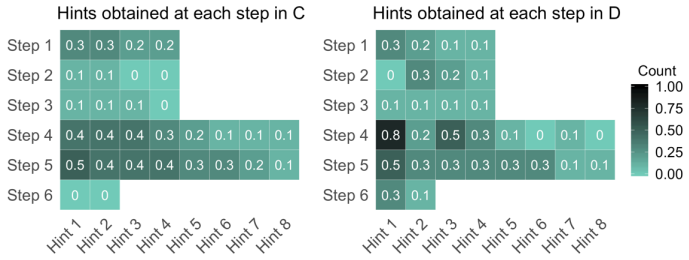


Figure 7: Average number of hints at each step obtained by children in condition C (left) and condition D (right).

By examining the conversation logs, we found that the interaction between the wizard and the 18 children in condition D comprised a total of 927 messages, or an average of 51.5 messages per child. On average, a child received 40.4 ($SD = 27.5$) messages from the wizard and sent 11.1 ($SD = 10.8$) messages. 200 (21.6%) messages were sent by participants and the remaining 727 (78.4%) were sent by the wizard. As Table 4 shows, only 55.5% of messages were tutoring; the wizard spent the rest of the time building the relationship with the child (*small talk*) and encouraging them (*encouragement*).

Children's Preferred Modes of Interaction with the Chatbot

As a reminder, the chatbot supported both typed and spoken conversation. Of the total 11.1 ($SD = 10.8$) messages children sent, an average of 2.3 ($SD = 5.1$) were entered through speech recognition, accounting for 20.7% of messages sent. 15 out of 18 (83.3%) children expressed that they liked typing much better. They commented that "I wanted to type. I don't like to speak." and "it [typing] feels better than speaking, and sometimes when you're speaking it just does different words." Two people (11.1%) preferred typing and speaking about the same. Only one person (5.6%) preferred speaking since "it's hard to type on the tablet."

Perceptions of the chatbot were generally positive. The most common adjectives children used to describe the chatbot were *happy*, *friendly*, *funny*, *helpful*, and *respectful*. One child commented that "he [the chatbot] wants to chat a lot - friendly, fun to play with", and another child mentioned "he [the chatbot] is a smart partner and a nice one". Eleven children (61.1%) thought they were talking to a robot because "humans can program the chatbot" and "if you type random letters, it [chatbot] would be exactly the same; he didn't acknowledge my response and just said what he would say." Four children (22.2%) believed the chatbot was played by a human.

DISCUSSION

Our study offers implications for members of the interaction design community interested in leveraging the power of stories and conversational agents to create engaging experiences for educational applications. In particular, by integrating narrative with conversational feedback to support children's math learning, we demonstrated the promise of applying features in conjunction, rather than in isolation. We found consistent evidence from the User Engagement Scale and video coding that this integrated approach could make learning more engaging, with pre-post math assessments and retention tests additionally indicating the educational efficacy of the approach in both the short-term and long-term. In contrast, we did not observe as strong boosts in interest or knowledge from a hint system based on conventional designs. In fact, we were surprised to find the conventional hint system design negatively impacted user engagement and may have offset the benefits of narrative.

The discrepancies observed between the hint and chatbot conditions support a proposal that learning is a social endeavor. We saw that feedback that is too abrupt, direct, and lacking in interpersonal connection can negatively impact learning, while a supportive chatbot that builds rapport with learners leads to better experiential and educational outcomes, similar to prior findings on other non-narrative settings [47, 59, 69]. It is possible that the fantasy world created by the narrative offered an approachable learning environment and a mechanism for learner-agent bonding. The politeness incorporated into agent responses likely also contributed to learning gains observed in the chatbot condition, which aligns with prior work [37, 74]. In addition, studies have shown that learners do not always have the meta-cognition to know when to ask for help [3, 4, 5]. It is possible that the social interactions and discourse between learners and the chatbot helped them identify their confusions and, in turn, promoted question-asking that led to improved engagement and learning. Overall, we identified advantages to a chatbot approach given their ability to iteratively and naturally gauge students' understanding [41] by holding casual conversations [78], delivering hints out of order [9], and interacting with learners socially.

We also observed that students did not improve on learning fractions or multiplication as much as they did on volume. We suspect this may be because more opportunities to learn the concept of volume were available compared to multiplication. Many participants obtained hints on volume; however, multiplication was only explicitly taught in hint 7 in step 4, which few children obtained (Figure 7). This cannot explain the low learning gain observed in fractions though, as learners heavily obtained fraction hints. It could be that mastering this knowledge component requires abstract thinking that was not adequately conveyed by our designed hints. Also, our learning task was centered around measuring the volume of a box using a piece of chocolate as a unit cube, and participants were able to manipulate the box and the chocolate to solve the problem. It is likely that both the hint's instructional value [27, 34, 68] and the manipulable objects [55] that served as concrete examples [39] contributed to learning volume.

Lastly, another somewhat surprising finding is that supplementing narratives with a conventional hint system did not improve learning outcomes or engagement. In condition C, students were given hints passively, whereas in condition D, students interacted with the chatbot to actively solve the learning task. Based on the ICAP framework [14], we could consider students' behaviors in condition C as *passive/attentive* and in D as *interactive/collaborative*. According to this framework, when students' engagement increases from *passive* to *active* to *constructive* to *interactive*, their learning also increases — which is consistent with and helps to explain our results.

Limitations and Future Work

In this study, opportunities for interaction and practice were limited to one session focusing on volume. It will be desirable for future work to investigate additional concepts to better assess the extent of learning gains and the generalizability of results, including for subject matter from other domains such as science, reading, writing, or the arts. A number of other directions also stand out to continue the design, deployment, and evaluation of narrative-centered learning systems augmented with contextualized feedback. For example, one clear next step is to implement and test a functional chatbot using natural language processing (NLP) techniques.

Further, it is important to conduct longitudinal studies with additional chapters and content, both to examine the extent to which our observed effects were due to the novelty of the technology as well as to go beyond optimizing for task "performance" and instead design scaffolds for "learning" that may encourage or even embrace failure as part of promoting more permanent changes in comprehension and skills [70]. Designing for longer-term learning experiences will also motivate the exploration of how students' relationships with chatbots evolve over time and how tutoring systems can remain responsive as a learner grows in age and knowledge [59]. There is also an opportunity to design personalized features that help counteract the "one-size-fits-all" educational paradigm, e.g., by adapting to a child's skills, interests, and even location.

CONCLUSION

Many students are not motivated to learn or are failing to succeed in formal educational settings [58]. Today's children need novel, engaging, and effective methods that cultivate their passion for learning. Our results suggest that a narrative-based learning system augmented with chatbot-mediated feedback may not only help encourage children to engage in non-formal learning but also help improve the educational outcomes of those activities.

SELECTION AND PARTICIPATION OF CHILDREN

This study was approved by the Stanford Institutional Review Board. We recruited 72 children in grades 3–5 through posts on NextDoor, university mailing lists, and word of mouth. We carefully explained the procedure including video recordings and data usage to all guardians and children and obtained their written consent. Throughout the study, the child was accompanied by at least one researcher at all times and was provided with water and snacks (with permission from parents). Researchers were instructed to inform children throughout the session, especially if they displayed signs of distress, that they could take breaks anytime and discontinue the study without consequences. We offered $75 Amazon gift cards as compensation for participating families regardless of completion.

ACKNOWLEDGEMENTS

We thank TAL for their gift funding and Tracy Cai, Glenn Davis, Liwei Jiang, Chung Wui Kang, Paula Moya, Bryce Tham, Ellen Wang, Angelica Willis, and Justin Xu for their additional help.

REFERENCES

[1] Khan Academy. 2019. Khan Academy. (2019). https://www.khanacademy.org/ Accessed: 2019-08-20.

[2] Deanne M Adams, Richard E Mayer, Andrew MacNamara, Alan Koenig, and Richard Wainess. 2012. Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of educational psychology* 104, 1 (2012), 235.

[3] Vincent Aleven and Kenneth R Koedinger. 2000. Limitations of student control: Do students know when they need help?. In *International conference on intelligent tutoring systems*. Springer, 292–303.

[4] Vincent Aleven, Ido Roll, Bruce M McLaren, and Kenneth R Koedinger. 2016. Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education* 26, 1 (2016), 205–223.

[5] Vincent Aleven, Elmar Stahl, Silke Schworm, Frank Fischer, and Raven Wallace. 2003. Help seeking and help design in interactive learning environments. *Review of educational research* 73, 3 (2003), 277–320.

[6] Isabel Machado Alexandre, David Jardim, and Pedro Faria Lopes. 2010. Maths4Kids: telling stories with Maths. In *Proceedings of the Intelligent Narrative Technologies III Workshop*. 1–6.

[7] National Governors Association and others. 2010. Common Core State Standards. *Washington, DC* (2010).

[8] Ryan S.J.d. Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[9] Jonathan Bassen, Bharathan Balaji, Michael Schaarschmidt, Jay Painter, Dawn Zimmaro, Alex Games, Ethan Fast, Candace Thille, and John Mitchell. 2020. Reinforcement Learning for the Adaptive Scheduling of Educational Activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

[10] Christopher C Blakesley. 2013. The role of narrative in the design of an educational game. *Dissertation abstracts international section A: Humanities and social sciences* (2013).

[11] Cyril Brom, Michal Preuss, and Daniel Klement. 2011. Are educational computer micro-games engaging and effective for knowledge acquisition at high-schools? A quasi-experimental study. *Computers & Education* 57, 3 (2011), 1971 – 1988. DOI:http://dx.doi.org/https://doi.org/10.1016/j.compedu.2011.04.007

[12] Rick Busselle and Helena Bilandzic. 2009. Measuring Narrative Engagement. *Media Psychology* 12, 4 (2009), 321–347. DOI: http://dx.doi.org/10.1080/15213260903287259

[13] Emma Carey, Francesca Hill, Amy Devine, and Denes Szucs. 2017. The Modified Abbreviated Math Anxiety Scale: A Valid and Reliable Instrument for Use with Children. *Frontiers in Psychology* 8 (2017), 11. DOI: http://dx.doi.org/10.3389/fpsyg.2017.00011

[14] Michelene T.H. Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[15] Sunhee Choi and Richard E Clark. 2006. Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of educational computing research* 34, 4 (2006), 441–466.

[16] Isaac Chuang and Andrew Ho. 2016. HarvardX and MITx: Four years of open online courses–fall 2012-summer 2016. *Available at SSRN 2889436* (2016).

[17] Andrew Clements. 1999. *Frindle*. Simon and Schuster.

[18] Merrill Cook. 2016. State of the MOOC 2016: A year of massive landscape change for massive open online courses. *Online Course Report* (2016). https://www.onlinecoursereport.com/state-of-the-mooc-2016-a-year-of-massive-landscape-\change-for-massive-open-online-courses/

[19] Diana I Cordova and Mark R Lepper. 1996. Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of educational psychology* 88, 4 (1996), 715.

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

[21] Carol S Dweck. 2008. *Mindset: The new psychology of success*. Random House Digital, Inc.

[22] Hans Magnus Enzensberger. 1998. *The number devil: A mathematical adventure*. Macmillan.

[23] Facebook. 2019. Facebook React. (2019). https://github.com/facebook/react

[24] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-based learning: Possible benefits and problems. *Communications* 34, 4 (2009), 429–447.

[25] Joshua Grossman, Zhiyuan Lin, Hao Sheng, Johnny Tian-Zheng Wei, Joseph Jay Williams, and Sharad Goel. 2019. MathBot: Transforming Online Resources for Learning Math into Conversational Interactions. *AAAI 2019 Story-Enabled Intelligence* (2019).

[26] Hasuka. 2019. Hasuka GraphQL. (2019). https://github.com/hasura/graphql-engine

[27] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[28] Kathy Hirsh-Pasek, Jennifer M Zosh, Roberta Michnick Golinkoff, James H Gray, Michael B Robb, and Jordy Kaufman. 2015. Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest* 16, 1 (2015), 3–34.

[29] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[30] Carnegie Learning Inc. 2019a. MATHia. (2019). https://www.carnegielearning.com/products/software-platform/mathia-learning-software/

[31] Singapore Math Inc. 2019b. Singapore Math Books. (2019). https://www.singaporemath.com/ Accessed: 2019-08-20.

[32] Worcester Polytechnic Institute. 2019. ASSISTments. (2019). https://new.assistments.org/

[33] Cheryl I Johnson and Richard E Mayer. 2009. A testing effect with multimedia learning. *Journal of Educational Psychology* 101, 3 (2009), 621.

[34] Cheryl I Johnson and Heather A Priest. 2014. 19 The Feedback Principle in Multimedia Learning. *The Cambridge handbook of multimedia learning* (2014), 449.

[35] W Lewis Johnson and James C Lester. 2018. Pedagogical Agents: Back to the Future. *AI Magazine* 39, 2 (2018).

[36] W Lewis Johnson, Jeff W Rickel, James C Lester, and others. 2000. Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in education* 11, 1 (2000), 47–78.

[37] W Lewis Johnson and Paola Rizzo. 2004. Politeness in tutoring dialogs:"run the factory, that's what I'd do". In *International Conference on Intelligent Tutoring Systems*. Springer, 67–76.

[38] Greg Jones and Scott Warren. 2009. The Time Factor: Leveraging Intelligent Agents and Directed Narratives in Online Learning. *Innovate: Journal of Online Education* 5, 2 (2009), 2.

[39] Jennifer A Kaminski, Vladimir M Sloutsky, and Andrew F Heckler. 2008. The advantage of abstract examples in learning math. *Science* 320, 5875 (2008), 454–455.

[40] Bilge Karacora, Morteza Dehghani, Nicole Kramer-Mertens, and Jonathan Gratch. 2012. The influence of virtual agents' gender and rapport on enhancing math performance. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

[41] Alice Kerlyl, Phil Hall, and Susan Bull. 2006. Bringing chatbots into education: Towards natural language negotiation of open learner models. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 179–192.

[42] Yanghee Kim and Amy L Baylor. 2006. A social-cognitive framework for pedagogical agents as learning companions. *Educational technology research and development* 54, 6 (2006), 569–596.

[43] Adena M Klem and James P Connell. 2004. Relationships matter: Linking teacher support to student engagement and achievement. *Journal of school health* 74, 7 (2004), 262–273.

[44] Kwan Min Lee, Namkee Park, and Seung A. Jin. 2006. *Narrative and interactivity in computer games*. Routledge Taylor & Francis Group, 304–322. DOI: http://dx.doi.org/10.4324/9780203873700

[45] James C Lester, Hiller A Spires, John L Nietfeld, James Minogue, Bradford W Mott, and Eleni V Lobene. 2014. Designing game-based learning environments for elementary science education: A narrative-centered learning perspective. *Information Sciences* 264 (2014), 4–18.

[46] Michael Madaio, Justine Cassell, and Amy Ogan. 2017. The impact of peer tutors' use of indirect feedback and instructions. In *Making a Difference: Prioritizing Equity and Access in CSCL, 12th International Conference on Computer Supported Collaborative Learning (CSCL)*, Vol. 1. Philadelphia, PA: International Society of the Learning Sciences.

[47] Michael Madaio, Kun Peng, Amy Ogan, and Justine Cassell. 2018. A climate of support: a process-oriented analysis of the impact of rapport on peer tutoring. In *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS)*, Vol. 1. London, UK: International Society of the Learning Sciences.

[48] Richard E. Mayer. 2009. *Multimedia Learning* (2 ed.). Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511811678

[49] Richard E Mayer. 2014. *Computer games for learning: An evidence-based approach*. MIT Press.

[50] Scott W McQuiggan, Jennifer L Robison, and James C Lester. 2010. Affective transitions in narrative-centered learning environments. *Journal of Educational Technology & Society* 13, 1 (2010), 40–53.

[51] Scott W Mcquiggan, Jonathan P Rowe, Sunyoung Lee, and James C Lester. 2008. Story-based learning: The impact of narrative on learning experiences and outcomes. In *International Conference on Intelligent Tutoring Systems*. Springer, 530–539.

[52] Microsoft. 2019. Microsoft TypeScript. (2019). https://github.com/microsoft/TypeScript

[53] Roxana Moreno and Richard E Mayer. 2000. Engaging students in active learning: The case for personalized multimedia messages. *Journal of educational psychology* 92, 4 (2000), 724.

[54] Roxana Moreno and Richard E Mayer. 2004. Personalized messages that promote science learning in virtual environments. *Journal of educational Psychology* 96, 1 (2004), 165.

[55] Patricia S Moyer. 2001. Are we having fun yet? How teachers use manipulatives to teach mathematics. *Educational Studies in mathematics* 47, 2 (2001), 175–197.

[56] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28 – 39. DOI: http://dx.doi.org/https://doi.org/10.1016/j.ijhcs.2018.01.004

[57] Jaclyn Ocumpaugh. 2015. Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual. *New York, NY and Manila, Philippines: Teachers College, Columbia University and Ateneo Laboratory for the Learning Sciences* 60 (2015).

[58] U.S. Department of Education. 2011. Partnering for Education Reform. (2011). https://www.ed.gov/news/speeches/partnering-education-reform Accessed: 2019-08-13.

[59] Amy Ogan, Samantha Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012. Rudeness and rapport: Insults and learning gains in peer tutoring. In *International Conference on Intelligent Tutoring Systems*. Springer, 11–21.

[60] Louise E Parker and Mark R Lepper. 1992. Effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of personality and social psychology* 62, 4 (1992), 625.

[61] Nancy E Perry. 1998. Young children's self-regulated learning and contexts that support it. *Journal of educational psychology* 90, 4 (1998), 715.

[62] Postgres. 2019. Postgres. (2019). https://www.postgresql.org/

[63] Vladimir Propp. 2010. *Morphology of the Folktale*. Vol. 9. University of Texas Press.

[64] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 2019a. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 357.

[65] Sherry Ruan, Angelica Willis, Qianyao Xu, Glenn M. Davis, Liwei Jiang, Emma Brunskill, and James A. Landay. 2019b. BookBuddy: Turning Digital Materials Into Interactive Foreign Language Lessons Through a Voice Chatbot. In *Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (L@S '19)*. ACM, New York, NY, USA, Article 30, 4 pages. DOI: http://dx.doi.org/10.1145/3330430.3333643

[66] Gregory Schraw. 1997. Situational interest in literary text. *Contemporary Educational Psychology* 22, 4 (1997), 436–456.

[67] Noah Schroeder, Rachel Barouch Gilbert, and Olusola Adesope. 2011. Pedagogical Agents and Learning: Are They Worth the Cost?. In *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Association for the Advancement of Computing in Education (AACE), 2078–2083.

[68] Valerie J Shute. 2008. Focus on formative feedback. *Review of educational research* 78, 1 (2008), 153–189.

[69] Tanmay Sinha and Justine Cassell. 2015. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*. 13–20.

[70] Nicholas C Soderstrom and Robert A Bjork. 2015. Learning versus performance: An integrative review. *Perspectives on Psychological Science* 10, 2 (2015), 176–199.

[71] Android Studio. 2019. Android Debug Bridge (adb). (2019). https://developer.android.com/studio/command-line/adb

[72] Joanna Szurmak and Mindy Thuna. 2013. Tell me a story: The use of narrative as a tool for instruction. *Conference of the Association of College and Research Libraries* 2013 (2013), 546–552.

[73] Candace A Walkington. 2013. Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology* 105, 4 (2013), 932.

[74] Ning Wang, W Lewis Johnson, Richard E Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2008. The politeness effect: Pedagogical agents and learning outcomes. *International journal of human-computer studies* 66, 2 (2008), 98–112.

[75] David Jackman Whitin. 1992. *Read Any Good Math Lately?: Children's Books for Mathematical Learning, K-6*. Heinemann Educational Books.

[76] Pieter Wouters, Christof Van Nimwegen, Herre Van Oostendorp, and Erik D Van Der Spek. 2013. A meta-analysis of the cognitive and motivational effects of serious games. *Journal of educational psychology* 105, 2 (2013), 249.

[77] Michael F Young, Stephen Slota, Andrew B Cutter, Gerard Jalette, Greg Mullin, Benedict Lai, Zeus Simeoni, Matthew Tran, and Mariya Yukhymenko. 2012. Our princess is in another castle: A review of trends in serious gaming for education. *Review of educational research* 82, 1 (2012), 61–89.

[78] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Proceedings of the 17th annual meeting of the Special Interest Group on Discourse and Dialogue*. 55–63.

[79] Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 514–527.