

Understanding Attention to Adaptive Hints in Educational Games: An Eye-Tracking Study

Cristina Conati · Natasha Jaques · Mary Muir

Published online: 18 October 2013

© International Artificial Intelligence in Education Society 2013

Abstract This paper presents a user study that investigates the factors affecting student attention to user-adaptive hints during interaction with an educational computer game. The study focuses on Prime Climb, an educational game designed to provide individualized support for learning number factorization skills in the form of textual hints based on a model of student learning. We use eye-tracking data to capture user attention patterns on the game adaptive-hints and present results on how user performance, hint timing, and attitude toward getting help all affect the degree to which students attend to hints. We also show that improved attention to hints is a worthwhile goal, at least in Prime Climb, because when they are attended to hints can positively affect a student's performance with the game.

Keywords Educational games · Adaptive hints · Eye-tracking · Attention to hints

Introduction

Educational games are an increasingly popular paradigm embedding pedagogical activities in highly engaging, game-like interactions. However, there is still limited evidence on their pedagogical potential (e.g. de Castell and Jenson 2007; Van Eck 2007; Linehan et al. 2011). One possible reason for these results is that most educational games are designed based on a one-size-fits-all approach, rather than being able to respond to the specific needs of individual students. Hence, researchers have started investigating how to provide individualized support to learning during game play (e.g., Conati and Klawe 2002; Conati and Manske 2009; Easterday et al. 2011; Peirce et al. 2008; Rowe et al. 2011). It can be challenging to provide this support because it requires a careful trade-off between fostering learning and maintaining engagement. In this paper, we aim to provide insight into how to approach this challenge by performing an analysis of which factors affect student attention to the adaptive support provided by Prime Climb, an educational game for number factorization skills. The

C. Conati (✉) · N. Jaques · M. Muir
University of British Columbia, Vancouver, British Columbia, Canada
e-mail: conati@cs.ubc.ca

form of support we focus on is textual *adaptive hints*, namely hints designed to gradually help students through specific educational activities when they have difficulties proceeding on their own¹.

Adaptive textual hints (which we will refer to simply as “adaptive hints” from now on) are one of the most widespread forms of adaptive interventions in Intelligent Tutoring Systems (Woolf 2008). However, there is an increasing body of research showing their possible limitations, from students *gaming the system*, i.e., using the hints to get quick answers from the ITS (see (Baker et al. 2008) for an overview), to *help avoidance*, i.e. students not using hints altogether (e.g., Alevén et al. 2004; Roll et al. 2006). In this paper, we are interested in investigating the latter issue. More specifically, we seek to supply initial answers to the following research questions:

- 1) Do students attend to adaptive hints that they have not explicitly requested?
- 2) If they do, which factors related to game play (e.g., move correctness, interaction time) or student differences (e.g. student’s initial domain knowledge, attitude toward receiving help) affect a student’s tendency to attend to the unsolicited adaptive hints?
- 3) Is attention to these hints useful, i.e. does it impact game performance?

This research has three main contributions to the ITS field. First, while previous work on help avoidance focused on capturing and responding to a student’s tendency to avoid requesting hints (e.g., Alevén et al. 2004; Roll et al. 2006), here we investigate how students react when the hints are provided unsolicited.

A second contribution is that we look at attention to adaptive hints during interaction with an edu-game, whereas most previous work on student usage (or misusage) of hints has been in the context of more structured problem solving activities.

The third contribution of our work is that we use eye-tracking data to study user attention patterns to adaptive-hints. Others have used eye-tracking data to study attention to relevant components of an ITS, e.g. to Open Learner Models (Bull et al. 2007; Mathews et al. 2012), to adaptive animations (Loboda and Brusilovsky 2010) and, most related to our work, to ITS feedback messages (Gluck et al. 2000). The distinguishing feature of our work is that we perform a more detailed analysis of *which factors* affect user attention to adaptive interventions, as well as whether differences in attention affect performance with the system.

This research was first presented in Muir and Conati (2012). Here, we expand on that work by providing detailed information on the methodology we used for collecting, validating and processing the relevant eye-tracking data. We also provide additional results on how students interacted with the game, as well as on how well participants’ subjective evaluations of their game experience are supported by action and gaze data of their in-game behaviors.

In the rest of the paper, we first discuss related work. Then, we describe Prime Climb, the educational game we use as a test bed for this research. Next, we

¹ Whereas there has been work on presenting hints via audio, in this paper we focus on the widespread textual presentation modality.

illustrate the user study we conducted for collecting gaze data, followed by a section describing how this data was processed. After discussing the study results, we conclude with a discussion of possible avenues of future research.

Related Work

Educational Games and Adaptive Feedback

User-adaptive edu-games have received increased attention as a way to improve edu-game effectiveness. Peirce et al. (2008) use both rule-based and probabilistic methods to create an adaptive component in the Elektra game called ALIGN, which provides adaptations in the form of feedback and hinting. A preliminary evaluation of Elektra showed that users felt positively about the game, but didn't provide any reliable results on learning outcomes and flow experience. Easterday et al. (2011) compared the effects of providing tutoring assistance as opposed to basic game-like feedback (e.g. notifying the user of errors and penalties for errors) in Policy World, an edu-game which teaches policy argumentation skills. Tutoring was provided in the form of knowledge-based feedback and required students to correct errors immediately. They found that providing assistance increased competence in pre-debate analysis steps (e.g. issue comprehension, diagrammatic representation and synthesis), as well as perceived interest, but did not affect performance on the actual debate tasks. There has also been extensive research on how to model relevant student cognitive, affective and meta-cognitive states in Crystal Island, a narrative-based adventure game for teaching microbiology (Rowe and Lester 2010; Robison et al. 2009; Sabourin et al. 2012).

Conati and Klawe (2002) propose using adaptive pedagogical agents to provide individualized support to students learning from Prime Climb, the edu-game for number factorization targeted in this paper. They argue that this adaptive support is most effective if based on student models that can capture both student learning (e.g. Manske and Conati 2005), as well as affective reactions (Conati and Maclaren 2009). Conati and Manske (2009) compared a version of Prime Clime with no adaptivity against two versions that differed in the sophistication of the adaptive hints provided, as well as in the accuracy of the student model used to provide them based solely on assessing learning during game playing. While no significant difference in learning was found between the three conditions, they observed that students paid more attention (based on estimating the time the hints were open) to the hints provided by the less sophisticated adaptive game version. They attribute this result to the fact that this version provided fewer and simpler hints (and therefore fewer interruptions). In our work, we extend the work by Conati and Manske (2009) by using a more accurate measure of attention, eye-tracking data, to better understand if and how users attend to Prime Climb's adaptive interventions.

Providing customized feedback or instruction is one of the distinguishing characteristics of Intelligent Tutoring Systems (ITS). One common way this feedback is given is by using adaptive incremental textual hints, especially in the context of problem solving, when a tutor follows a student's individual solution steps and aims

to provide support on the individual steps as the need arises². Incremental hints have been used, for instance, in most model-tracing Cognitive Tutors (Anderson et al. 1995), in the Andes tutoring system (Conati et al. 2002) and in some constrained-based Tutors (Mitrovic 2012), but their effectiveness is in question because of evidence that students can misuse them. There are two main categories of hint misusage that have been investigated so far. The first is *gaming the system*, where we see students repeatedly asking for help or purposely entering wrong answers to get to bottom-out hints, which explicitly tell the student how to perform a problem-solving step and move on. Baker et al. (2008) describe this behavior and compare 6 systems (described in more detail in Alevin et al. (2004), Baker et al. (2004), Beal et al. (2006), Beck (2005), Johns and Woolf (2006) and Walonoski and Heffernan (2006)) that can accurately detect in real-time when a student is exhibiting this behaviour and intervene to reduce it. Baker et al. (2008) and Shih et al. (2010) also found that there are two distinct categories of gaming the system (harmful and non-harmful) based on the learning gain seen by the students, and investigated detectable student behaviors that can differentiate these two different “gaming” outcomes. Goldin et al. (2012) investigate the effectiveness of student-requested hints on local problem solving performance (i.e., individual problem solving steps) as opposed to overall learning, and how effectiveness is mediated by student proficiency and hint type.

The second type of hint misusage uncovered in the literature is *help avoidance*, where students avoid asking for help even when it is needed. Alevin et al. (2004) presented a model that could detect both gaming the system and help avoidance. Roll et al. (2006) embed this model in the Help tutor, an ITS that can generate hints designed to improve students’ help seeking behaviour, in addition to hints that help with the target problem solving activities. They showed that the Help Tutor was able to reduce both help avoidance and help abuse. Unfortunately, this change failed to translate to improved domain learning, possibly because the improvement in help usage was the result of students simply following the Help Tutor’s advice rather than learning the principles of good help seeking behavior themselves.

Little work has been done on understanding if and how students process adaptive hints that they have not elicited from the system, however Roll et al. (2006) suggest that students often ignore these hints. A similar hypothesis was brought forward by Conati and Manske (2009) based on preliminary results on student attention to hints in Prime Climb; however, attention was not measured using an eye-tracker but was calculated based on the time the hint was open on the screen.

Eye-Tracking in ITS

There has been a rising interest in using eye-tracking in ITS research, following two main directions.

² An alternative approach that has been explored in the literature is to devise tutors that engage students in a tutorial dialogue to provide the necessary feedback or instruction, by leveraging natural language processing techniques. These tutors are outside the scope of this paper, because by their nature they encompass a different set of possible issues than those raised by the type of incremental textual hints addressed here.

One research direction investigates eye-tracking data as a direct source of information for student modelling and personalized instruction. Eye-tracking data has so far been used to direct the adaptive behavior of an ITS in real-time by capturing simple gaze patterns indicating attention (or lack thereof) to relevant interface elements, including: an available pedagogical agent and target didactic material (Wang et al. 2006; D’Mello et al. 2012); words during a reading task (Sibert et al. 2000); available tutor feedback (Anderson 2002). There has also been work on leveraging gaze data in student models that capture higher level student states such as learning (Kardan and Conati 2012, 2013), meta-cognition in terms of self-explanation (Conati and Merten 2007), and affect in terms of motivation (Qu and Johnson 2005). While these gaze-enhanced models of higher level states have been validated in terms of accuracy, so far they have not been integrated in an ITS.

The second research direction in leveraging gaze data in ITS attempts to understand relevant student behaviors and processes through off-line analysis of the gaze data, by investigating how students attend to specific elements of an ITS interface. Gluck et al. (2000) performed off-line analysis of eye-tracking data obtained with a simplified version of an ITS for algebra, and showed that this data could quite reliably disambiguate domain-specific strategies even when they led to the same problem-solving steps. They also showed that students did not attend to as many as 40 % of the system’s error feedback messages, although they did not provide reasons for this effect. Muldner et al. (2009) investigated whether pupil size can provide information on a student’s meta-cognitive and affective states while working with a tutor that supports analogical problem solving. They found that users had significantly larger pupil size when expressing positive vs. negative affect, as well as when engaging in self-explanation as opposed to other forms of reasoning. Bull et al. (2007) and Mathews et al. (2012) used gaze data to understand if and how students attend to different visualizations of their student model (or Open Learner Model, OLM). OLMs have been investigated as a way to aid learning by scaffolding reflection and self-assessment. The main results of this work indicate that different ways to visualize an OLM trigger attention to different aspects of a learner’s performance, e.g. misconceptions vs. levels of knowledge. Loboda and Brusilovsky (2010) used off-line gaze analysis to understand the effect of adaptation in cWADE, an ITS that supports expression evaluation in the C programming language through explanatory animated visualizations. The user-adaptive version of cWADE adapts the speed of the animations to a student’s progress through the available material (i.e., more progress results in a faster pace of animations). An exploratory eye movement analysis showed the adaptive version engaged students more and attracted their attention more. The work described in this paper extends the use of off-line analysis of gaze information to understand not only if and how users attend to an educational game’s adaptive interventions, but also which factors may affect these behaviors.

The Prime Climb Educational Game

In Prime Climb, students practice number factorization by pairing up to climb a series of mountains. The current version of Prime Climb is Web based, so players can login

remotely. Each mountain is divided into numbered hexagons (see Fig. 1) and players must move to numbers that do not share common factors with their partner's number, otherwise they fall. Players begin on a hexagon at the base of the mountain that is randomly selected by the game. Since each player cannot move to a square that is more than two hexagons away from the partner, the two players must cooperate to reach the top of the mountain. Each player can make more than one move before turning the control to the other player. The object of the game is to reach the top of as many mountains as possible, out of a total of 12.

To help students, Prime Climb includes the Magnifying Glass, a tool that allows players to view the factorization for any number on the mountain in the device at the top-right corner of the interface (see Fig. 1). Prime Climb also provides individualized textual hints, both on demand and unsolicited. Unsolicited hints are provided in response to student moves and are designed to foster student learning during game playing by (i) helping students when they make wrong moves due to lack of factorization knowledge; (ii) eliciting reasoning in terms of number factorization when students make correct moves due to lucky guesses or playing based on game heuristics.

Prime Climb relies on a probabilistic student model to decide when incorrect moves are due to a lack of factorization knowledge vs. distraction errors and when good moves reflect knowledge vs. lucky guesses. The student model assesses the student's factorization skills for each number involved in game playing, based on the student's game actions (Manske and Conati 2005). Prime Climb gives hints at incremental levels of detail, if the student model predicts that the student doesn't know how to factorize one of the numbers involved in the performed move. The hint sequence includes a *tool* hint that encourages the student to use the magnifying glass tool to see relevant factorizations ("You can use the magnifying glass to see the factors of the number you clicked on"). If the student needs further help, Prime Climb

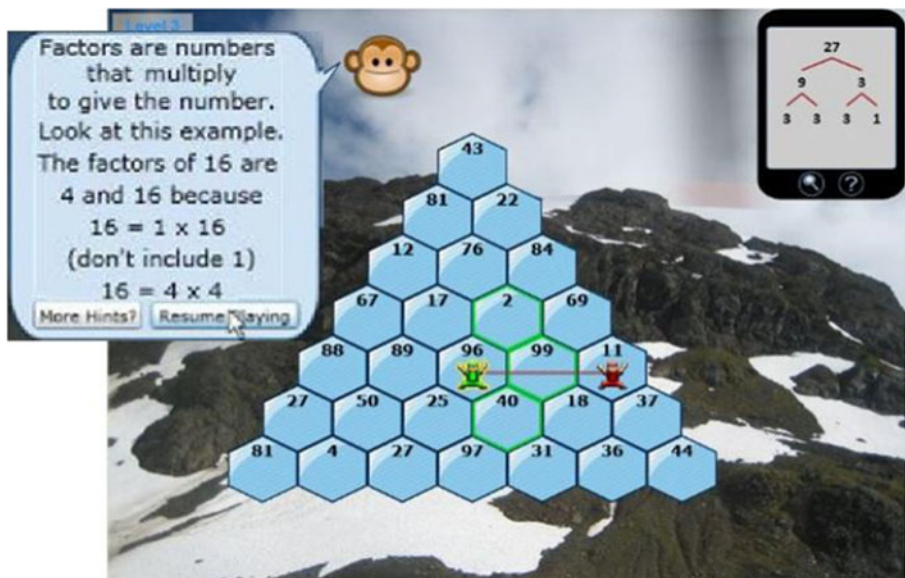


Fig. 1 The prime climb interface

gives *definition* hints designed to re-teach “what is a factor” via explanations and generic examples (e.g., see Fig. 1). There are two different factorization definitions: “*Factors are numbers that divide evenly into the number*” and “*Factors are numbers that multiply to give the number*”. The game alternates which definition to give first and presents the second the next time it needs to provide a definition hint. The examples that accompany the definitions change for every hint and are designed to help illustrate the given definitions while still leaving it to the student to find the factorization of the numbers relevant to the performed move. Finally, Prime Climb provides a *bottom-out* hint giving the factorization of the two numbers involved in the move (e.g., “*You fell because 84 and 99 share 3 as a common factor. 84 can be factorized as...*”). The basic wording of the *bottom-out* and *tool* hints does not change. Students can access the next available hint by clicking on a button at the bottom of the current hint (See Fig. 1). Otherwise, hints are given in progression as the student model calls for a new hint. After the entire sequence of hints has been given, it starts again at the beginning with another *tool* hint. A hint is displayed until the student selects to access the next hint or to resume playing (by clicking a second button available at the bottom of the hint).

It should be noted that the Prime Climb *bottom-out* hints focus on making the student understand her previous move in terms of factorization knowledge; they never provide explicit information on how to move next. Thus, the Prime Climb hints are less conducive to a student gaming the system than *bottom-out* hints giving more explicit help (e.g. Baker et al. 2008). As a matter of fact, previous studies with Prime Climb show that students rarely ask for hints. Most of the hints the students see are unsolicited.

User Study on Attention to Hints

Participants and Study Design

We recruited 13 participants (six female) from grades 5 and 6 (six participants in grade 5). Participants came to our research laboratory to participate in the study. Recruitment was conducted through flyers distributed in a local school, at youth outreach events held by our department, as well as at local sports camps. Each child was compensated with a \$10 gift card for a local children’s bookstore.

Prime Climb was run on a Pentium 4, 3.2 GHz machine with 2GB of RAM, with a Tobii T120 eye-tracker acting as the primary screen. The Tobii T120 eye-tracker is a non-invasive desktop-based eye-tracker embedded in a 17" display. It collects binocular eye-tracking data at a rate of 120 Hz. In addition to the eye-tracking data, it collects all keystrokes and mouse clicks made. It also collects video data of the user’s face.

Participants started by completing a pre-test which tested their ability to identify the factors of individual numbers (16 numbers tested overall) and identify the common factors between two of the numbers from the first part (5 pairs of numbers tested). After completing the pre-test, participants underwent a familiarization and calibration phase with the Tobii eye-tracker (described in more detail below). Next they played Prime Climb with an experimenter playing as their partner. Following the

protocol we adopted in previous studies, the experimenter was instructed to play as neutrally as possible, trying to avoid making mistakes (although mistakes did happen on some of the mountains with larger numbers) and to avoid leading the climb too much.

Students played the game until they climbed all mountains. During this period, we also had a second investigator observing a secondary monitor to detect eye-tracking data collection issues as they occurred. Finally, participants took a post-test analogous to the pre-test and completed a questionnaire to obtain their subjective feedback on their game experience.

Assessment Tools

The pre-test and post-test used for knowledge assessment in this study were the same used by Manske and Conati (2005). They consist of two sections. The first section includes 16 questions on the factorization of a number. Students need to select the factors of the number from a list of options, where multiple options may be correct. The second section consists of 5 multiple-choice questions gauging knowledge of the concept of common factors. As with the factorization questions, students select the common factors of two numbers from a list of possibilities. Ten of the 16 numbers found in the factorization section also appear in the common factor questions, to discern when errors in the latter are due to lack of understanding of the concept of common factors as opposed to missing knowledge on how to factorize the numbers involved. The test was marked by giving one mark for each correctly circled response and taking one mark away for each incorrectly circled response. As a result, it is possible for a student to receive a negative score on the test if they select more incorrect answers than correct ones. The highest possible mark on this test was 31, while the lowest possible mark was -105.

After the post-test, each student was given a written questionnaire also based on questionnaires used in previous Prime Climb studies (Manske and Conati 2005). The questionnaire included items tapping, among other things, a participant's attitude towards receiving help and perception of the Prime Climb agent's adaptive interventions. Items were scored on a Likert scale where 1 indicated "Strongly Disagree" while 5 indicated "Strongly Agree".

Arrangements to Minimize Gaze-Data Loss

The collection of eye-tracking data with non-intrusive devices like the Tobii is susceptible to error due to excessive head movement of the subject, as well as participants looking away from the screen, blinking, resting one's head on one's hands, and other actions which can cause the eye-tracker to lose track of their eyes. These sources of error can be especially relevant when dealing with children involved in a game-like activity. In this section, we describe the two methods we used to minimize loss of eye-gaze data in our study, which we believe may be useful for other researchers who plan to embark on similar research.

The first method consisted in having a second experimenter observe and respond to eye gaze traces in real-time using a secondary 17" monitor (Tobii experimenter view). When the experimenter detected loss of eye-gaze data, the subject was asked to

shift position until the eye-tracker started collecting data properly again. We are aware that these interventions may somehow have affected students' overall attention to the game, but because the experimenter never made any comments on what students should be paying attention to on the screen, we don't expect that the interventions changed how students attended to the hints and why.

Secondly, each participant was exposed to an eye-tracker familiarization phase, designed to make participants aware of how their actions affected the ability of the eye-tracker to collect data. The familiarization phase relies on the Tobii's pre-calibration display (shown in Fig. 2), which allows participants to see their eyes being captured (the two white dots in Fig. 2). The color of the bar at the bottom of the display gives feedback on the quality of the data being collected: green indicates that the eye-tracker is able to capture data; yellow indicates that there are difficulties with detecting eye gaze but data is still being collected; red indicates that the eye-tracker is unable to capture the eye. The bar on the right side of the display shows how far the subject's eyes are from the monitor. When the eyes get too close or too far, the eye-tracker encounters difficulty in identifying the eyes and therefore the bottom bar will turn yellow and then red.

We introduced a familiarization phase based on an observation made during pilot runs of the study: a participant who spent more time looking at Tobii's pre-calibration

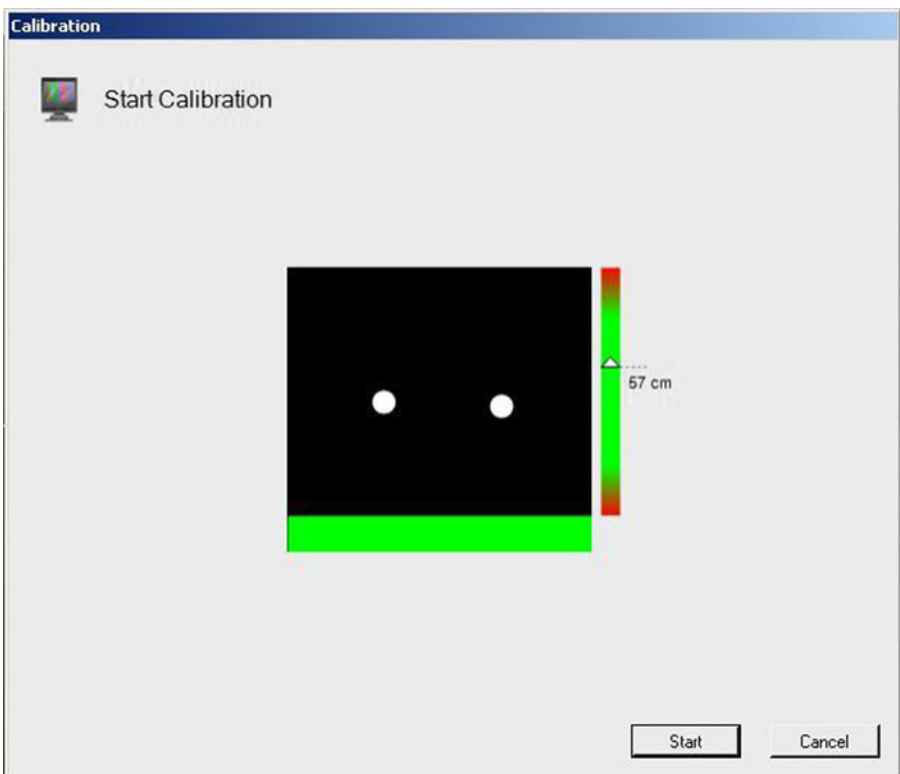


Fig. 2 Pre-calibration screen which allows participants to become more aware of the capabilities of the eye-tracker

screen seemed to be more mindful than other pilot participants of her position in front of the screen during the game and generated excellent gaze data. Thus, during the familiarization phase, participants were instructed to “play” with the pre-calibration display by moving in their seat to see how much they could move before the eye tracker could not detect their gaze. Participants were also asked to rest their head on their hand and observe the effect on gaze tracking accuracy. In addition to making participants more aware of how their movements affected the eye-tracking process, this phase also made it easier for them to understand what to do when asked to shift their position during game play because the experimenter detected that the eye-tracker had stopped collecting data.

Processing Eye-Tracker Data

Eye-gaze information is provided in the form of *fixations* (i.e., eye-gaze remains at one point on the screen) and *saccades* (i.e., eye-gaze moves quickly from one fixation point to another), which we can use to derive attention patterns. Both fixations and saccades are made up of multiple *samples* from the eye-tracker. Fixations are derived from consecutive samples that relate to the eye focusing on a single place on the screen; saccades are derived from all consecutive samples between two fixations. For each recorded sample, the Tobii eye-tracker stores its coordinates, quality (e.g., whether that sample had data for one or both eyes or none) and pupil size. In addition, it includes the length and coordinates of the fixation with which that sample is associated. Tobii also generates a summary of the recorded fixations i.e., the fixation timestamp, duration and coordinates.

As the game interaction can be long (up to 65 min) and dynamic, we need to divide the game play into portions that are easier to analyze. As we are interested in investigating if and how students are using the hints provided to them during game play, we chose to focus on the portions during which a hint is available to the player. These short time periods provide a fairly static view of the game as players are unable to make any moves during this time frame. This method simplifies data analysis because we don't have to account for objects moving during the interaction.

Eye-Tracking Measures

To analyze the attention behaviors of our study participants with respect to the received adaptive hints, we define an area of interest (*Hint AOI*) that covers the text of the hint message. We adopt a standard metric used in eye-tracking research to measure overall attention (Goldberg and Kotval 1999; Goldberg and Helfman 2010), namely *total fixation time* (e.g., overall time a subject's gaze rests on the Hint AOI for each displayed hint). It should be noted that because in Prime Climb students are actually unable to continue playing while a hint is being displayed (they must first manually dismiss it), the amount of time the hint is open on screen is correlated with *total fixation time* ($r=.711$, $N=484$, $p=.01$). *Total fixation time*, however, gives a more precise picture of student attention to hints. For instance, it helps distinguish the situation depicted in Fig. 3, (where student attention while the hint is open is mostly on the hint itself) from the situation in Fig. 4 (where student attention while the hint is

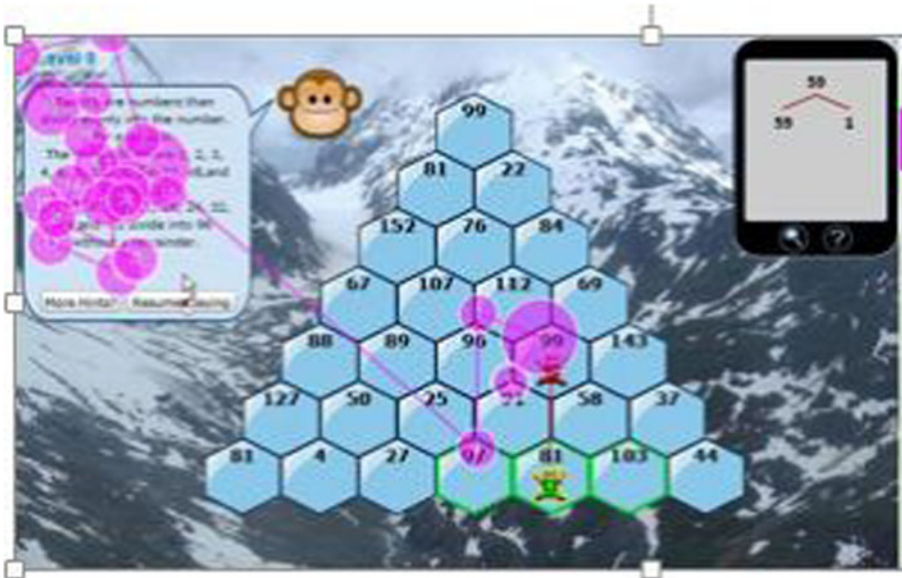


Fig. 3 Attention on the hint

open is mostly away from the hint). The circles in the figures represent fixation points, and their size expresses duration.

Total fixation time is also a more powerful source of information than *hint display time* for a student model to detect situations in which students try to game the system. If included in a student model, *Hint display time* could determine if the student

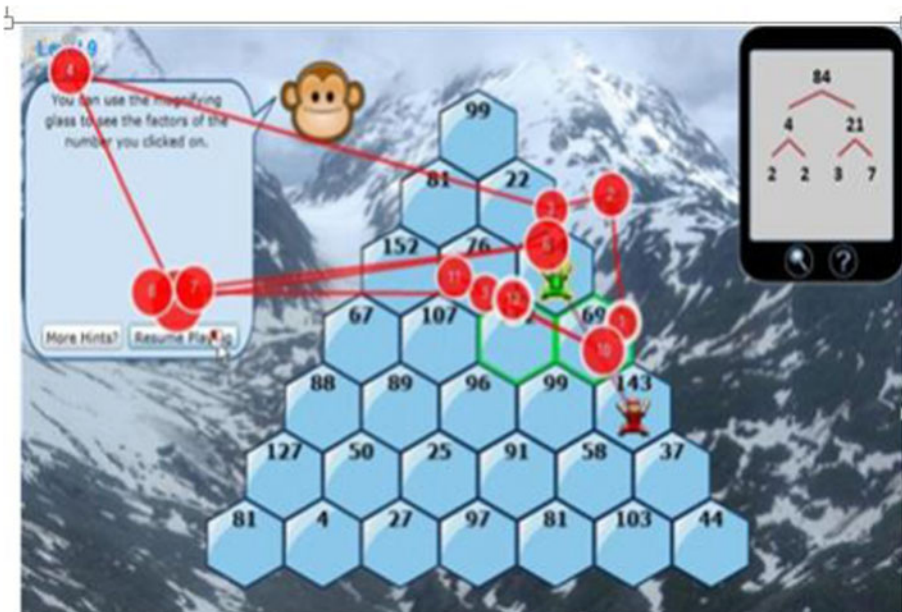


Fig. 4 Attention away from the hint

missed reading a hint before asking for a new one only when the student asks for a new hint too quickly. But it could not handle a situation in which a hint is open for a while but the student is barely looking at it, like in Fig. 4, before asking for the next hint. Eye tracking allows this behaviour to be detected. The limitations of *hint display time* would become even more of a problem when students are not required to explicitly dismiss a hint before they can continue playing, because it would make it easier for students to ignore the hint while it is open.

One weakness of *Total fixation time* is that it does not provide detailed information on how a hint is actually processed, because it cannot differentiate between a player who stares blankly at a hint vs. one who carefully reads each word. Furthermore, it is not ideal to compare attention to the different types of hints in Prime Climb because they have different lengths on average (15 words for *tool* hints; 17 words for *bottom-out* hints; 36 words for *definition* hints). Thus, in our analysis of attention to the Prime Climb hints we also use the ratio of fixations per word (*fixations/word* from now on). This metric is independent of hint length and gives a sense of how carefully a subject scans a hint's text. *Fixations/word* is 1 when the subject fixates once on each word of the text and decreases when a reader starts skipping words.

Finally, in order to assess how quickly students are able to reallocate attention to hints during gameplay, we include *time to first fixation* as our last measure. *Time to first fixation* is the time in milliseconds it takes for a student to fixate on the Hint AOI after the hint is first displayed on the screen. Lower values reflect a quicker shift in attention to the hint. Using these three measures, we will evaluate general attention to a hint, the amount of care the students are taking in reading the hint text, and the ability of the hint to attract students' attention.

We also considered including additional eye tracking measures (as done, for instance, by Kardan and Conati (2012) and Bondareva et al. (2013)). The *scan path length* measures the Euclidean distance between fixations in pixels and can be used to compare scanning behavior where longer scan paths can indicate less efficient or less careful scanning. For instance, Goldberg and Kotval (1999) found that when they compared user's scan path length on a good and poor interface, the poor design led to longer scan paths. However, in our case, scan path length comparison between hints would not be ideal because, like *fixation time*, it would be affected by the varying length of our hints. The *number of transitions* between the Hint AOI and other salient elements of the game interface can also be useful in looking at the specific attention patterns involving hints, e.g. looking back and forth between the Hint AOI and the mountain AOI while reading a hint. However, because this is an exploratory study with a limited number of participants, we cannot include too many dependent measures in our analysis without losing power (see below for details on the statistical models used). We felt that the three measures described above are the best set to start with for a first exploration of the factors that affect attention to hints.

Data Validation

Despite the measures we took to reduce participants' movements that could interfere with eye-tracking, data loss cannot be entirely eliminated, especially when dealing with children, who generally cannot sit in the same position for long. This data loss is in the form of samples marked as invalid by the eye-tracker, which may include

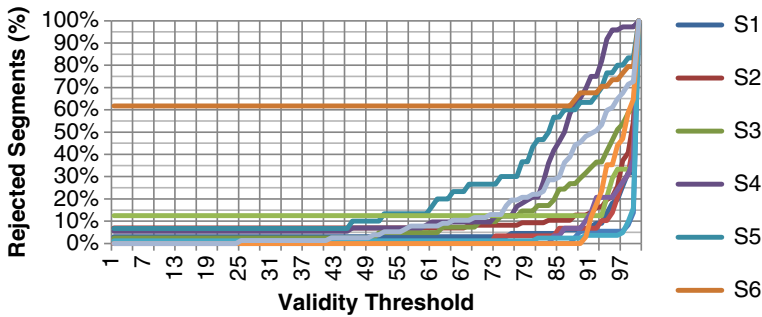


Fig. 5 Percentage of segments (hints) discarded for different threshold values

samples collected while the participant is looking off-screen. Since our analysis focuses only on participants' attention patterns when hints are displayed on the screen, we need to make sure that we have sufficient valid gaze information during these *segments* of interest (i.e., that we have *valid segments*), rather than during the overall interaction with the game.

A standard method to define a segment of tracked gaze data as valid is to ascertain whether the proportion of valid samples in the segment is greater than a preset threshold. To determine the threshold, we plotted the percentage of segments that would be rejected per participant, for different threshold values.

Figure 5 shows the percentage of segments that would be excluded from analysis for each subject given different possible threshold values (where the higher the threshold is set, the more segments will be rejected). As the figure shows, one subject (S6) has a severe loss of data (over 60 % of the segments having no data collected at all), and was thus excluded from data analysis. Based on the remaining participants, we choose to set 75 % as our threshold for excluding a segment from analysis (i.e. if a segment contains less than 75 % of the possible eye gaze samples, it would be considered invalid). At this threshold, we find that the majority of the participants have a low percentage of rejected segments (less than 20 %), while the worst subject (S5) still has 70 % valid segments. Therefore, this threshold is a good compromise between retaining a sufficient number of hint segments for data analysis and ensuring that the gaze data over these segments is of good quality.

Gaze data was processed using an earlier version of EMDAT, an open source package for off-line and on-line gaze data analysis developed in our lab³.

Results

In this section, we first present summary statistics on participants' performance with Prime Climb, as derived from pre/post test scores and from the analysis of the available action logs. Next, we present an analysis based on gaze data aimed at answering the research questions we introduced in the discussion section:

- 1) Do students attend to adaptive hints that they have not explicitly requested?

³ <http://www.cs.ubc.ca/~skardan/EMDAT/index.html>

- 2) If they do, which factors related to game play (e.g., move correctness, interaction time) or student differences (e.g. student's initial domain knowledge, attitude toward receiving help) affect a student's tendency to attend to the unsolicited adaptive hints?
- 3) Is attention to these hints useful, i.e. does it impact game performance?

Finally, we report results on how well some of the students' observed action and gaze patterns match with their subjective assessment of their game experience, measured from their responses to the study post-questionnaire.

Descriptive Statistics on Game Play

The study game sessions lasted 33 min on average ($SD=15$). There was no improvement from pre- to post-test performance, with participants scoring an average of 74 % ($SD=31$ %) in the pre-test, an average of 72 % ($SD=31$ %) on the post-test and an average percentage learning gain of -0.02 ($SD=0.06$). It should be noted that six of the participants performed worse in the post-test than in the pre-test, but in general when a student went down in score, it was usually due to missing one of the factors of a number that they got correct on the pre-test rather than giving an incorrect answer (i.e. circling a number that is not a factor). This suggests that player fatigue might have contributed to the poor results, causing students to not take as much care on the post-test as they could have.

Consistent with previous Prime Climb studies, students rarely asked for help. One student asked for four hints, two students asked for hints twice, and two other students requested one hint, for a total of 8 requested hints. Prime Climb, however, generated a total of 476 unsolicited hints; an average of 51 hints per player ($SD=23$), with an average frequency of 37 s ($SD=44$). Thus, lack of system interventions can be ruled out as a reason for lack of learning. If anything, it is possible that the hints happened too frequently, resulting in reduced student attention (we will discuss this point in more detail in the next section). It should be noted that the two participants who received the most frequent hints (S13 and S11) were the participants with the lowest pre- and post-test score, indicating that lack of knowledge was a factor in these participants receiving frequent hints.

Participants made an average of 17.5 % incorrect moves ($SD=4.5$). Less than half of these ($M=37.7$ %, $SD=11.4$) were made after receiving a hint. Figure 6 shows the

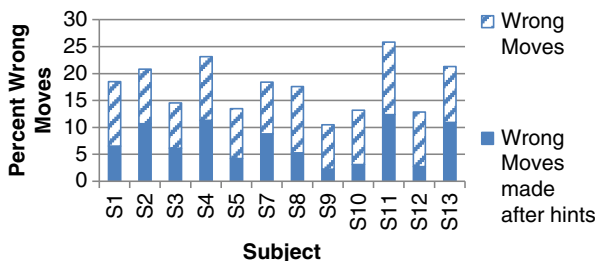


Fig. 6 Wrong moves made by the subjects including the proportion of wrong moves that were made after receiving a hint (solid blue portion)

number of incorrect moves and the percentage of these moves that followed a hint for each participant. A move follows a hint if it is the next action made after a hint has been displayed.

On average, participants used the magnifying glass 18 times ($SD=20$). This large standard deviation indicates that while some players used the tool frequently, others used it very infrequently. In fact, 2 participants did not use the magnifying glass tool at all, while 3 more only used it 5 times. Participants followed a *Tool* hint (which directs them to make use of the magnifying glass tool) 20 % of the time on average, with a high standard deviation (18). This number ranged as high as 56 %, indicating that, at least for some participants, the *Tool* hints were successful in triggering the use of the magnifying glass. The *Tool* hint is the only hint for which we can provide this type of statistic, because it is the only one that suggests an explicit action available in Prime Climb. What we can do for all hints, however, is to use the gaze data collected during the study to ascertain whether participants paid attention to them and under which circumstances, as we discuss in the next section.

Attention to Hints and Factors Affecting It

As we mentioned in the previous section, Prime Climb generated unsolicited hints frequently, raising the question of whether the frequent hints possibly interfered with game playing and led students to ignore them.

In order to answer this question we first compared average fixation time on each hint type with the *expected reading time* (calculated using the 3.4 words/second rate from Just (1986)), which is the time it would take an average-speed reader to read the hint. Figure 7 shows that average fixation time is much shorter than *expected reading time*. On the other hand, the high standard deviation in all three measures shows that there is a great deal of variability in reading time, depending on the hint. Students are taking longer to read certain hints, indicating a trend of selective attention. In the rest of this section, we investigate which factors influenced a student's decision to attend a hint or not.

One obvious factor is whether the hints generated were justified, i.e., whether the probabilistic student model that drives hint generation is accurate in assessing a student's number factorization knowledge. We can only answer this question for the numbers in the game that were also included in the pre- and post-tests, which are about 10 % of all the numbers covered in Prime Climb. The model *sensitivity* on test numbers (i.e., the proportion of actual positives

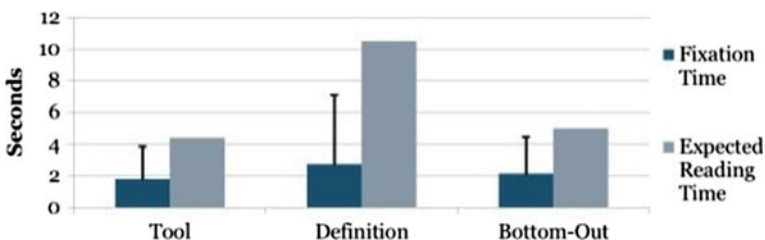


Fig. 7 Average fixation time for prime climb hint types

which are correctly identified as such) is 89 %, indicating that the model generally did not underestimate when a student knew a test number and thus it is unlikely to trigger hints when they were not needed. It should be noted, however, that whereas for test numbers the student model is initialized with prior probabilities derived from test data from previous studies, for all the other numbers in Prime Climb, the model starts with generic prior probabilities of 0.5. Thus, the model's assessment of how student factorization knowledge on these numbers evolved during game play was likely to be less accurate than for test numbers and may have generated unjustified hints.

Bearing this in mind, we looked at the following additional factors that may influence student attention to hints in our dataset.

- *Move Correctness* indicates whether the hint was generated in response to a correct or to an incorrect move.
- *Time of Hint* sets each hint to be in either the first or second half of a student's interaction with the game, defined by the median split over playing time.
- *Hint Type* reflects the three categories of Prime Climb hints: *Definition*, *Tool* and *Bottom-out*.
- *Attitude* reflects student's general attitude towards receiving help when unable to proceed on a task, based on student answers to a related post-questionnaire item ("I want help when I am stuck"), rated using a Likert-scale from 1 (strongly disagree) to 5 (strongly agree). To compensate for the limited number of data in each category due to our small dataset, we divided these responses into three categories: *Want help*, *Neutral* and *Wanted no help*, based on whether the given rating was greater than, equal to, or less than 3, respectively.
- *Pre-test score* represents the student percentage score in the pre-test as an indication of student pre-existing factorization knowledge.

In order to utilize the eye tracking data on all hints collected during the course of the study, we chose to use a Mixed Effects Model analysis. Unlike a standard General Linear Model, Mixed Effects Models can handle correlated data, as in the case when there are repeated measures from the same subjects (Wainwright et al. 2007). Mixed Effects Models also provide advantages over the traditional Repeated Measures ANOVA because they are more robust to missing data, which is ideal for noisy eye tracking data (Toker et al. 2013). We ran a mixed model analysis for each of the three dependent measures described above; *total fixation time*, *fixations/word* and *time to first fixation*. Each mixed model is a 2 (*Time of Hint*) by 3 (*Hint Type*) by 2 (*Move Correctness*) by 3 (*Attitude*) model, with pre-test score as a covariate. We report significance at the .05 level after applying a Bonferroni correction to adjust for familywise error. In order to calculate the R^2 effect sizes of each fixed effect in the model, Snijders and Bosker (1994) suggest using a measure based on the increase in between- and within-subjects variance accounted for by each factor, over the null model. We use this method, although it suffers from the shortcoming that the R^2 value can sometimes become negative, if the within-subjects variance is decreased at the cost of increasing the between-subjects variance (Snijders and Bosker 1994). An R^2 of .01 is considered small, while .09 is considered a medium-sized effect and .25 is considered a large effect.

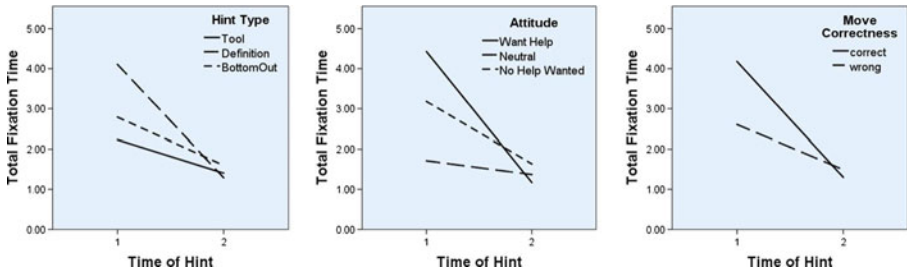


Fig. 8 Interaction effects for total fixation time involving time of hint

Factors That Affect Attention to Hints Measured by Total Fixation Time

In our first model, we used total fixation time as the dependent measure. We found the following interaction effects⁴:

- *Time of Hint* and *Hint Type*, $F(2,662.299)=8.550$, $p<.005$, $R^2=.097$ (see Fig. 8, left). Fixation time drops for all hint types between the first and second half of the game. The drop, however, is statistically significant only for *definition* hints, suggesting that these hints became repetitive and were perceived as redundant, despite the inclusion of varying examples that illustrate the definitions.
- *Time of Hint* and *Attitude*, $F(2,662.300)=4.920$, $p=.024$, $R^2=.067$ (see Fig. 8, middle). In the first half of the game, those with a neutral attitude ($M=1.71$, $SD=2.19$) had a significantly lower total fixation time than those who wanted help ($M=4.42$, $SD=5.63$), $t(179)=5.417$, $p<.005$ and those who did not want help ($M=3.18$, $SD=3.71$), $t(138)=2.616$, $p=.030$. There is a non-significant trend of higher attention for students who wanted help compared to students who do not want help. For both of these groups, fixation time dropped significantly in the second half of the game ($M=1.17$, $SD=1.33$, $t(189)=5.460$, $p<.005$ for wanting help students; $M=1.64$, $SD=2.22$, $t(180)=3.443$, $p<.005$, for those who do not want help), but did not change significantly for players with a neutral attitude.
- *Time of Hint* and *Move Correctness*, $F(1,662.300)=10.306$, $p=.003$, $R^2=.093$ (see Fig. 8, right). Initially, students attend more to hints occurring after a correct move, although this effect disappears by the second half of the game. We find the increased attention to hints following correct moves somewhat surprising, because we would have expected these hints to be perceived as redundant and thus attended to less than hints following incorrect moves. It is possible however, that the very fact that hints after correct moves were unexpected attracted the students' attention and this surprise effect faded as the game progressed.

⁴ We also found significant main effects for *Time of Hint* and *Hint Type* but we don't discuss them in detail because they are further qualified by the detected interactions.

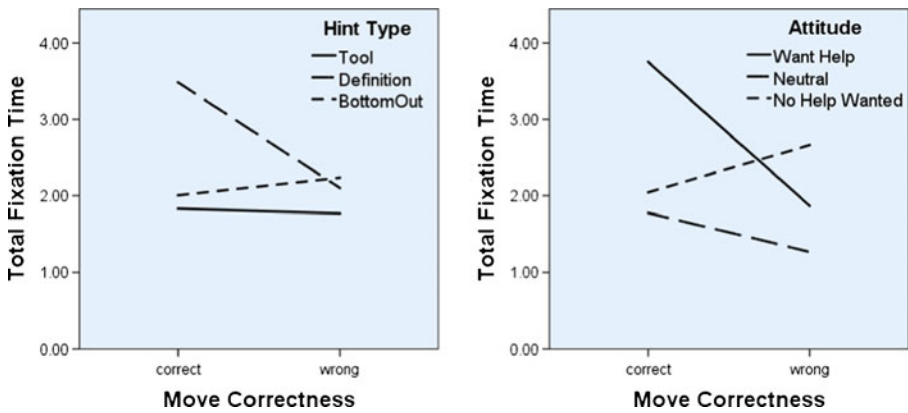


Fig. 9 Interaction effects for total fixation time involving move correctness

- *Hint Type* and *Move Correctness*, $F(2,662.300)=5.028$, $p=.021$, $R^2=.023$ (see Fig. 9, left). Players had significantly higher fixation time on *definition* hints caused by correct moves than on those caused by incorrect moves⁵. There were no statistically significant differences between fixation times for correct vs. incorrect moves for the other two hint types.
- *Attitude* and *Move Correctness*. $F(2,662.299)=9.877$, $p<.005$, $R^2=.004$ (see Fig. 9, right). For students who want help, fixation time is higher for hints that occur after a correct move, mirroring the trend of increased attention to these unexpected hints. This trend is reversed for students who do not want help; they attend more to hints occurring after an incorrect move. This could be because the students who want help are willing to use these hints to confirm that they are on the right track, especially if they guessed at the correct move, but students who do not want help only wish to use hints when they are failing to move correctly on their own.

Factors That Affect Attention to Hints Measured by Fixations/Word

To gain a better sense of how students looked at hints when they were displayed, we ran a second Mixed Model with the same independent measures described above (*Time of Hint*, *Hint Type*, *Move Correctness*, *Attitude* and *pre-test scores*) and fixations/word as the dependent measure. It should be noted that *total fixation time* and *fixations/word* are correlated, $r=.643$, $N=484$, $p<.001$, as we would expect if the students tend to read the content of the hints. *Fixations/word* is included in order to account for the effect of hint length on the previously reported findings. We found three main effects:

- *Move Correctness*, $F(1,605.148)=14.103$, $p<.005$, $R^2=.013$ (see Fig. 10, left). Fixations/word was higher for those hints following a correct move, consistent with the surprise effect of hints after correct moves found in the previous section.

⁵ There is also a significant difference between fixation time on *definition* hints after correct moves and the other two type of hints after correct moves, but this difference is likely an effect of definition hints being longer, as we discussed in “User Study on Attention to Hints”.

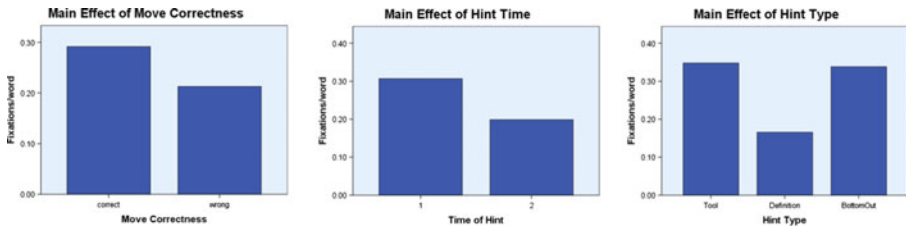


Fig. 10 Main effects for fixations/word

- *Time of Hint*, $F(1,605.150)=22.829$, $p<.005$, $R^2=.030$ (see Fig. 10, middle). Fixations/word was lower for hints occurring in the second half of the interaction, consistent with the pattern of decreased total fixation time in the second half of the game described in the previous section.
- *Hint Type*, $F(2,605.148)=36.064$, $p<.005$, $R^2=.079$ (see Fig. 10, right). *Definition* hints (Avg. 0.17, $SD=0.22$) had a statistically significantly lower fixations/word than either *Tool* ($M=0.35$, $SD=0.38$) or *Bottom-out* hints ($M=0.34$, $SD=0.32$), possibly because after a few recurrences of definition hints students tended to stop reading them fully, since the definition portion does not change.

We also found an interaction effect involving *Move Correctness* and *Hint Type*, $F(2,605.150)=12.276$, $p<.005$, $R^2=.122$ (see Fig. 11). Fixations/word on *Bottom-out* hints drops significantly between those given after a correct move and those given after an incorrect move. This result confirms the positive effect that *Move Correctness* seems to have on attention to hints found in the previous section for *definition* hints. Here, the effect possibly indicates that students are scanning *Bottom-out* hints after correct moves carefully in order to understand why they are receiving this detailed level of hint when they are moving well.

Factors That Affect Attention to Hints Measured by Time to First Fixation

A mixed model with *time to first fixation* as the dependent variable and the same factors/co-variates as the previous models showed only one significant main effect,

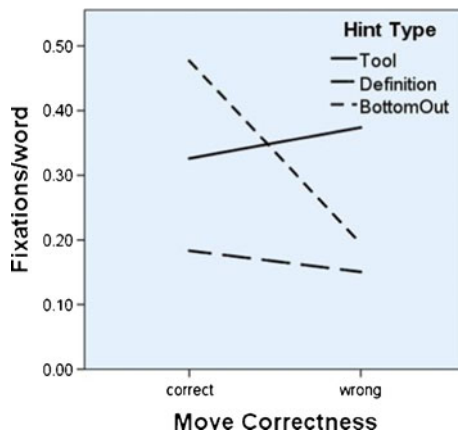


Fig. 11 Interactions effects for fixations/word

related to *Move Correctness*, $F(1,466.672)=5.823$, $p=.048$, $R^2=.010$ (see Fig. 12). This result indicates that it takes less time for students to fixate on hints that occur after a correct move and is consistent with the “surprise” effect of correct hints suggested by the results on fixation time and fixations/word.

Factors That Affect Attention: Discussion

All of the factors that we explored, except for *Pre-test Scores (Time of Hint, Hint Type, Attitude, Move Correctness)* affected to some extent attention to the Prime Climb hints, providing insight into how adaptive hints can be delivered more effectively.

We found, for instance, that attention to hints decreases as the game proceeds and the drop is highest for *definition* hints, suggesting that these hints are too repetitive and should be varied and improved in order to remain informative and engaging as the game proceeds. We also found that hints given after a correct move tend to elicit attention more than those given after an incorrect move, likely because they are perceived as surprising. This is an interesting result, because in Prime Climb hints are always provided when there is an indication from the student model that the student does not fully grasp the factorization knowledge underlying her last move, even if the move is correct. When we devised this strategy, we were concerned that receiving hints after a correct move may be awkward for a student and may result in the student ignoring the hints. However, our results indicate that this is not necessarily the case, at least in the first part of the game and for students with a positive attitude toward receiving help. It is possible that hints occurring after correct moves were acting like positive feedback for these students (Mitrovic and Martin 2000). Positive feedback (e.g. statements like “Correct!” or “Well done”) is usually given after a correct answer and allows a student to reduce uncertainty about an answer by confirming that it is correct (Barrow et al. 2008). This is especially useful if the student was unsure or guessing at the correct answer, as is likely the case when the Prime Climb student model triggers a hint after a correct move. Thus, our results show that it is possible to

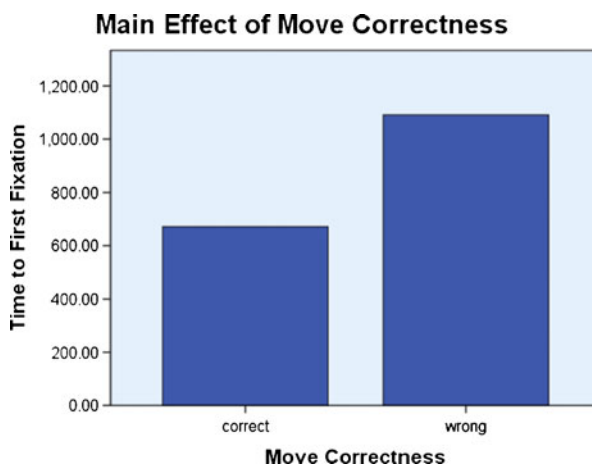


Fig. 12 Main effect on time to first fixation

Table 1 Logistic regression results for *Move Correctness After Hint*

	B (SE)	p	95 % CI for odds ratio		
			Lower	Odds ratio	Upper
Fixations/word	0.98 (0.44)	0.03	1.12	2.68	6.39

leverage both correct and incorrect moves for fostering students' reflection when needed, but they also show that student attitude toward receiving help should be taken into account to increase the effectiveness of these hints. We found that students with a positive attitude towards receiving help tended to pay more attention to hints after correct moves, students with a negative attitude towards help showed more attention to hints after incorrect moves, and students with a neutral attitude showed limited attention to hints overall. Thus, strategies should be investigated to increase attention to hints that are tailored to the specific attention shortcomings generated by each type of attitude towards receiving help.

In the next section, we show initial evidence that improving attention to hints as discussed here is a worthwhile endeavour because it can improve student interaction with the game.

Effect of Attention to Hints on Game Playing

In this section, we look at whether attention to hints impacts students' actions and performance with Prime Climb.

We start by focusing on the effect of attention to hints on correctness of the player's subsequent move. As our dependent variable, *Move Correctness After Hint*, is categorical (e.g., the move is either correct or incorrect), we use logistic regression to determine if *Total Fixation Time*, *Fixations/word* and *Hint Type* are significant predictors of *Move Correctness After Hints*⁶.

Table 1 shows the results of running logistic regression on the data. A Hosmer-Lemeshow test of goodness of fit was not significant, $\chi^2(8)=7.105$, $p>.05$, indicating the model fit the data well. As can be seen in Table 1, *Fixations/word* is the only significant predictor of *Move Correctness After Hints*. The odds ratio greater than 1 indicates that, as fixations/word increases, the odds of a correct move also increase. This suggests that when the players read the hints more carefully, their next move is more likely to be correct. The results of the logistic regression also indicate that the type of hint students pay attention to does not impact move correctness. This finding is consistent with the fact that, in Prime Climb, *bottom-out* hints do not provide direct information on what to do next; they only explain how to evaluate the player's *previous* move in terms of number factorization and this information cannot be

⁶ The data points in our dataset are not independent, since they consist of sets of moves generated by the same students. Lack of independence can increase the risk of making a type 1 error due to overdispersion (i.e., ratio of the chi-square statistic to its degrees of freedom is greater than 1), but this is not an issue in our data set. ($\chi^2=6.41$, $df=8$).

directly transferred to the *next* move. Still, it appears that students benefit from paying attention to the hints, since when they attend to the hints they tend to make fewer errors on subsequent moves. This finding suggests that further investigation on how to increase student attention to hints is a worthwhile endeavor, because it can improve student performance with the game and possibly help trigger student learning.

Next, we look at whether attention to the *Tool* hints affected the subject's decision to follow the hint's advice and use the magnifying glass. We again use logistic regression to determine whether *Total Fixation Time* and *Fixations/word* are significant predictors of the categorical dependent variable *Use Magnifying Glass*. In this case, neither measure was a significant predictor. A possible explanation for this result is that the *Tool* hints are essentially all the same (aside for minor variations in wording), thus reading one *Tool* hint may be enough to get the message and subsequent hints simply act as reminders, that is they can be effective in triggering magnifying glass usage even if the subject just glances quickly at the hint.

Comparing Students' Behaviors and Their Subjective Assessments

In this section, we look at some of the answers provided by participants in the study post-questionnaire, to see if they reflect the behaviors derived from the study action log and gaze data. We found that students who agreed with the statement "I used the magnifying glass" tended to use the magnifying glass more frequently, $r(12)=.693$, $p=.013$. There was also a strong, significant correlation between fixations/word and participants' self-reported agreement with the statement "The agent's hints were useful for me", $r(12)=.655$, $p=.021$. The correlation between fixations/word and participants' self-reported agreement with the statement "I read the hints given by the agent" was still rather strong, but only marginally significant, $r(12)=.514$, $p=.087$, possibly due to the small size of the data set. All in all, these results provide encouraging evidence that students were answering the post questionnaire rather truthfully, despite general concerns that exist regarding the reliability of these instruments to elicit information from study participants.

Conclusions and Future Work

Many Intelligent Tutoring systems provide adaptive support to their students in the form of adaptive textual hints that gradually help students through specific educational activities when they have difficulties proceeding on their own. Despite the widespread adoption of adaptive hints, there is substantial evidence of possible limitations. Some students use the hints to get quick answers from the ITS (i.e., they game the system), others don't look for or want hints altogether. This paper presented a user study to investigate student attention to user-adaptive hints during interaction with Prime Climb, an educational computer game for number factorization. In particular, we aimed to find initial answers to the following questions:

- 1) Do students attend to adaptive hints that they have not explicitly requested?

- 2) If they do, which factors related to game play (e.g., move correctness, interaction time) or student differences (e.g. student's initial domain knowledge, attitude toward receiving help) affect a student's tendency to attend to the unsolicited adaptive hints?
- 3) Is attention to these hints useful, i.e. does it impact game performance?

This work contributes to existing research on student use and misuse of adaptive hints in ITS by looking at how students react to hints when they are provided unsolicited by the system, as opposed to being explicitly requested by the student or obtained via gaming strategies. A second contribution is that, to the best of our knowledge, this work is the first to focus on adaptive hints provided by an edu-game, i.e., in a context in which it is especially challenging to provide didactic support, because this support can interfere with game playing. An additional contribution is that we use eye-tracking data to analyze student attention to hints, whereas most previous research on this topic relied on time-based measures. An exception is the work by Gluck et al. (2000), who also used eye-tracking to provide initial evidence that students do not pay much attention to an ITS's adaptive feedback. Our work, however, provides a more in-depth gaze-based analysis of which factors affect attention.

We found that attention to hints is affected by a variety of factors related to student performance in the game, hint timing and context, as well as attitude toward receiving help in general. We also found that attention to hints affects game performance in Prime Climb (i.e. correctness of the players' moves), thus indicating that improving attention can be beneficial for students. The next step in this research will be to leverage the findings in this paper to improve the design and delivery of the Prime Climb hints. First, we plan to extend the Prime Climb student model to use eye-tracking data in real-time for assessing if a student is attending to hints. To attract student attention when it is lacking, we plan to leverage our findings on the factors that affect attention to hints. For instance, the finding that attention to *definition hints* (i.e. hints providing re-teaching of relevant factorization concepts) decreases when they are provided during the second half of the game can be taken into account to make these hints more relevant and interesting if they have to be provided at that time (e.g. by varying the presentation of the hints, or by better justifying why they are still presented). In light of our results, we will also extend the Prime Climb student model to incorporate information about a student's attitude towards receiving help and investigate strategies to increase attention to hints based on this attitude.

In this paper, we looked at three basic eye-tracking measures to gauge attention to hints in Prime Climb: *Total Fixation Time*, *Fixations/word* and *Time to First Fixation*. Another avenue of future research is to consider more sophisticated attention patterns, e.g., transitions between hints and other relevant areas on the Prime Climb interface, such as the mountain and the Magnifying glass. One option to identify which patterns should be considered as indicative of productive attention is to mine them from data. This process involves first using a form of clustering to identify groups of users that behave similarly in terms of how they process hints visually, and then identifying possible relationships of these patterns with learning, similarly to what Kardan and Conati (2012) have done with interface actions. Finally, action and eye-data can be

combined to build real-time classifiers that leverage both sources for predicting student learning as the interaction proceeds and intervening if learning is predicted to be low. Kardan and Conati (2013) discuss one such classifier for an interactive simulation for university students. We aim to investigate if and how this approach generalizes to more game-like environments, as exemplified by Prime Climb.

References

- Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In *Intelligent Tutoring Systems* (pp. 227–239). Macció, Brazil. Retrieved from <http://www.springerlink.com/index/RTCY8JNL39PWQKFJ.pdf>.
- Anderson, J. R. (2002). Spanning seven orders of magnitude: a challenge for cognitive modeling. *Cognitive Science*, 26(1), 85–112.
- Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
- Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In *Intelligent Tutoring Systems* (pp. 531–540). Macció, Brazil. Retrieved from <http://www.springerlink.com/index/LWBCJ45L817QXFWM.pdf>.
- Baker, R., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*, 18(3), 287–314.
- Barrow, D., Mitrovic, A., Ohlsson, S., & Grimley, M. (2008). Assessing the impact of positive feedback in constraint-based tutors. In *Intelligent Tutoring Systems* (pp. 250–259). Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-69132-7_29.
- Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. In *Proceedings of the 21st National Conference on Artificial Intelligence* (Vol. 21, pp. 2–8). Boston. Retrieved from <http://www.aaai.org/Papers/AAAI/2006/AAAI06-024.pdf>.
- Beck, J. E. (2005). Engagement tracing: using response times to model student disengagement. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp. 88–95). Amsterdam.
- Bondareva, D., Conati, C., Feyzhi-Behnagh, R., Harley, J., Azevedo, R., & Bouchet, F. (2013). Inferring Learning from Gaze Data during Interaction with an Environment to Support Self-Regulated Learning.
- Bull, S., Cooke, N., & Mabbott, A. (2007). Visual attention in open learner model presentations: An eye-tracking investigation. In *User Modeling 2007* (pp. 177–186). Presented at the Proceedings of UMAP 2007, the 11th International Conference on User Modeling, Springer-Verlag. Retrieved from <http://www.springerlink.com/index/p2272121j9673136.pdf>
- Conati, C., & Klawe, M. (2002). Socially intelligent agents in educational games. In K. Dautenhahn, A. Bond, D. Canamero, & B. Edmonds (Eds.), *Socially intelligent agents - creating relationships with computers and robots* (213–220). Kluwer Academic Publishers.
- Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, 19(3), 267–303.
- Conati, C., & Manske, M. (2009). Evaluating adaptive feedback in an educational computer game. In *Intelligent Virtual Agents* (pp. 146–158). Retrieved from <http://www.springerlink.com/index/y1p6146183727716.pdf>.
- Conati, C., & Merten, C. (2007). Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. *Knowledge-Based Systems*, 20(6), 557–574.
- Conati, C., Gertner, A., & Vanlehn, K. (2002b). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12(4), 371–417.
- D’Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: a gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies*, 70(5), 377–398.
- De Castell, S., & Jenson, J. (2007). Digital games for education: When meanings play. *Intermedialités: Histoire et théorie des arts, des lettres et des techniques*, (9). Retrieved from <http://www.erudit.org/revue/im/2007/v/n9/1005533ar.html?vue=resume>.

- Easterday, M., Aleven, V., Scheines, R., & Carver, S. (2011). Using tutors to improve educational games. In *Artificial Intelligence in Education* (pp. 63–71). Retrieved from <http://www.springerlink.com/index/P1321423258N254P.pdf>.
- Gluck, K., Anderson, J., & Douglass, S. (2000). Broader Bandwidth in Student Modeling: What if ITS Were “Eye” TS? In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems* (pp. 504–513). Springer-Verlag. Retrieved from <http://www.springerlink.com/index/B5WQD63CA47N29VJ.pdf>
- Goldberg, J. H., & Helfman, J. I. (2010). Comparing information graphics: a critical look at eye tracking. In *Proceedings of BELIV'10, the 3rd Workshop: BEyond time and errors: novel evaluation methods for Information Visualization* (pp. 71–78). Retrieved from <http://dl.acm.org/citation.cfm?id=2110203>
- Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics*, 24(6), 631–645.
- Goldin, I. M., Koedinger, K. R., & Aleven, V. (2012). Learner Differences in Hint Processing. *EDM*. Retrieved from <http://www.pitt.edu/~goldin/pubs/goldin-2012-different-learners.pdf>
- Johns, J., & Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the 21st National Conference on Artificial Intelligence* (Vol. 21, p. 163). Boston. Retrieved from <http://www.aaai.org/Papers/AAAI/2006/AAAI06-026.pdf>
- Just, M. C. (1986). *The Psychology of Reading and Language Comprehension*. Boston. (1986). Boston.
- Kardan, S., & Conati, C. (2012). Exploring gaze data for determining user learning with an interactive simulation. In *Proceedings of UMAP 2012, the 20th International Conference on User Modeling, Adaptation, and Personalization* (pp. 126–138). Springer. Retrieved from <http://www.springerlink.com/index/4810592K35003701.pdf>
- Kardan, S., & Conati, C. (2013). Comparing and Combining Eye Gaze and Interface Actions for Determining User Learning with an Interactive Simulation. In *User Modeling, Adaptation, and Personalization* (pp. 215–227). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-38844-6_18.
- Linehan, C., Kirman, B., Lawson, S., & Chan, G. (2011). Practical, appropriate, empirically-validated guidelines for designing educational games. In *Proceedings of CHI 2011, the 2011 Annual Conference on Human factors in Computing Systems* (pp. 1979–1988). Retrieved from <http://dl.acm.org/citation.cfm?id=1979229>.
- Loboda, T. D., & Brusilovsky, P. (2010). User-adaptive explanatory program visualization: evaluation and insights from eye movements. *User Modeling and User-Adapted Interaction*, 20(3), 191–226.
- Manske, M., & Conati, C. (2005). Modelling learning in educational games. In *Proceedings of AIED 2005, the 12th International Conference on Artificial Intelligence in Education*. Amsterdam: University of British Columbia. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.6247&rep=rep1&type=pdf>.
- Mathews, M., Mitrovic, A., Lin, B., Holland, J., & Churcher, N. (2012). Do Your Eyes Give It Away? Using Eye Tracking Data to Understand Students’ Attitudes towards Open Student Model Representations. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 424–429). Springer-Verlag. Retrieved from <http://www.springerlink.com/index/6T47G3L65482N153.pdf>.
- Mitrovic, A. (2012). Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User Modeling and User-Adapted Interaction*, 22(1–2), 39–72.
- Mitrovic, A., & Martin, B. (2000). Evaluating the effectiveness of feedback in SQL-Tutor. In *Advanced Learning Technologies, 2000. IWALT 2000. Proceedings. International Workshop on* (pp. 143–144). Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=890591.
- Muir, M., & Conati, C. (2012). An Analysis of Attention to Student-Adaptive Hints in an Educational Game. In *Intelligent Tutoring Systems* (pp. 112–122). Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-30950-2_15.
- Muldner, K., Atkinson, R., & Burleson, W. (2009). Investigating the utility of eye-tracking information on affect and reasoning for user modeling. *User Modeling, Adaptation, and Personalization*, 138–149.
- Peirce, N., Conlan, O., & Wade, V. (2008). Adaptive educational games: Providing non-invasive personalised learning experiences. In *Proceedings of DIGTEL 2008, the second IEEE International Conference on Digital Games and Intelligent Toys Based Education* (pp. 28–35). Banff, Canada. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4700726.
- Qu, L., & Johnson, W. L. (2005). Detecting the learner’s motivational states in an interactive learning environment. In *Proceedings of AIED 2005, the 12th International Conference on Artificial Intelligence in Education* (pp. 547–554). IOS Press.
- Robison, J. L., McQuiggan, S. W., & Lester, J. C. (2009). Modeling task-based vs. affect-based feedback behavior in pedagogical agents: An inductive approach. In *Proceedings of the 2009 conference on*

- Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 25–32).
- Roll, I., Aleven, V., McLaren, B., Ryu, E., Baker, R., & Koedinger, K. (2006). The Help Tutor: Does Metacognitive Feedback Improve Students' Help-Seeking Actions, Skills and Learning? In *Intelligent Tutoring Systems* (pp. 360–369). Retrieved from <http://www.springerlink.com/index/a1861j14616g3n62.pdf>.
- Rowe, J., & Lester, J. (2010). Modeling user knowledge with dynamic Bayesian networks in interactive narrative environments. In *Proceedings of the Sixth Annual AI and Interactive Digital Entertainment Conference* (pp. 57–62). Retrieved from <http://www.aaai.org/ocs/index.php/AIIDE/AIIDE10/paper/download/2149/2545>.
- Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 21(1), 115–133.
- Sabourin, J., Shores, L., Mott, B., & Lester, J. (2012). Predicting student self-regulation strategies in game-based learning environments. In *Intelligent Tutoring Systems* (pp. 141–150). Retrieved from <http://www.springerlink.com/index/67VU6G07664083HV.pdf>.
- Shih, B., Koedinger, K., & Scheines, R. (2010). Unsupervised discovery of student learning tactics. In *Proceedings of the Third International Conference on Educational Data Mining*. Retrieved from http://pact.cs.cmu.edu/pubs/Shih,%20Koedinger,%20Scheines-edm2010_.pdf.
- Sibert, J. L., Gokturk, M., & Lavine, R. A. (2000). The reading assistant: eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (pp. 101–107). ACM Press. Retrieved from <http://dl.acm.org/citation.cfm?id=354418>.
- Snijders, T., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, 22(3), 342–363.
- Toker, D., Conati, C., Steichen, B., & Carenini, G. (2013). Individual User Characteristics and Information Visualization: Connecting the Dots through Eye Tracking. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems, (CHI 2013)*. Retrieved from <http://www-devel.cs.ubc.ca/~conati/my-papers/CHI2013CameraReady.pdf>.
- Van Eck, R. (2007). Building artificially intelligent learning games. *Games and simulations in online learning: Research and development frameworks*, pp. 271–307.
- Wainwright, P. E., Leatherdale, S. T., & Dubin, J. A. (2007). Advantages of mixed effects models over traditional ANOVA models in developmental studies: a worked example in a mouse model of fetal alcohol syndrome. *Developmental Psychobiology*, 49(7), 664–674.
- Walonoski, J., & Heffernan, N. (2006). Prevention of off-task gaming behavior in intelligent tutoring systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 722–724). Jhongli, Taiwan. Retrieved from <http://www.springerlink.com/index/e2j662w5x15x3707.pdf>.
- Wang, H., Chignell, M., & Ishizuka, M. (2006). Empathic tutoring software agents using real-time eye tracking. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications* (pp. 73–78). ACM Press. Retrieved from <http://dl.acm.org/citation.cfm?id=1117346>.
- Woolf, B. P. (2008). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.