Hint Systems May Negatively Impact Performance in Educational Games

Eleanor O'Rourke, Christy Ballweber, and Zoran Popović
Center for Game Science
Department of Computer Science & Engineering, University of Washington
{eorourke, christy, zoran}@cs.washington.edu

ABSTRACT

Video games are increasingly recognized as a compelling platform for instruction that could be leveraged to teach students at scale. Hint systems that provide personalized feedback to students in real time are a central component of many effective interactive learning environments, however little is known about how hints impact player behavior and motivation in educational games. In this work, we study the effectiveness of hints by comparing four designs based on successful hint systems in intelligent tutoring systems and commercial games. We present results from a study of 50,000 students showing that all four hint systems negatively impacted performance compared to a baseline condition with no hints. These results suggest that traditional hint systems may not translate well into the educational game environment, highlighting the importance of studying student behavior to understand the impact of new interactive learning technologies.

Author Keywords

Hint systems, educational games, behavioral analytics.

ACM Classification Keywords

H.5.0 Information interfaces and presentation: General

INTRODUCTION

Video games are famous for their ability to motivate players to solve challenging, complex problems. As a result, there is a growing interest in leveraging games to teach serious content to students [14, 22]. Games have many features that make them well-suited for learning: they can adapt to meet individual needs, provide interactive learning experiences, and scale easily [22]. Furthermore, educational games have been shown to increase student's motivation [25] and time-on-task [20]. Many commercial video games utilize real-time hint systems to provide help to struggling players. Personalized feedback is recognized as an important part of the learning process [16], and hint systems could be leveraged to improve the effectiveness of educational games. However, little is known about how hints impact learning-related behavior in these environments.

L@S 2014, March 04 - 05 2014, Atlanta, GA, USA

http://dx.doi.org/10.1145/2556325.2566248

Hint systems are a central component of intelligent tutoring systems, and many hint design questions have been studied in this context. Studies have shown that students perform better when tutoring systems provide hints [6]. The most effective hints respond when students make mistakes, providing a scaffolded sequence of increasingly concrete information [4, 6, 27]. However, little is known about how these approaches will translate into the educational game environment. Games are typically played voluntarily, and they provide students with open-ended learning experiences. As a result, students' motivations to struggle, persist, use hints, and eventually learn are very different in educational games than in cognitive tutors. Understanding the impact of hint systems in this type of environment would therefore be valuable.

In this work, we study the effects of hint systems on player behavior in the educational game *Refraction*. We draw inspiration for our designs from successful hint systems used in intelligent tutoring systems and commercial video games. Although many factors influence hint design, we chose to focus on how hints are presented to players and the content they provide. We explore the importance of these characteristics by conducting a factorial experiment on the educational website BrainPOP. An analysis of data from 50,000 students shows that all four hint designs negatively impact performance compared to a baseline condition with no hints. Rewarding players with earned hints slightly improved player persistence, but this had no positive effect on performance.

Our findings suggest that the hint system designs that are effective in existing educational tools may not directly translate into educational video games. The learning environment provided by online games designed to be played at scale has many unique characteristics. Students participate voluntarily and can quit at any point, so maintaining engagement is paramount. Asking for help has a known negative connotation in games [5, 11], which may reduce the impact of hints in this setting. Further research is needed to understand the full generality of these results, however we hope these findings encourage developers to carefully consider the design of hint systems, particularly when engagement is a key component of the learning experience.

BACKGROUND

Personalized feedback is recognized as a central part of the one-on-one learning experience [16], and providing effective hints to students has long been a goal of the educational technology community [27]. Many questions relating to hint system design have been studied in the context of intelligent tu-

toring systems (ITS), however little is known about how these designs will translate into educational games. In this section, we review related research from the intelligent tutoring systems and video games literature.

Hints in Intelligent Tutoring Systems

Intelligent tutoring systems, designed to emulate one-on-one tutoring environments, have been studied for decades. Interactive feedback and hint systems are typically a central component of ITSs [27, 4], and research shows that students perform better in these environments when hints are provided [6]. Corbett and Anderson compared three hint designs in their LISP tutor to a version with no hints, and found that students in all three hint conditions performed over 20% better on a post-test than students in the no-hints condition [13]. As a result, many hint design questions have been studied in the context of ITSs. We focus on three design questions highlighted as important in the ITS literature: when should hints be given to students, what content should hints contain, and how should hints be presented [27].

In an overview of ITS behavior, VanLehn states that hints should only be given to students when they really need help [27]. Automated tutors should refuse to give help to until the student becomes frustrated or stops making progress, at which point a hints should be given automatically. Murray et al. implemented a policy based on this model in the DT Tutor [23], but their solution was computationally expensive and required excessive student data, making it unsuitable for real-world applications [27]. Most ITSs avoid this problem by only providing hints when students explicitly ask for help [27, 3]. One downside of this solution is that students frequently abuse help when it is not needed or refuse help when they are struggling [2]. However, Aleven and Koedigner show that these problems can be reduced if the ITS also teaches metacognitive and help-seeking strategies [2, 3].

Determining what content hints should contain is another important design question. In ITSs, hints often revel a step in a multi-step procedure. VanLehn states that the step should be chosen according to the following criteria: the step must be correct, it must be a step the student has not completed, it must align with the student's current solution plan, and it must honor the instructor's preferences about the desired solution method [27]. However, providing this type of hint requires access to a real-time model of the student's progress and the desired solution, which may not be available in all interactive learning systems. Arroyo et al. compared the effectiveness of concrete hint content that referred to physical examples and symbolic hint content that used numeric expressions and operations. They found that children with low cognitive ability perform better with concrete hints, while children with high cognitive ability perform better with symbolic hints [7].

ITSs typically present hint content textually, and provide a sequence of scaffolded hints that the student can click through to access increasingly specific help [27]. This sequence often begins with a "pointing hint", which highlights information that is relevant to completing the next step in a procedure [16]. This is followed by a "teaching" hint, which presents factual information required to complete the step, and finally a "bottom-out" hint that reveals how to complete the next step [16]. Khan Academy's practice problem system provides scaffolded hints that include richer content such as images and diagrams [17]. Arroyo et al. studied the effectiveness of interactive hints that require student response, and found that girls performed better with interactive hints but that boys performed better with standard textual hints [7, 8].

The hint systems that have been effective in intelligent tutoring systems provide a valuable theory around which to base the design of hints for other interactive learning systems. However, it is not currently known how these hint designs will translate into large-scale learning environments such as educational games. The effects of hints on student motivation and engagement are not well understood, and help buttons may not be effective in immersive environments. We seek to address these questions through our work studying the impact of hints in educational games.

Hints in Video Games

A central challenge in video game design is to support players as they develop the skills to solve increasingly complex problems. Games frequently depend on hint systems to help players who are stuck or frustrated. Some commercial games, such as *Bad Piggies* (Rovio Entertainment Ltd. 2012) and *SquareLogic* (TrueThought LLC 2009), provide hints through interfaces similar to the "hint buttons" found in ITSs. However, previous research has shown that there is a strong negative connotation to asking for help in games. Andersen et al. found that adding a help button to *Refraction* caused players to complete 12% fewer levels and play for 15% less time, even though only 31% of players actually clicked the button [5]. Big Fish Studios discovered that players avoided using the hint system they designed for *Drawn: Dark Flight* because they saw hints as punishment [11]. They were able to reverse this effect by replacing the word "hint" with "advice." These results suggest that the standard "hint button" interface for presenting on-demand help may not be appropriate in games where maintaining player engagement is crucial.

As a result, many designers have chosen to closely integrate hint systems in the game environment. Many immersive role-playing games, such as *The Legend of Zelda* (Nintendo 1998) and *Superbrothers: Sword & Sworcery EP* (Superbrothers and Capybara Games, 2012), provide help through characters that players encounter in the game world. These characters typically give textual hints during dialogs with the player. Other games, such as *Professor Layton* (Level-5 2007) and *Puzzle Agent* (Telltale Games 2010), allow players to purchase hints with earned game items such as coins or points. While these designs have many different properties, they all incorporate hints into the game environment and remove the negative connotation of asking for help by making players feel like they deserve to use the hints they have collected.

Despite these examples of successful hint systems in commercial games, many games do not provide hints. Game development companies often depend on external user-generated resources such as walkthroughs and wikis to support struggling players. These resources cost nothing to produce, however research in human-computer interaction has

shown that context-insensitive help systems present many challenges for users, and are typically not as effective as help that is directly integrated into software products [1, 18, 19]. We believe that well-designed hint systems that are integrated into the game environment have more potential for success, particularly in the educational context where providing high-quality help resources is crucial.

Research Questions

The goal of this work was to gain a better understanding of how hints impact player engagement, performance and behavior in the educational game *Refraction*. To explore this question, we designed a factorial experiment comparing hint designs that varied along two axes: hint *presentation* and hint *content*. We chose to study these characteristics because they were highlighted as important in both the ITS and game literature. We implemented two methods of presenting hints to players: one that embeds hints into game levels and another that awards players "earned" hints as they progress. We chose to explore integrated hint designs rather than a standard "hint button" interface to reduce the negative connotation of asking for help, which has been shown to be a problem in video games [5, 11]. We were interested in learning how hint presentation would affect both student engagement and performance in *Refraction*.

Research Question 1: *Will player behavior differ based on hint presentation?*

Our two methods of presenting hints to players require that the system give a single hint at a time, rather than a sequence of scaffolded hints. We were therefore interested in studying the effectiveness of different types of hint content. We designed two hint models based on designs commonly used in ITSs: concrete "bottom-out" hints that tell players exactly which game pieces to use, and abstract "pointing" hints that highlight information relevant to the solution. We expected concrete hints to help players progress in the current level because they provide directly relevant information. However, we thought that abstract hints would produce better long-term performance because they provide suggestions that can generalize more easily.

Research Question 2: *Will concrete hints produce better immediate performance than abstract hints?*

Research Question 3: *Will abstract hints produce better long-term performance than concrete hints?*

We also included a baseline version of *Refraction* with no hints in our study to measure how the presence of hints affects player behavior. We expected all four hint systems to improve player performance compared to this baseline, since they provide access to help resources that are otherwise unavailable for struggling players.

Research Question 4: *Will all four hint systems improve performance over the baseline?*

To study these research questions, we released all five versions of *Refraction* to BrainPOP, a popular educational website for elementary school students that provides a game portal [9]. We analyzed data from 50,000 players to determine the impact of incorporating hint systems into the game.

Figure 1. A level of Refraction. The goal is to use the pieces on the right to split lasers into fractional pieces and redirect them to satisfy the target spaceships. All spaceships must be satisfied at the same time to win.

Order	Topic
1	Directionality of bender pieces
2	Make halves by splitting in two
3	Make thirds by splitting in three
4	Make fourths by splitting in two twice
5	Make eighths by splitting in two three times
6	Make both a half and a fourth
7	Split in three first to make both a third and a sixth
8	Split in three first to make both a sixth and a ninth

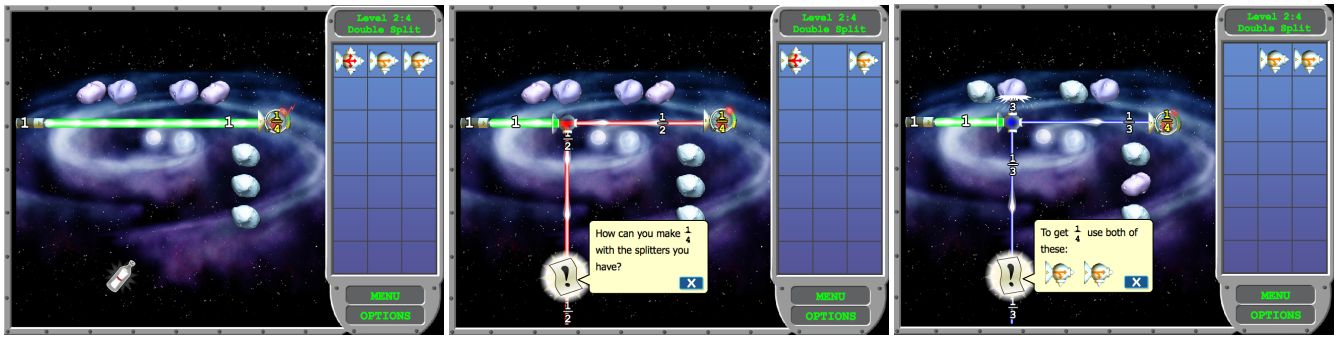
Table 1. The eight milestone concepts covered by the level progression. The progression includes an introductory level and evaluation level for each concept, used to measure student performance.

EXPERIMENT DESIGN

To explore the impact of hint systems on player behavior, we implemented four distinct hint systems in the educational game *Refraction*. These hint systems varied in how they presented hints to players, and what type of hint content they provided. We also implemented a baseline version of the game without hints. In this section, we describe each of our designs in detail and discuss how they are integrated into the *Refraction* game environment.

Refraction

This educational puzzle game was designed by game researchers and learning science experts at the Center for Game Science to teach fraction concepts to elementary school students. To play, a child must interact with a grid that contains laser sources, target spaceships, and asteroids, as shown in Figure 1. The goal of the game is to satisfy target spaceships by splitting the laser into the correct fractional amounts and avoiding asteroids. The player uses pieces that either change the laser direction or split the laser into two or three equal parts to achieve this goal. To win, the player must correctly satisfy all the target spaceships at the same time. *Refraction* has been successful at attracting elementary school students, and has been played over 250,000 times on the educational website BrainPOP since its release in April 2012.



(a) (b) (c)

Figure 2. Screenshots of the embedded hints. Figure 2(a) shows an uncollected message in a bottle on the Refraction grid. Figure 2(b) shows an abstract hint being displayed. Figure 2(c) shows an concrete hint being displayed.

(a) (b) (c)

Figure 3. Screenshots of the earned hints. Figure 3(a) shows the message that displays when a hint is earned. Figure 3(b) shows a concrete hint being displayed. Figure 3(c) shows what the robot button looks like when there are no earned hints available.

For this study, we implemented a *Refraction* level progression that included 35 levels covering 14 mathematical concepts. In casual online gaming environments, the majority of players leave after a short period of time. Previous research shows that children play *Refraction* on BrainPOP for about three minutes on average [24]. As a result, we focus on the eight concepts described in Table 1 in our evaluation. For each concept, we designed an introductory level that presented the material for the first time. This level was directly followed by an evaluation level designed to give player a chance to apply their understanding of the concept. In both the introductory and evaluation levels, students had to choose between pieces that would make the correct target fraction and pieces that would make incorrect fractions. We use these levels to evaluate students' understanding of the eight fraction concepts.

Hint Presentation

The two hint presentation modalities we implemented for this experiment are based around designs that have been successful in commercial games. We avoided presenting hints to players using a simple "hint button", even though this method is commonly used in interactive learning environments, due to the negative connotations that players have with asking for help in games [5, 11]. Players are required to explicitly solicit hints in both of our designs, however we integrate these hints directly into the game environment. In one design, we embed hints into *Refraction* levels and in the other we reward players with earned hints as they progress.

Our first hint presentation design encourages players to use hints by embedding them directly into *Refraction* levels and making them part of the game environment. This design was inspired by hints given in games like *The Legend of Zelda* (Nintendo 1998) where hints are encountered as the player explores the game world. Since *Refraction* is a puzzle game, rather than an immersive role-playing game, we added hints as a new type of level object that can appear on the game grid. We used an icon depicting a message in a bottle to represent the hint, as shown in Figure 2(a). To open the hint, the player must direct the laser through the bottle, as shown in Figure 2(b). Any fractional value of laser will open the hint, to make the hints simple to access. However, we placed hint bottles in grid squares off the main path in each level to ensure that the player had to actively solicit the hint. Embedded hints were added to the introductory levels of each of the 14 concepts in the level progression to encourage players to use help as they were learning. Players were only able to view hints on these 14 levels; there was no way to view a hint on a level that did not contain a bottle object.

Our second hint presentation design rewards players with "earned" hints as they progress to make them feel that they deserve help. This design is inspired by the hint systems in

Concept	Concrete	Abstract
Bender Piece	Use this piece: **piece**	Which side of the piece can the laser enter?
Single Split	To get **fraction** split the laser using: **piece**	How much power do the ships need to be happy?
Multiple Splits	To get **fraction** use both of these: **piece piece**	How can you make **fraction** with the splitters you have?
Multiple Fractions	To get **first fraction** use **piece** Then split again with **piece** to get **second fraction**	How can you make **first fraction** first, and then **second fraction** with the splitters you have?
Split Ordering	To get both **first fraction** and **second fraction** use this first:	Which splitter should you use first to make both ships happy?
Multiple Lasers	This level has two lasers! To get **fraction** use both of these: **piece piece**	This level has two lasers! How can you make **fraction** with the splitters you have?
Fractional Laser Source	The lasers have smaller values! Use both of these: **piece piece**	The lasers have smaller values! How much power do the ships need to be happy?
Fractional Laser with Splits	To get **fraction** split the **laser fraction** laser using: **piece**	How can you split **laser fraction** to make **fraction** with the splitters you have?
Multiple Fractional Lasers	To get **fraction** split the **laser fraction** laser using: **piece**	Which laser should you split to get **fraction**?

Table 2. A template defining hint content for the abstract and concrete hints. Text shown in red bold that says "piece" is replaced with an image of the piece associated with this hint. An example of this presentation is shown in 2(c). Red bold text that says "fraction" is replaced by the appropriate fractional value. An example of this presentation is shown in 2(b).

games like *Professor Layton* (Level-5 2007) in which players trade collected items for hints. Since *Refraction* is a casual puzzle game, we decided to implement a simple hint reward system where players earn hints at the beginning of certain levels. We created a cute robot character who serves as a guide and awards hints to players. We chose to personify hint-related messages because studies show that players are more willing to accept feedback from personified characters that neutral interfaces [21]. The robot notifies the player when a hint has been earned using a motivating message, shown in Figure 3(a). The player can view this earned hint at any time by clicking on the robot button, shown in Figure 3(b). The button is grayed out when no hints are available, as in Figure 3(c). Hints are earned at the beginning of each of the 14 levels that introduce new concepts to encourage player to use help resources on these levels, however players can chose to save earned hints and use them on any future level.

Hint Content

We designed two types of hint content that varied in the level of abstraction of the information they provided to players. Since our methods of presenting hints to players were designed to give the player a single hint at a time, we were not able to provide a scaffolded sequence of hints as is common in intelligent tutoring systems. We were therefore interested in learning about the impact of different types of hint content. We designed two types of hints based on designs that are commonly used in ITSs: concrete "bottom-out" hints and abstract "pointing" hints [16, 27].

The concrete hints tell the player exactly which pieces are required in the solution. Since it is difficult to reference pieces using text, we inserted images of the pieces into our hint content. The hint did not tell players where to put the pieces on the grid, but knowing which pieces are needed to make the correct fraction removes most of the problem's difficultly. An example concrete hint is shown in Figure 2(c). The abstract pointing hints were designed to highlight important information related to solving the problem. These hints were framed as questions that players should be asking themselves to arrive at the correct solution. These hints did not reference specific pieces or any other concrete game objects, so they did not include images. An example abstract hint is shown in Figure 2(b).

While we would have liked to follow the ITS standard and provide personalized hints to each player based on a cognitive model, we do not currently have access to real-time models of player progress and level solutions. As a result, we designed a single hint for each *Refraction* level that targets the central challenge presented by that level. While these hints may not always address the player's current struggle, they provide information integral to solving the level that the player could use to build a solution. Table 2 provides a complete template for the concrete and abstract hints used to cover the 14 concepts included in the *Refraction* level progression.

METHOD

To explore the impact of hint systems on player behavior in educational games, we studied how students play the five versions of *Refraction* on the educational website BrainPOP [9]. BrainPOP is best known for its curriculum resources, however it also provides an educational game portal intended for use in the classroom. The BrainPOP Educators community has over 210,000 members [10], and the website is used as a resource in around 20% of elementary schools in the United States (Traci Kampel, personal communication).

A central benefit of using the BrainPOP game portal to study hint systems is that it provides access to a large, diverse population of students. Interactions with hint systems are sparse, occurring only when students struggle and ask for help. As a result, studying hint usage across a large data set of students with varied backgrounds is valuable. Furthermore, these students do not know that they are part of an experiment, so our findings reflect their natural behavior. However, one downside of this resource is that we know very little about the children who visit BrainPOP or the contexts in which they play. We cannot collect any demographic information, and while we know that the website is primarily used in schools, we cannot tell whether children are playing in the classroom, in a computer lab, or at an after-school program. We mitigate the effects of these uncontrolled variables by evenly distributing them between conditions through the use of randomization and large sample sizes.

To collect our data, we set up *Refraction* to randomly assign new players to one of the five versions of the game, and logged all interactions players made with the game or its interface. We only included new players who were not already

familiar with *Refraction* in our analysis, and we only used data from a player's first session to control for issues with shared computers in schools. To track players, we stored their player ids and progress in the Flash cache. This allowed us to selectively include new players and exclude return sessions. One drawback of this method is that a player who clears the cache or switches computers will be treated as a new player by our system. However, since the Flash cache is inconvenient to clear and this action deletes all saved game progress, we consider this risk to be small.

Our data set contains 79,895 players, and was collected between August 29, 2012 and November 20, 2012. *Refraction* was featured on the front page of BrainPOP's game portal between September 3rd and 7th 2012, allowing us to attract a large number of players during that period. Since the data sets for the five conditions contain different numbers of players, we randomly selected 10,000 players from each condition to include in our analysis.

DATA ANALYSIS AND RESULTS

We study the impact of our hint systems by analyzing a number of outcome measures that capture player engagement, performance, and hint usage behavior. We describe each of these metrics in detail below. Before performing this analysis, we evaluated the Kolmogorov-Smirnov test to assess the normality of our data, and found that it was statistically significant for all of our outcome measures. We therefore used non-parametric statistical methods for our analysis.

Our study has a 2x2 factorial design with two between-subjects factors: hint *presentation* with levels earned and embedded, and hint *content* with levels concrete and abstract. Our design therefore warranted a non-parametric factorial analysis. For our dichotomous outcome measures, we used a binomial logistic regression and a Cramer's V measure of effect size. For our continuous outcome measures, we applied the Aligned Rank Transform [15, 26] procedure, which aligns and ranks non-parametric data so that a standard ANOVA model can be used to perform a factorial analysis. For each main effect or interaction, the ART procedure aligns the data such that only that main effect or interaction remains, and then ranks the aligned data. A standard ANOVA model can then be used on the ranked data to measure the effect for which it was aligned. Unlike the conventional rank transform [12], the ART procedure is known to preserve the integrity of interaction effects and not inflate Type I errors. We used the ARTool program to align and rank our data [28]. We used an Eta-Squared (η^2) measure of effect size for these continuous outcome measures.

We also included a baseline condition with no hints in our study. We use pairwise comparisons to measure differences between our four hint designs and this baseline condition. Since this produces a large number of comparisons, we risk inflating alpha. To address this potential issue, we sum *p*-values across the four comparisons for each outcome measure to ensure that the combined alpha still falls below the 0.05 threshold. We use a Wilcoxon rank sums test and an Cramer's *V* measure of effect size for our our nominal variables, and a Kruskal Wallis test and an *r* measure of effect size for pairwise comparisons of our continuous variables.

We report effect sizes in addition to *p*-values to show the magnitude of the differences between our populations, since we are likely to find many significant differences due to our large sample sizes. For the Cramer's *V* and *r* measures of effect size, effects with values less than 0.1 are considered *very small*, 0.1 are *small*, 0.3 are *moderate*, and 0.5 or greater are *large*. For the η^2 measure of effect size, effects with values less than 0.01 are considered *very small*, 0.01 are *small*, 0.06 are *moderate*, and 0.14 or greater are *large*.

Earned hints improved engagement

We expected our hint designs to have an impact on player engagement, and specifically thought that concrete hints would keep players more engaged than abstract hints. To evaluate this hypothesis, we measured how long children played *Refraction*. BrainPOP offers a large variety of games, many of which teach fraction concepts, so we expected players to quit *Refraction* if they became bored or frustrated. Therefore, our time played metric captures how long players are willing to persist in the game before choosing to leave, an approximation of their level of engagement.

We calculate active time played by counting the number of seconds each player spends in the game, excluding menu navigation and idle periods with more than thirty seconds between actions. Any time spent reading hints was included in the active time. Our analysis showed that hint *presentation* had a significant main effect on active time played ($F(1,39996)=17.29$, $p<0.0001$, $\eta^2=0.001$). Children with earned hints played significantly longer, a median of 192 seconds compared to 178 seconds for players with embedded hints. There was no main effect of hint *content* ($F(1,39996)=2.45$, *n.s.*) or the interaction *presentation*content* ($F(1,39996)=3.75$, *n.s.*). We also used pairwise comparisons to measure the differences between our four hint conditions and the baseline condition. We found that players in both of the earned hint conditions played significantly longer than players in the baseline condition, as shown in Table 3. However, the combined alpha for our three significant comparisons is 0.0516, slightly above the 0.05 threshold.

We also analyzed the amount of time players spent reading hints to determine whether this could explain the observed differences in total time played. To calculate the total time each player spent reading hints, we counted the number of seconds between each "hint opened" and "hint closed" action and summed across all hint views. Our analysis showed that hint *presentation* had a significant main effect on the amount of time spent reading hints ($F(1,39996)=272.74$, $p<0.0001$, $\eta^2=0.07$). Players with embedded hints spent more time reading hint content, a median of 4.77 seconds compared to 2.62 second for players with earned hints. Hint *content* also had a significant main effect on the amount of time spent reading hints ($F(1,39996)=20.14$, $p<0.0001$, $\eta^2=0.04$). Players with concrete hints read for an average of 3.67 seconds, compared to 3.19 seconds for players with embedded hints. The *presentation*content* interaction had no significant effect ($F(1,39996)=0.81$, *n.s.*).

Condition	Time Played		Unique Levels Completed		Won Intro / Started Intro		Won Eval / Started Intro	
No Hints Baseline	$N = 20,000$	$n.s.$	$N = 20,000$	$p = 0.0015$	$N = 62,476$	$p = 0.0056$	$N = 62,476$	$p = 0.0002$
Concrete Embedded Hints	$Z = 0.12$		$Z = -3.17$	$r = 0.01$	$\chi^2 = 7.66$	$V = 0.01$	$\chi^2 = 14.31$	$V = 0.02$
No Hints Baseline	$N = 20,000$	$n.s.$	$N = 20,000$	$p < 0.0001$	$N = 61,448$	$p < .0001.$	$N = 61,448$	$p < 0.0001$
Abstract Embedded Hints	$Z = -1.16$		$Z = -5.88$	$r = 0.03$	$\chi^2 = 77.06$	$V = 0.04$	$\chi^2 = 45.18$	$V = 0.03$
No Hints Baseline	$N = 20,000$	$p = 0.0037$	$N = 20,000$	$n.s.$	$N = 64,169$	$p < .0001$	$N = 64,169$	$p = 0.0141.$
Concrete Earned Hints	$Z = 2.90$	$r = 0.01$	$Z = 0.74$		$\chi^2 = 26.55$	$V = 0.02$	$\chi^2 = 6.02$	$V = 0.01$
No Hints Baseline	$N = 20,000$	$p = 0.0478$	$N = 20,000$	$n.s.$	$N = 63,482$	$p < .0001$	$N = 63,482$	$p < .0001$
Abstract Earned Hints	$Z = 1.99$	$r = 0.01$	$Z = -0.77$		$\chi^2 = 64.44$	$V = 0.03$	$\chi^2 = 26.32$	$V = 0.02$

Table 3. Results from the pairwise comparisons between our four hint conditions and a baseline condition with no hints. We report results for the following four metrics: the amount of active time played, the number of unique levels completed, the percentage of players who start introductory levels and win, and the percentage of players who start introductory levels and win the corresponding evaluation level.

These results suggest that earned hints increase engagement, motivating children to continue playing for significantly longer. Furthermore, this increase cannot be explained by the time children spent reading hints. Players with earned hints spent significantly less time reading hint content those with embedded hints. This effect is very small, however previous studies have measured negative effects on time played after adding hints [5], so it is encouraging to see hint systems produce even neutral effects on engagement.

All hint systems negatively impacted performance

We expected all four of our hint systems to improve player performance because hints provide access to otherwise unavailable help. We thought that abstract hints would have a stronger positive effect on long-term performance than concrete hints because they provide more generalizable information. Since children play *Refraction* on BrainPOP for such a short period of time, we were unable to formally assess player performance using pre- and post-tests. Instead, we measure players' ability to complete game levels, since successfully solving *Refraction* puzzles requires some understanding of the concepts taught by the game.

First, we calculated the number of unique levels each player completed by counting levels with win events. Level completion rates are closely tied to the amount of time a player spends in the game, so we expected our results to mirror the time played results. Our analysis did show that *presentation* had a significant main effect on unique levels completed ($F(1,39996)=36.01$, $p<0.0001$, $\eta^2=0.007$), with players in the earned hint conditions completing a median of 7 levels, compared to 6 for players in the embedded hint conditions. Hint *content* also had a main effect on unique levels completed ($F(1,39996)=5.01$, $p<0.05$, $\eta^2=0.004$). The median number of levels completed was 6 for both conditions, but players with concrete hints completed an average of 7.85 levels, compared to 7.59 for players with abstract hints. The *presentation*content* interaction did not have a significant effect ($F(1,39996)=0.03$, *n.s.*). Pairwise comparisons with the baseline condition, included in Table 3, showed that players in the two embedded hint conditions completed significantly fewer levels than players in the baseline.

Next, we analyzed how players performed on the first eight levels that introduced new concepts in our *Refraction* level progression. For each introductory level, we calculated the percentage of players who started the level and went on to win. We only included players who started each introductory level in this analysis to control for differences in play time across conditions. Then, we averaged the percentages across all eight introductory levels to create a single combined metric. Our analysis showed that hint *presentation* had a significant main effect on the percentage of players who won introductory levels ($\chi^2(1,N=156,988)=19.98$, $p<.0001$, $V=0.01$). Players with embedded hints won more often, 88.52% of the time compared to 87.84% of the time for players with earned hints. Hint *content* also had a main effect on the introductory level win rate ($\chi^2(1,N=156,988)=11.78$, $p<.001$, $V=0.01$). Players with concrete hints won more often, 88.50% of the time compared to 88.06% of the time for players with abstract hints. The *presentation*content* interaction did not have a significant effect ($\chi^2(1,N=40,000)=0.81$, *n.s.*). Pairwise comparisons showed that players in the baseline condition performed significantly better than players in all four hint conditions, winning 89.45% of the time. See Table 3.

We used a similar metric to analyze performance on the eight corresponding evaluation levels. For each evaluation level, we calculated the percentage of players who started the introductory level for that concept and went on to win the evaluation level. This metric was designed to capture how well a player's understanding of a newly introduced concept transfers to a second puzzle. Again, we averaged the percentages across all eight evaluation levels to create a single combined metric. Hint *presentation* did not have a significant main effect on performance in the evaluation levels ($\chi^2(1,N=156,988)=0.09$, *n.s.*), and neither did hint content ($\chi^2(1,N=156,988)=2.83$, *n.s.*). However, the *presentation*content* did have a significant effect on performance in the evaluation levels $\chi^2(1,N=156,988)=5.64$, $p<.05$, $V=0.01$). Players with concrete earned hints won the evaluation level 77.01% of the time, players with concrete embedded hints won 76.55% of the time, players with abstract earned hints won 76.10% of the time, and players with abstract embedded hints won 75.52%. Pairwise comparisons showed that players in the baseline condition performed significantly better, winning 77.82% of the time. See Table 3.

These results directly oppose our expectations. We expected all of our hint systems to improve performance because they provide struggling players with help that was otherwise unavailable. However, players in all four hint conditions completed fewer levels than we anticipated given their average time played. It is possible that players complete fewer levels in the same amount of time due to the time they spend reading hint content. However, this does not explain why hints

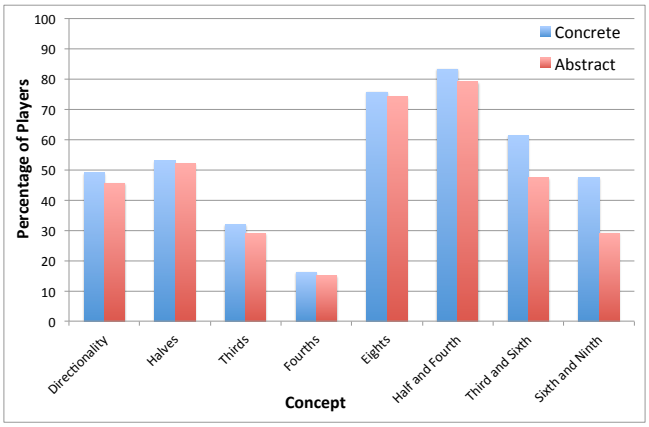


Figure 4. The percentage of players who opened Embedded hints on the eight introductory levels.

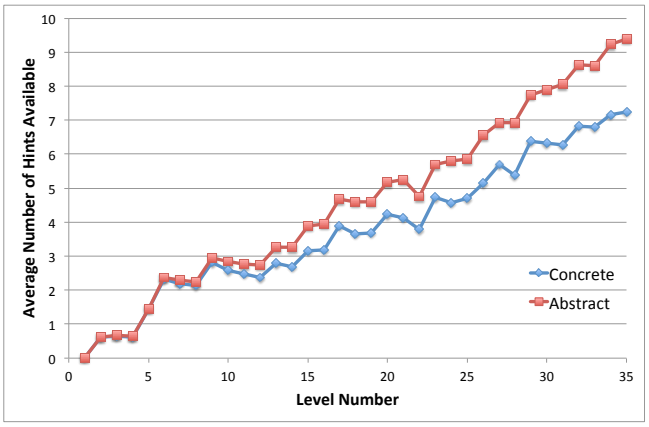


Figure 5. The average number of earned hints stored for each level in *Refraction*.

did not help players win the introductory levels. The hint presentation structure ensures that every player has access to a hint during every introductory level, so we would expect hints to improve performance on those levels. We also hypothesized that abstract hints would benefit player performance more than concrete hints, but we observed the opposite effect. Players with concrete hints performed better than those with abstract hints. While the sizes of these effects are small, hints certainly did not improve performance as we were expecting.

Embedded hints are viewed more often than earned hints

We were interested in learning how the four hint systems were used by players, and in identifying any differences in usage patterns. We were not sure how hint presentation would affect hint usage, but we expected players to use concrete hints more often than abstract ones because they provide more directly applicable information. To investigate how players used hints, we analyzed a variety of outcome measures.

First, measured the number of hints viewed by players in each condition by counting the number of hints opened by each player. Our analysis showed that *presentation* had a significant main effect on the number of hints viewed ($F(1,39996)=469.07$, $p<0.0001$, $\eta^2=0.05$). While the median number of hints viewed for both presentations was 1, players with embedded hints viewed an average 2.67 hints compared to 1.58 for players with earned hints. Hint *content* also had a significant main effect on hints viewed ($F(1,39996)=4.20$, $p<0.05$, $\eta^2=0.07$). Again, the median number of hints viewed in both conditions was 1, but players with concrete hints viewed 2.26 hints on average, compared to 1.98 viewed by players with abstract hints. The *presentation*content* interaction did not have a significant effect ($F(1,39996)=1.06$, *n.s.*).

Next, we looked at how hints were used in the two embedded hint conditions. While we knew that embedded hints were viewed more often than earned ones, we wanted to measure what percentage of players chose to open embedded hints. We calculated the percentage of players who opened the message-in-a-bottle hints in the eight introductory levels. We found that nearly 50% of players viewed embedded hints on average. However, the graph in Figure 4 shows that hint usage varies drastically depending on the concept, indicating that players are more likely to view hints for difficult concepts. We also found that *content* had a significant effect on the percentage of players who viewed hints ($\chi^2(1,N=60,874)=78.16$, $p<.0001$, $V=0.04$). 49.04% of players with concrete hints viewed introductory level hints on average, compared to 47.27% of players with abstract hints.

We also looked at how hints were used in the two earned hint conditions. Players could view earned hints on any level, not just the eight levels that introduced new concepts. To learn how players chose to use earned hints, we measured how many hints they had saved on average. The graph in Figure 5 shows the average number of saved hints available for each level in the progression. It clearly shows that players hoard earned hints, because the average number of hints saved increases as players progress through the game. We also analyzed whether hint content had any effect on saving behavior. For each player, we summed the number of earned hints the player had available during each level, and divided by the total number of levels played to calculate the average number of hints saved. We found that hint *content* had a significant effect on the number of hints saved ($Z=4.65$, $p<.0001$, $r=0.03$). Players were more likely to use concrete hints; those in the concrete hints condition had a median of 0.67 hints available, compared to 0.75 for those in the abstract hints condition.

After discovering that players hoard earned hints, we were interested in learning whether the hints they do view are used judiciously. To explore this question, we calculated whether players view earned hints on levels in which they are struggling. First, we computed the median number of moves made on each level across all five conditions. We considered a player to struggle if she made more than the median number of moves for that level. We found that players used hints on levels in which they struggled nearly 75% of the time. We also found that hint content had a significant effect on whether players used hints judiciously ($\chi^2(1,N=20,000)=141.59$, $p<.0001$, $V=0.08$). Players with concrete hints only used hints when they were struggling 74.32% of the time, compared to 80.40% of the time for players with abstract hints.

These results make sense intuitively. Embedded hints are viewed often because they have no value. However, earned hints can be saved for difficult future levels, so it is not surprising that players are reluctant to use them. Players in the earned hint conditions wait until they are struggling to view hints, while players with embedded hints view them when they are available. These results also suggest that concrete hints are more valuable than abstract hints, since players view concrete hints more often across both presentation types.

Concrete hints were more helpful than abstract hints

The analysis of hint usage in the earned and embedded hint systems suggests that players value concrete hints more highly than abstract ones. We were therefore interested in learning whether concrete hints were more effective at helping players complete levels than abstract hints.

To investigate the impact of hints on level completion rates, we analyzed the percentage of players who won levels in which they viewed hints. We found that over 75% of players won the levels in which they viewed hints in all four conditions. Hint *presentation* had a significant main effect on the win rate (χ^2(1,N=58,352)=1439.11, $p<.0001$, V=0.16), with players who see embedded hints winning 89.64% of the time, compared to 78.42% for players with earned hints. This is likely because players in the earned hint conditions only choose to use hints when they are really struggling, since hints have value. Hint content also had a main effect on the win rate (χ^2(1,N=58,352)=250.88, $p<.0001$, V=0.07), with players who see concrete hints winning 86.42% of the time, compared to 81.96% of the time for players with abstract hints. This suggests that the concrete hints are more useful to players in the immediate level than abstract hints.

We also explored how quickly the hints allowed players to win levels. For the players who won the level after viewing a hint, we calculated the number of moves they made between viewing the hint and winning. We hypothesized that more effective hints would help players win more quickly. We found that players in all four conditions made a median of 12 or fewer moves between viewing a hint and winning. A move is defined as any interaction the player makes with the game, such as picking up or placing a piece, so 12 moves corresponds to moving six pieces. Our analysis showed that *presentation* had a significant main effect on the number of moves F(1,49229)=955.98, $p<0.0001$, η^2=0.040). Players with embedded hints won levels more quickly, in a median of 8 moves compared to 10 for players with earned hints. Hint *content* also had a significant main effect on the number of moves (F(1,39996)=295.81, $p<0.0001$, η^2=0.032). Players with concrete hints won more quickly, in a median of 8 levels compared to 10 for players with abstract hints. The interaction *presentation*content* also had a significant effect on the number of moves (F(1,49229)=91.88, $p<0.0001$, η^2=0.033).

These results suggest that concrete hints are more helpful than abstract hints, allowing players to win levels more frequently and with fewer moves. Presentation also impacted the effectiveness of hints, however we believe this is because players with earned hints only chose to use hints when they were really struggling.

CONCLUSION

In this work, we study the impact of four hint system designs on player motivation, performance, and behavior in the educational game *Refraction*. An analysis of 50,000 students who played on the educational website BrainPOP revealed that all four designs negatively impacted performance when compared to a baseline condition without hints. While the size of these effects is small, hints clearly did not improve student learning as we had expected. A factorial analysis of the four hint designs showed that players with earned hints persist longer than players with embedded hints, suggesting that this hint presentation method may improve motivation. Students in the earned hint conditions hoarded their hints, only choosing to use them when they were really struggling. Players with embedded hints viewed hint content much more frequently. We also observed that players played a higher value on concrete hints, using them more often than abstract hints. The concrete hints were also more helpful, allowing players to win levels more often and more quickly than abstract hints.

These results highlight a number of design challenges associated with developing effective hint systems for educational games. However, one limitation of this work is that we cannot fully understand why these hints negatively impacted student performance. The hint content we designed provides information that is relevant to solving each *Refraction* level, and our analysis of hint usage suggests that hints help students solve levels. However, since our hint content did not adapt to the student's partial solution of the level, it is likely that our hints did not always address the student's current confusion. This could have influenced the effectiveness of our hints, particularly if students expected more personalized feedback.

Further work is needed to determine why these hints negatively affected player behavior in *Refraction*, and whether other types of hints could have more benefit. However, it is clear from our results that hint systems do not uniformly improve student performance in educational games. This finding is surprising, particularly given the success of hints in intelligent tutoring systems. However, educational games provide a very different learning environment than cognitive tutors, in which children play voluntarily and expect to be highly engaged. These factors could affect how hints are perceived by students. Further research is needed to understand the full generality of these results, however we expect our findings to translate to other learning environments with similar characteristics. We hope these results will encourage developers to consider the design of hint systems carefully, particularly when engagement is a key component of the learning experience.

ACKNOWLEDGMENTS

We thank creators of *Refraction* for making this work possible. In particular, we would like to recognize Marianne Lee and Brian Britigan who created the art for this study. We also thank Allisyn Levy of BrainPOP for promoting *Refraction* and helping us collect data. This work was supported by the Office of Naval Research grant N00014-12-C-0158, the Bill and Melinda Gates Foundation grant OPP1031488, the Hewlett Foundation grant 2012-8161, Adobe, and Microsoft.

REFERENCES

1. Adams, E. The designer's notebook: Eight ways to make a bad tutorial. *Gamasutra* (2011).
2. Aleven, V., and Koedinger., K. R. Limitations of student control: Do students know when they need help? In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, ITS '00, Springer (Berlin, 2000), 292–303.
3. Aleven, V., McLaren, B., Roll, I., and Koedinger, K. Toward meta-cognitive tutoring: A model of help. *International Journal of Artificial Intelligence in Education 16* (2006), 101–130.
4. Aleven, V., Stahl, E., Schworm, S., Fischer, F., and Wallace, R. Help seeking and help design in interactive learning environments. *Review of Educational Research 73*, 3 (2003), 277–320.
5. Andersen, E., O'Rourke, E., Liu, Y.-E., Snider, R., Lowdermilk, J., Truong, D., Cooper, S., and Popović, Z. The impact of tutorials on games of varying complexity. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 59–68.
6. Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences 4*, 2 (1995), 167–207.
7. Arroyo, I., Beck, J., Woolf, B. P., Beal, C. R., and Schultz, K. Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, ITS '00, Springer-Verlag (London, UK, UK, 2000), 574–583.
8. Arroyo, I., Beck, J. E., Beal, C. R., Wing, R., and Woolf, B. Analyzing students' response to help provision in an elementary mathematics intelligent tutoring system. In *In R. Luckin (Ed.), Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments* (2001), 34–46.
9. BrainPOP. http://www.brainpop.com/.
10. BrainPOP Educators. http://www.brainpop.com/educators/home//.
11. Campbell, C. Casual meets core for a drink: Developing drawn. *Gamasutra* (2010).
12. Conover, W. J., and Iman, R. L. Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician 35*, 3 (1981), 124–129.
13. Corbett, A. T., and Anderson, J. R. Feedback control and learning to program with the CMU LISP tutor. In *Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL* (1991).
14. Gee, J. P. *What Video Games Have to Teach Us About Learning and Literacy*. St. Martin's Press, 2008.
15. Higgins, J. J., and Tashtoush, S. An aligned rank transform test for interaction. *Nonlinear World 1*, 2 (1994), 201–211.
16. Hume, G., Michael, J., Rovick, A., and Evens, M. Hinting as a tactic in one-on-one tutoring. *Journal of the Learning Sciences 5*, 1 (1996), 23–49.
17. Khan Academy. http://www.khanacademy.org/.
18. Knabe, K. Apple guide: A case study in user-aided design of online help. In *CHI '95 Conference companion on Human factors in computing systems*, ACM (New York, NY, USA, 1995).
19. Lau, T., Bergman, L., Castelli, V., and Oblinger, D. Sheepdog: Learning procedures for technical support. In *IUI '04 Proceedings of the 9th international conference on Intelligent user interfaces*, ACM (New York, NY, USA, 2004).
20. Lee, J., Luchini, K., Michael, B., Norris, C., and Soloway, E. More than just fun and games: assessing the value of educational video games in the classroom. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, ACM (New York, NY, USA, 2004), 1375–1378.
21. Lee, M. J., and Ko, A. J. Personifying programming tool feedback improves novice programmers' learning. In *Proceedings of the seventh international workshop on Computing education research*, ICER '11, ACM (New York, NY, USA, 2011), 109–116.
22. Mayo, M. J. Video Games: A Route to Large-Scale STEM Education? *Science 323* (2009), 79–82.
23. Murray, R. C., VanLehn, K., and Mostow, J. Looking ahead to select tutorial actions: A decision-theoretic approach. *International Journal of Artificial Intelligence in Education 14*, 3-4 (2004), 233–278.
24. O'Rourke, E., Butler, E., Liu, Y., Ballwebber, C., and Popović, Z. The effects of age on player behavior in educational games. In *Foundations of Digital Games*, FDG '13 (2013).
25. Ricci, K. E., Salas, E., and Cannon-Bowers, J. A. Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology 8*, 4 (1996), 295–307.
26. Salter, K. C., and Fawcett, R. F. The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation 22*, 1 (1993), 137–153.
27. VanLehn, K. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education 16* (2006), 227–265.
28. Wobbrock, J. O., Findlater, L., Gergle, D., and Higgins, J. J. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (New York, NY, USA, 2011), 143–146.