Available online at www.sciencedirect.com

SciVerse ScienceDirect

Int. J. Human-Computer Studies 70 (2012) 377–398

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Gaze tutor: A gaze-reactive intelligent tutoring system

Sidney D'Mello[a,b,*], Andrew Olney[c], Claire Williams[c], Patrick Hays[c]

[a]Department of Computer Science, 384 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556, USA
[b]Department of Psychology, 384 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556, USA
[c]Institute for Intelligent Systems and Psychology Department, 202 Psychology Building, University of Memphis, Memphis, TN 38152, USA

Received 13 October 2010; received in revised form 25 January 2012; accepted 26 January 2012
Communicated by C. Penstein
Available online 7 February 2012

Abstract

We developed an intelligent tutoring system (ITS) that aims to promote engagement and learning by dynamically detecting and responding to students' boredom and disengagement. The tutor uses a commercial eye tracker to monitor a student's gaze patterns and identify when the student is bored, disengaged, or is zoning out. The tutor then attempts to reengage the student with dialog moves that direct the student to reorient his or her attentional patterns towards the animated pedagogical agent embodying the tutor. We evaluated the efficacy of the gaze-reactive tutor in promoting learning, motivation, and engagement in a controlled experiment where 48 students were tutored on four biology topics with both gaze-reactive and non-gaze-reactive (control condition) versions of the tutor. The results indicated that: (a) gaze-sensitive dialogs were successful in dynamically reorienting students' attentional patterns to the important areas of the interface, (b) gaze-reactivity was effective in promoting learning gains for questions that required deep reasoning, (c) gaze-reactivity had minimal impact on students' state motivation and on self-reported engagement, and (d) individual differences in scholastic aptitude moderated the impact of gaze-reactivity on overall learning gains. We discuss the implications of our findings, limitations, future work, and consider the possibility of using gaze-reactive ITSs in classrooms.

Keywords: Affective computing; Affect-sensitive ITS; Boredom; Disengagement; Eye tracking; Gaze-sensitive dialogs; Intelligent tutoring systems (ITSs); Zoning out

1. Introduction

Intelligent tutoring systems (ITSs) have emerged as effective tools to promote active knowledge construction by capitalizing on the merits of one-on-one human tutoring in an automated fashion (Graesser et al., in press; Psotka et al., 1988; Sleeman and Brown, 1982; Woolf, 2009). ITSs are increasingly being used in classrooms all over the United States, and the ones that have been successfully implemented and tested have produced learning gains with average effect sizes ranging from .79 to 1 sigma[1] (Corbett, 2001; Koedinger et al., 1997; VanLehn, 2011; VanLehn et al., 2007). When compared to classroom instruction and other naturalistic controls, the 1.0 effect sizes obtained by ITSs are superior to the .39 sigma effect for computer-based training, the .50 sigma effect for multimedia, and the .40 sigma effect obtained by novice human tutors (Cohen et al., 1982; Corbett, 2001; Dodds and Fletcher, 2004; Wisher and Fletcher, 2004).

*Corresponding author at: Department of Computer Science, 384 Fitzpatrick Hall, University of Notre Dame, Notre Dame, IN 46556, USA. Tel.: +1 901 378 0531.
E-mail addresses: sdmello@nd.edu (S. D'Mello), aolney@memphis.edu (A. Olney), mcwllams@memphis.edu (C. Williams), dphays@memphis.edu (P. Hays).

[1]An effect-size measures the strength of a relationship between two variables. Cohen's d (see below) is a common measure of effect size in standard deviation units between two samples with means M_1 and M_2 and standard deviations s_1 and s_2. According to Cohen (1992), effect sizes approximately equal to .3, .5, and .8 represent small, medium, and large effects, respectively. $d=(M_1-M_2)/\sqrt{(s_1^2+s_2^2)/2}$. In learning contexts, an effect size of 1.0 sigma is roughly equivalent to an improvement of one letter grade.

?show 1071-5819/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2012.01.004

Despite their impressive successes, it is important to note that ITSs are not the panacea for all the problems associated with learning. Although most ITSs are effective at supporting students' cognitive needs, until recently, they have made less of an effort to promote students' engagement, motivation, and interest in learning. This is a serious limitation that reduces the efficacy of these systems because engagement, motivation, and interest are precursors to learning, effortful problem solving, and deep thinking (Berlyne, 1978; Craig et al., 2004; Csikszentmihalyi, 1990). Students might begin a learning session with an ITS with some level of interest and enthusiasm, but boredom inevitably creeps in as the session progresses, when the novelty of the system and content fades, and when they have difficulty comprehending the material (Csikszentmihalyi, 1990; D'Mello and Graesser, in press; Larson and Richards, 1991; Mann and Robinson, 2009; Moss et al., 2008; Pekrun, 2010; Pekrun et al., 2010). When boredom strikes, students' interest wanes to a point where they give up and eventually disengage from the learning session. At this point, any further instruction is essentially futile.

Recent work, albeit outside of learning contexts, has investigated how engagement can be maintained in computer based interventions over extended periods of time (e.g., months) (Bickmore et al., 2010; Bickmore and Picard, 2005). For example, in a series of studies measuring long term engagement with an animated agent for a health intervention, Bickmore et al. (2010) demonstrated that increasing the variability of the agent's behavior and adding a backstory to the agent increased the amount of time users spent with the system. Some research has also examined whether polite or direct strategies are more effective at maintaining engagement. When users are switching tasks, polite interruptions, as measured by the annoyingness of computer beep/alert sounds, has been shown to increase user compliance in a health care intervention, whereas impolite interruptions had the opposite effect (Bickmore et al., 2007). However, in negotiation settings, angry/threatening statements such as "This is a ridiculous offer, it really pisses me off" were found to lead to greater user concessions than neutral or happy/non-threatening statements (de Melo et al., 2010). The different effects suggest that task-specific factors, such as multi-tasking vs. single-tasking, may strongly influence whether polite/non-threatening agent behaviors enhance successful outcomes more than annoying/threatening behaviors.

One perspective on this recent work is to consider the problem of engagement as two complimentary subproblems operating on different timescales. The first is disengagement repair. Disengagement occurs within a session and prevents the user from completing that session successfully. Disengagement repair requires refocusing the user's attention and increasing his or her motivation to complete the task at hand. The second engagement subproblem is maintaining sustained engagement across multiple sessions. Sustained engagement requires making the sessions compelling enough so that the risk of disengagement is minimized within a session and attrition across sessions is low. These two subproblems are related, since a user who becomes disengaged during a single session may be less likely to engage in a future session. However, work on sustained engagement in computer-based interventions has not directly addressed disengagement repair.

The present paper focuses on disengagement repair strategies within the context of learning environments. Before articulating the specific disengagement-repair strategy we have implemented, we review some findings on the prevalence, antecedents, and consequences of boredom and disengagement during learning.

1.1. Boredom and disengagement during learning

When compared to cognitive constructs such as attention and memory, or basic emotions such as anger and disgust (Ekman, 1992), the scientific research on boredom during complex learning is relatively sparse and scattered. For example, the number of studies on boredom and engagement in educational contexts is negligible when compared to the approximately 1000 studies on test anxiety (Hembree, 1988; Pekrun et al., 2010; Zeidner, 2007). Nevertheless, some theoretical models of the cognitive and affective processes that underlie boredom have emerged (Larson and Richards, 1991; Mann and Robinson, 2009). The *understimulation* model (Perkins and Hill, 1985) posits that boredom arises when the student is physiologically and cognitively under-aroused, presumably due to the monotony of repetitive tasks that have been habituated (e.g., solving numerous algebraic problems once the basic concepts have been mastered). The *forced-effort* model (Larson and Richards, 1991; Robinson, 1975) claims that students will experience more boredom when they are required to invest considerable mental effort in tasks that are beyond their control (e.g., forced to suffer through a lecture when there is no intrinsic motivation to learn).

According to Pekrun's *control-value* theory of emotions, subjective appraisals of control and value of a learning activity are the critical predictors of engagement (Hulleman et al., 2008; Pekrun, 2010; Pekrun et al., 2006). Subjective control pertains to the perceived influence that a student has over the activity, while subjective value represents the perceived value of the outcomes of the activity. Boredom occurs when perceived value or control are low, as would be the case when an unmotivated student (low value) is attempting to solve math problems that far exceed his or her ability (low control) (Csikszentmihalyi, 1975). Boredom has also been hypothesized to occur when control is too high, as is the case when skills greatly outweigh challenges and the student is understimulated (Pekrun et al., 2010).

In addition to these theoretical perspectives, boredom has recently been gaining some attention in studies that investigate the links between affect and cognition during learning (Baker et al., 2010; Beck, 2005; Cocea and Weibelzahl, 2009; D'Mello and Graesser, in press; Drummond and Litman, 2010; Moss et al., 2008; Pekrun et al., 2010). Available data

suggest a number of conclusions pertaining to the incidence and effects of boredom during learning. These conclusions are summarized below.

1.1.1. Prevalence of boredom

Boredom is one of the most frequent affective states that students experience during learning, irrespective of the learning context, content area, task, student population, and method used to track affect (see D'Mello, in preparation for a meta-analysis). Boredom is not only prevalent with computer learning environments, but is also observed in human–human tutoring sessions. For example, an analysis of several tutoring sessions with expert human tutors indicated that students spent a fourth of the time merely socially attending to the tutor instead of actively learning the material (Lehman et al., 2008).

1.1.2. Hindrance to learning and performance

As could be expected, boredom negatively correlates with learning gains (Craig et al., 2004; D'Mello and Graesser, 2011; Forbes-Riley and Litman, 2011; Schutz and Pekrun, 2007), presumably because bored students have trouble focusing attention (Fisher, 1993; Thackray, 1981) or are simply not willing to process the material at deeper levels of comprehension.

1.1.3. Persistent temporal quality

Boredom adopts a persistent temporal quality upon activation (D'Mello and Graesser, 2011), where students wallow in their ennui and are less likely to be reengage in the material. This form of persistent boredom is a negative predictor of learning gains. More importantly, the typical tutorial interventions (e.g., feedback, hints) are not very effective in alleviating boredom, indicating that novel pedagogical and motivational strategies are required to increase task persistence.

1.1.4. Gateway into negative affect

Consistent with predictions of the forced-effort model (Larson and Richards, 1991), bored students are more likely to transition into frustration (D'Mello and Graesser, 2010a). Frustration is another affective state that is harmful to learning (Linnenbrink and Pintrich, 2002). Persistent frustration can also transition into boredom if the student is stuck and simply gives up.

1.1.5. Catalyst for harmful behaviors

Bored students also engage in problematic behaviors such as going off-task, zoning out, intentionally misusing the learning environment (i.e., gaming the system), or simply becoming careless. These behaviors, and boredom in general, lead to lower self-efficacy, diminished interest in educational activities, increased attrition and dropout, and eventually lead to poorer learning (Baker et al., 2010; Cocea et al., 2009; Craig et al., 2004; Drummond and Litman, 2010; Moss et al., 2008).

1.1.6. Long-term effects of boredom

In addition to the short-term effects of negligible or even negative learning gains, boredom in educational activities is diagnostic of lower self-efficacy, lack of motivation in learning, hostility and dissatisfaction towards school, abnormal behavior in school, lower work satisfaction, and diminished work output (Fogelman, 1976; McGiboney and Carter, 1988; Perkins and Hill, 1985; Robinson, 1975; Wasson, 1981).

Given this sketch of the harmful effects of boredom on learning, it is important for ITSs to be more than mere cognitive machines, because preventing waning attention, zoning out, disengagement, and boredom are critically important for learning (Calvo and D'Mello, 2010; del Soldato and du Boulay, 1995; Woolf, 2009). Fortunately, as highlighted in the next section, there has been a recent emergence of research along this front.

1.2. Disengagement diagnosis and repair

A number of research groups have been addressing the problem of building learning environments that detect and respond to affective states such as boredom, confusion, frustration, and anxiety (Afzal and Robinson, 2009; Burleson and Picard, 2007; Chaffar et al., 2009; Conati and Maclaren, 2009; D'Mello and Graesser, 2010b; D'Mello et al., 2010b; Forbes-Riley et al., 2008; Robison et al., 2009; Woolf et al., 2010). These systems use state-of-the art sensing technologies and machine learning techniques to automatically detect student affect by monitoring facial features, speech contours, body language, interaction logs, language, and peripheral physiology (e.g., electromyography, galvanic skin response) (see Calvo and D'Mello, 2010 for an overview). These affect-sensitive systems then alter their pedagogical and motivational strategies in a manner that is dynamically responsive to the sensed affective states. Some of the implemented responses to student affect include affect mirroring (Burleson and Picard, 2007), empathetic responses (Woolf et al., 2010), and a combination of politeness, empathy, encouragement, and incremental challenge (D'Mello et al., 2010b).

Although these affective-response strategies have the potential of alleviating certain negative emotions (e.g., frustration), an effective response to boredom must address attention due to the inextricable link between attention and engagement (Fisher, 1993; Pekrun et al., 2010; Thackray, 1981). That is, engagement can be conceptualized as a state of involvement with a task such that concentration is intense, attention is focused, and involvement is moderate to complete (Baker et al., 2010; Csikszentmihalyi, 1975, 1990). Engagement is a multi-faceted construct encompassing both cognitive and affective components. Some of the cognitive aspects of engagement include attention and concentration, while the affective components consist of modulations in arousal and valence (D'Mello et al., 2007; Mandler, 1984; Pekrun et al., 2010).

Attention, which is one important cognitive component of engagement, is the focus of the present paper. Attention is critical because maintaining engagement in a learning activity requires attentional resources. Therefore, developing interventions that monitor periods of waning attention and attempt to encourage more productive use of attentional resources might be one promising way to increase engagement and promote learning. This paper tests this claim by using eye tracking to track periods of disengagement and uses gaze-sensitive dialogs in an attempt to reorient students' attention towards an animated pedagogical agent that embodies the tutor (as will be described below).

While eye tracking has historically been used in reading research (Rayner, 1998), it has been finding increased use in the context of learning and problem solving (van Gog and Scheiter, 2010). For example, eye tracking has been used to obtain novel insights into: (a) the split-attention effect (Schmidt-Weigand et al.), (b) saliency-based cueing (de Koning et al.), (c) expert–novice differences in attentional deployment during problem solving (Graesser et al., 2005), (d) text–diagram integration during comprehension (Hegarty and Just, 1993; Holsanova et al., 2009), and (e) meta-cognitive processes during learning (Conati and Merten, 2007).

These examples clearly illustrate the importance of eye tracking towards obtaining a low-level mechanistic account of the cognitive processes during learning. However, eye tracking can be more than a mere research tool. It can also be used as a means of improving learning gains. For example, in Attention Aware Systems (Roda and Thomas, 2006), eye tracking can be used to detect and alter users' attention to improve outcomes (Hyrskykari, 2006). Possible attention-sensitive responses include the modulation of information pacing (e.g., halting or slowing down the presentation of new information when the user is overwhelmed) and selection of modality style (e.g., present information auditorily when visual channel is busy) (Roda and Thomas, 2006; Toet, 2006). Other examples include real-time eye tracking: (a) to guide the behaviors of an animated pedagogical agent (Wang et al., 2006), (b) for attentional guidance during problem solving (van Gog et al., 2009), and (c) for student modeling (Conati and Merten, 2007). In a somewhat different vein, of present interest is the use of eye tracking to detect and alleviate disengagement during learning.

1.3. Overview of present research

Maintaining engagement in a learning activity requires attentional resources, hence, we explored the possibility of using students' eye gaze patterns to track attentional deployment, identify attentional failures (i.e., zoning out), and reorient attention to facilitate learning. As articulated above, engagement is a complex construct that encompasses both the mind and body. There might be aspects of engagement that might be manifested in physiology, facial expressions, and posture (D'Mello and Graesser, 2010b; Mota and Picard, 2003), and the present focus on gaze patterns risks overlooking these alternate manifestations of engagement. However, eye tracking has a long history as a tool to monitor patterns of attentional deployment or even a lack of attention (Asteriadis et al., 2009a, 2009b; Rayner and Fischer, 1996; Reichle et al., 2010; van Gog et al., 2009). For example, a lack of fixations on the text and an increased number of blinks have been associated with "mind wandering" (Smilek et al., 2010). Therefore, we have some confidence that using eye tracking to monitor attentional patterns will provide some insights into students' levels of engagement.

We developed a computer tutor that tracked engagement by monitoring students' eye gaze patterns while the tutor and the student engaged in a collaborative lecture (described in Section 2). The tutor assumed that the student was disengaged when he or she looked away from the screen for an extended period of time. It then attempted to reengage the student by providing messages that directly instructed the student to focus on the animated pedagogical agent that embodied the tutor. Our prediction was that monitoring and responding to disengagement with a gaze-sensitive (or gaze-reactive) tutor would yield superior learning gains compared to a non-gaze-sensitive (non-gaze-reactive) tutor.

We tested this prediction in an experiment where students were tutored on biology topics and the tutor either responded (experimental condition) or ignored (control condition) students' disengagement (tracked via their gaze patterns as described in Section 3). In addition to measuring learning gains, which is the primary dependent variable of interest, we also assessed whether responding to gaze patterns impacted students' motivation to learn and their self-reported engagement levels. Finally, we tested whether individual differences in prior knowledge, aptitude, and perceptions of learning biology from computer tutors moderated the effects of gaze-reactivity on learning, motivation, and engagement.

To our knowledge, the efficacy of gaze-reactive dialogs as an intervention to promoting learning, engagement, and motivation has not yet been systematically investigated. Therefore, the novelty of this work emerges from the development and evaluation of the first gaze-reactive ITS to diagnose and alleviate boredom.

We begin with a description of the gaze-reactive ITS, which we developed (Section 2), followed by the experimental protocol (Section 3), and the results of the experiment (Section 4). We conclude by taking stock of our findings, discussing limitations, and addressing the practical implications of this research for computer tutors deployed in real-world settings (i.e., classrooms and computer labs in schools).

2. Gaze-reactive dialog-based biology tutor

The ITS we implemented (Guru) is designed to tutor students on high school biology topics (e.g., cellular

respiration, mitosis, ecological succession) via natural language dialogs. Guru was designed to mirror the tactics, actions, and dialog of expert human tutors. The pedagogical and motivational strategies of Guru are informed by a detailed computational model of expert human tutoring. The model is developed from an analysis of 50 naturalistic tutoring sessions between students and expert human tutors (Cade et al., 2008). The computational model transcends various levels of granularity from (a) tutorial modes (i.e., pedagogically distinct phases in a session such as lecturing and scaffolding that last for several minutes and encompass multiple speech acts), to (b) collaborative patterns of dialog moves within individual modes (i.e., repetitive sequences of dialog moves that have particular pedagogical functions), to (c) individual dialog moves or speech acts (e.g. direct instruction, positive feedback, solidarity statement), and to (d) the language and gestures of tutors (D'Mello et al., 2010c). Understanding how elements (e.g., moves, modes) interact within and across levels is the essence of the computational model.

This paper focuses on one component of the Guru system, namely the collaborative lecturing module. This decision was motivated by two important factors. First, to our surprise, lectures were abundant in the expert tutoring sessions. In particular, when we segmented the tutoring sessions into eight dialog modes, lecturing was the second most frequent mode. Lectures comprised 22.1% of the modes and 30.2% of the dialog turns. Lectures were only surpassed by the scaffolding mode, which comprised 27.8% of the modes and 46.4% of the turns (Cade et al., 2008). Although there are a number of factors governing this high incidence of lectures (see D'Mello et al., 2010a for a discussion on this issue), what is important is the fact that an ITS that aspires to model expert human tutors (such as Guru) should implement the lecturing styles of these tutors to some extent.

Second, although these lectures serve an important pedagogical goal, disengagement is expected to be higher when tutors lecture because lecturing is much less interactive than other dialog modes such as scaffolded problem solving. Matters can only get worse when a computer, instead of an expert tutor, is delivering the lecture, so disengagement repair is critically relevant during lecturing.

A detailed description of the computational model and the implementation of Guru are beyond the scope of this paper. Instead, we focus on the components that are particularly relevant to our immediate goal of developing a tutor that is sensitive to students' gaze-patterns. These include: (a) a brief overview of the computational model of expert tutor lectures, (b) the implementation of the lectures in Guru, and (c) the implementation of the gaze-reactive component in Guru.

2.1. Modeling expert tutor's lectures[2]

An extensive analysis of the collaborative lecture strategies observed in our sample of 50 expert tutoring sessions is discussed in D'Mello et al. (2010a), so, we will focus on the major points here. In particular, there are two major clusters of dialog moves as illustrated in Fig. 1. The first cluster (*information-transmission*) is primarily concerned with the tutor delivering information to student (solid lines in Fig. 1). The tutor may assert some information through direct instruction and explanation (die), to which the student provides backchannel feedback via an acknowledgment (ack), and the tutor asserts more information (die). Alternatively, the tutor transmits some information (die), asks a comprehension gauging question (cgq) (e.g., "Do you understand?"). The student replies with an acknowledgment (ack) (e.g., "Yes sir.") or a metacomment (meta) (e.g., "No. I don't quite get it."), and more information is transmitted. These basic patterns associated with information-transmission account for 70.2% of the dialog moves during the lecture mode.

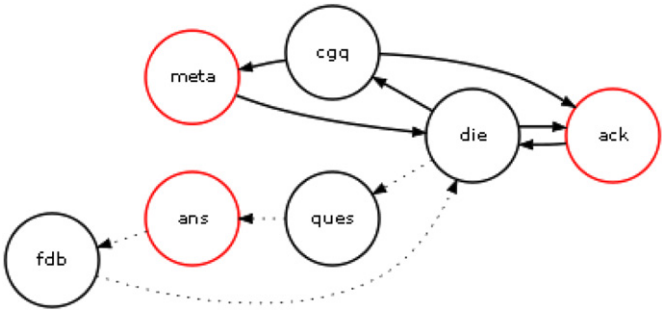

Fig. 1. Information-transmission and information-elicitation clusters. die–direct instruction by tutor, ack—acknowledgment by student, cgq–comprehension gauging question by tutor, meta—metacomment by student, ques—tutor poses question, ans—student answers question, fdb—tutor provides feedback.

The second cluster, or the *information-elicitation* cluster (dotted links in Fig. 1), consists of moves associated with attempts by the tutor to elicit information from the student. These moves are variations of the initiate respond evaluate (IRE) sequence (Mehan, 1979). The sequence begins by the tutor asking the student a question (ques) with prompts, pumps, forced choices, or simplified problems. The student responds with an answer (ans). The tutor evaluates the student's response and provides feedback (fdb) followed by more direct instruction (die). This cluster accounts for 18.6% of the moves during lectures.

In addition to these primary clusters that account for 88.6% of the dialog moves during lecturing, there is also an off-topic conversation cluster (9%), and a student-initiated question cluster (2.2%). These clusters were not implemented in the current version of the computer tutor.

To summarize, our analysis of lectures during expert tutoring sessions was not consistent with boring, extended, and long-winded explanations. Instead, we found that expert tutor lectures were highly *collaborative*, presumably because the expert tutors acknowledge that active participation, even during lectures, is key to learning and engagement (Chi et al., 2008; VanLehn et al., 2007).

It is important to mention one critical point pertaining to the above discussion of the computational model of the

[2]Sections 2.1 and 2.2 are adapted from D'Mello et al. (2010a).

expert tutor lectures. The present discussion was pitched at a very high level of granularity in the interest of brevity and because a detailed description of the model is available in an earlier publication (D'Mello et al., 2010c). Therefore, it is important to emphasize that although the model presented here captures the major patterns in the data, the actual model is an order of a magnitude more complex. For example, the tutor question node (ques) in Fig. 1 is actually an abstract category that represents six concrete tutor question moves: hints, prompts, pumps, forced-choice responses, new problems, and simplified problems. The student answer node (ans) is also an abstract category with concrete members including correct answers, partially correct answers, vague answers, error-ridden answers, and no answers. Expanding these abstract categories into concrete dialog moves yields the more detailed model with 29 nodes and 34 links (D'Mello et al., 2010c).

2.2. Implementing the collaborative lecture in Guru

We developed a lecture module in Guru for eight biology topics (e.g., cellular respiration, amino acids and RNA). As previously stated, Guru's lecturing strategies were designed to closely mirror the expert tutor lectures from our corpus. This was accomplished in two ways. First, the content of the lectures was obtained from transcripts of actual expert tutoring sessions. This made the lecture delivery style more conversational, informal, and presumably more engaging. Using expert tutoring transcripts also captured the sense of time pressure and urgency that comes with naturalistic tutoring. These tutors were tutoring students who had failed exams or who were preparing for one. Our expert tutors were direct, kept a steady pace, and put the students to work.

It should be noted that this form of *content mirroring* was only implemented in the prototype tested in this study. The prototype used a sequential script and did not adapt instruction to individual students, other than providing localized feedback. In contrast, the actual Guru system uses a semi-automated algorithm to extract its content from biology textbooks and other sources (Olney, 2010) and dynamically tailors instruction based on its model of the student's knowledge and abilities.

Second, the tutor closely modeled the collaborative lecturing tactics that were observed from our analysis of the human tutors (see Fig. 1). In particular, Guru primarily transmitted information (68% of the time) but occasionally provided cues for acknowledgments (e.g., "Right?", "ok?"), asked comprehension gauging questions, and prompted the student for answers (e.g., "X is a type of what?"). On average, the lectures contained 32% opportunities for student involvement, thereby yielding a 1:3 student to tutor dialog move ratio. We computed the tutor:student dialog move ratio from eight expert tutor lectures and correlated these with tutor:student dialog move ratio from Guru's implementation of these same lectures. The correlation was very high ($r=.97$), so we are quite confident that Guru does indeed model the collaborative lecturing styles of the expert human tutors.

Sample dialogs from the human tutors and Guru are presented in Table 1. In the actual lecture, the tutor introduces the topic (T1), uses a discourse marker (T2), asserts some information (T3), and then gives the student an opportunity to chime in (T4). The student provides an acknowledgment (S1), the tutor responds with a conversational OK (T5), asserts some more information (T6), and then prompts the student (T7). The student responds (S2), to which the tutor provides some feedback (T8), followed by an assertion, and so on (T9 and S3).

When Guru delivers a lecture, it preserves most of the conversational style from the actual lectures, asserts the same content, and gives the student an opportunity to type a response (see Table 1). The student has three different opportunities to type in a response, which is consistent with the 1:3 student to tutor dialog move ratio discussed above. On average, the expert human tutors articulated 790 words in each lecture, while Guru articulated an average of 718 words.

Table 1
Excerpts from human and computer lectures.

N	Actual lecture	Guru dialog
T1	Let's talk about mitosis.	Let's talk about mitosis.
T2	Ok.	Ok.
T3	Now, let's say here's a skin cell, he's just sitting around, and he needs to divide.	Now, let's say here's a skin cell, it's just sitting around, and it needs to divide.
T4	Someone's got to tell him, right?	Someone's got to tell him to divide, right?
S1	**Mm hmm.**	**< Student response >**
T5	Ok	Ok
T6	I mean, let's say a skin, skin cell is sitting around.	I mean, let's say a skin cell is sitting around.
T7	Do you think somebody needs to tell him to split, or do you think he can just say, oh, I think I'll split?	Do you think somebody needs to tell him to split, or do you think he can just say, oh, I think I'll split?
S2	**Tell him it's time?**	**< Student response >**
T8	Yeah!	
T9	Because, see, now folks need to get instructions, right?	Someone must tell him to split because he needs to get instructions, ok?
S3	**Mm hmm.**	**< Student response >**

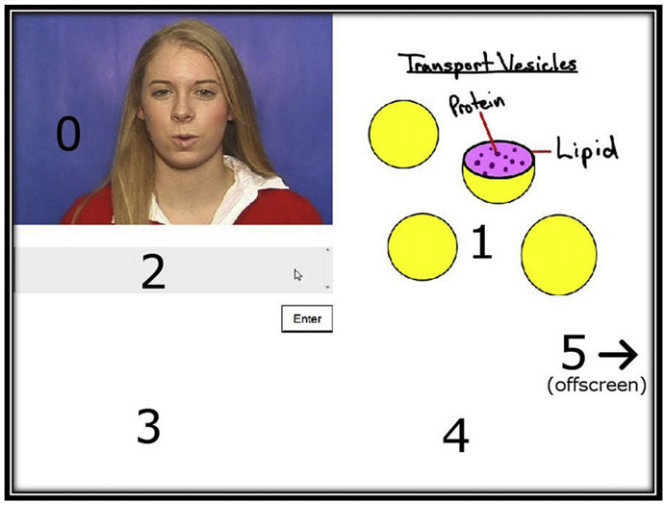

Fig. 2. Screen shot of interface.

The lectures were delivered via a simple conversational interface that consisted of an animated conversational agent that delivered the content of the lectures via synthesized speech, a media panel that displayed images relevant to the lectures, and an input box for students to type their responses (see Fig. 2). The animated agent was created by video recording an actor speaking and then analyzing the video to extract frames of visemes, the visual correlates of phonemes. During agent speech, lip sync was achieved by quickly switching the current image of the agent to viseme frames when viseme events were thrown by the speech engine (using Microsoft's Speech Application Programming Interface). To reduce jitter created by slight posture movements while the actor was producing various visemes, the mouth region of a "silent" agent image was identified, and subsequent visemes were bit-blitted onto that region. This had the effect of localizing viseme changes to the mouth region only. Additional blinking animations were created using the same technique.

A high quality speech engine, NeoSpeech© Kate, was used to produce speech. This speech engine automatically produces reasonable intonation for longer sentences and questions. Complex scientific words that were not in the speech engine's dictionary were manually added and checked for proper pronunciation. The resulting agent was very humanlike in appearance and speech. However the jitter-removal process described above had the consequence that the agent had no side to side body movement and could not perform non-speech-related facial expressions. Additionally no blending was performed between the current viseme and the next, so the agent's speaking appearance was exaggerated compared to normal human speech.

2.3. Gaze-reactive component

The gaze-reactive tutoring system can be considered to be the normal tutoring system augmented with gaze-sensitive user input. Gaze behavior, as measured by the Tobii T60™ eye tracker, has a maximal spatial resolution of 1024×768 pixels for the Guru interface. It is not clear whether this high degree of spatial resolution is meaningful when measuring a user's attention to a simple interface. Hence, to lower the dimensionality of the input, the 1024×728 screen (see Fig. 2) was divided into zones for the *tutor* (zone 0), the *image* (zone 1), the *text box* (zone 2), and the *blank areas* (zones 3 and 4). There was also an *off-screen* zone for gaze patterns that were not classified as falling into any of these five zones.

Interpreting user gaze behavior with respect to the Guru ITS requires a high-level user interface model that specifies Guru's interface states, the gaze behavior categories, and a set of gaze-sensitive responses. The Guru interface can be viewed as a finite state machine with three high-level states: tutor speaking, waiting for student response, and student typing response. Potentially each category of gaze behavior could lead to a different gaze-sensitive response in any of these three states, thereby yielding fifteen possible responses from the system.

Although the number of meaningful tutor responses can be significantly reduced if the goal is to retain the student's attention, it is difficult, if not impossible, to create precise rules for every state. For example, in the state of waiting for a student response, acceptable gaze behaviors might include focusing on the text entry box, the image, or the agent. Likewise, in the state of student response, requiring the student to look at the text entry box may be overly restrictive for touch-typists.

The gaze-reactive system was designed to accommodate variability in attentive behavior. As argued above, two of the three states admit a good deal of variability and appear to be less attractive states for triggering a gaze-reactive intervention. However, the state of tutor speaking is much less flexible, since an attentive student should be looking at either the agent or the image when the tutor is speaking. Moreover, one could argue that the state of tutor speaking is the most important state for maintaining student attention because the tutor introduces the bulk of new information and controls the conversation.

In light of these constraints, the gaze-reactive intervention was triggered when the following conditions were met:

- Tutor was speaking.
- Tutor was not in the middle of a gaze-reactive statement.
- The student had continuously not looked at the agent or image for more than five seconds.
- It had been more than 10 s since the last gaze-reactive statement.

For the present version of the tutor, the time parameters were fixed based on feedback during piloting; however, dynamically setting these parameters based on a particular student may offer additional adaptive advantages.

The gaze-reactive intervention was presented to the student in the following manner. The tutor stopped speaking in mid-sentence, paused for 1 s, and then delivered a gaze-reactive statement to refocus the student's attention, such as, "Please pay attention." The tutor then repeated the interrupted sentence from the beginning. Each gaze-sensitive response was randomly selected from a set of four predefined responses that are designed to reorient students' attention towards the tutor. We piloted with a number of gaze-sensitive responses but narrowed the final set to the following four responses: "Please pay attention," "I'm over here you know," "You might want to focus on me for a change," and "Snap out of it. Let's keep going."

As highlighted in Section 5, the present set of responses was selected to test the effect of *direct* statements that attempt to capture and reorient student's attention. It should be noted that the expert tutors we studied provided direct and immediate feedback to student responses and actions (D'Mello et al., 2010d), because there might be certain advantages to this form of feedback as discussed in the literature (Person et al., 1995). Nevertheless, other more indirect response options and alternate attentional reorientation strategies are discussed in Section 5.

It is important to make one final point regarding the use of eye tracking as a measure of attention. As evident from the description of the interface (Section 2.2), the tutor had an auditory component (i.e., it speaks its dialog moves) and a visual component (i.e., the image and the text box). Gaze-sensitive dialogs were triggered when the student did not fixate on the visual component (i.e., the image and the text box) for an extended period of time (i.e., more than 5 s). A student might have closed his or her eyes in order to tune out the visual component and could have directed all attentional resources towards the auditory component. In these instances, the system would make the incorrect inference that the student was disengaged because it only tracks visual attention. Although this is a limitation of our system, it should be noted that a significant portion of the dialogs occur within the context of an image. Therefore, both auditory and visual attention was required in order to fully engage with the Guru system.

2.4. System architecture

The system architecture of the tutor is presented in Fig. 3. Some details, such as the display of images, have been omitted for clarity. The two loops of information transmission and information elicitation can be traced to *Direct Instruction* and *Tutor Question*, respectively. Tutor questions cause the system to pause, allowing for student keyboard input. Student input is assessed for correctness (*Answer Assessments*), and then the appropriate feedback is delivered (*Feedback*) is generated by consulting the *Script*. Direct instruction statements do not cause the system to pause, so multiple tutor statements may be delivered before a tutor question is asked. Tutor text for both transmission and elicitation loops is maintained

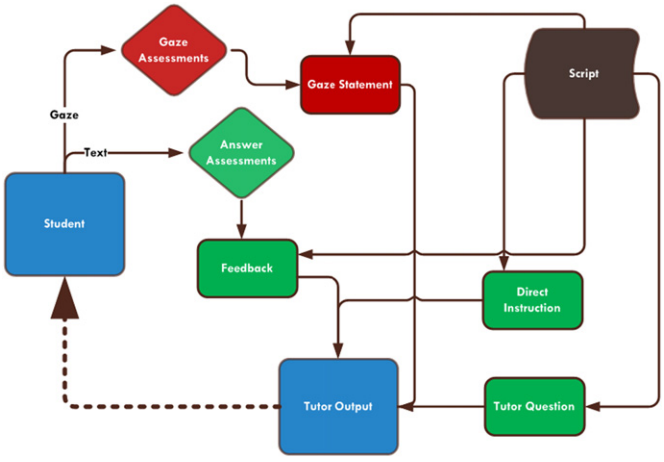

Fig. 3. System architecture.

in the *Script* database, which also specifies the order of delivery.

The most complex aspect of the system is the monitoring of gaze position and the decision to fire a gaze statement. Gaze position is continuously monitored and assessed via the *Gaze Assessments* module. In order to fire a gaze statement the tutor must be speaking (either a statement or a question) and the student must continuously not be looking on screen for an extended period of time. Thus a student who looks back on screen, even briefly, will reset a timer and prevent a gaze statement. When the system decides to fire a *Gaze Statement* and the appropriate text is obtained from the *Script*.

3. Method

3.1. Participants

Participants were 48 undergraduates from a southern university in the US who participated for course credit. There were 30 females and 18 males. There were 14 African-Americans, 27 Caucasians, 4 Hispanics, 2 Asians, and one Native Hawaiian/other Pacific Islander. Participants were between 18 and 36 years old, with a mean age of 20.6 years old ($SD=3.8$).

3.2. Design

The experiment utilized a within-subjects design where participants completed four biology lectures on the topics of: the Golgi body, cytoskeleton, phases of mitosis, and ecological succession. Two of these lectures were completed with the gaze-reactive tutor (GR) and the remaining two with the non-gaze-reactive (NGR) version of the tutor. Assignment of lecture topic to tutor version and order of lecture topic was counterbalanced across participants with a 4×4 Latin Square.

The order of tutor version followed two predefined sequences. Half of the participants interacted with the

GR tutor for the first lecture, the NGR tutor for the second, the GR for the third, and the NGR tutor for the last lecture (GR–NGR–GR–NGR pattern for the gaze-first group). The remaining 24 participants utilized a NGR–GR–NGR–GR pattern (gaze-second group).

3.3. Materials

3.3.1. Knowledge tests

The knowledge tests (used to measure learning gains) were 12-item multiple-choice tests with four alternative answers for each item. There were alternate test versions for pretest and post-test that were counterbalanced across participants. Specifically, half the participants received Test A as pretest and Test B as post-test, while the other half received Test B for the pretest and Test A for the post-test.

There were three questions for each lecture. *Prompt* questions tested participants on content for which the tutor explicitly prompted the student for a response. *Assertion* questions tested participants on content that the tutor explicitly asserted to the student via direct instruction but not with an explicit prompt. Finally, there were *deep reasoning* questions that required causal reasoning and inference rather than recall of shallow facts (e.g., prompt and assertion questions). When possible, the questions were designed to target different knowledge units within the same topic, so it would not be necessary for a student to answer an assertion question correctly before he or she could answer a deep reasoning question. Examples of prompt, assertion, and deep reasoning questions are presented in Appendix A.

3.3.2. Engagement measures

Participants' engagement levels were tracked after each lecture with the *Affect Grid* (Russell et al., 1989) and through a locally created *Post-Lecture Engagement Scale*. The Affect Grid is a validated single item affect measurement instrument consisting of a 9×9 (valence $\times$ arousal) grid, which are the primary dimensions that underlie affective experiences (Barrett, 2009). The arousal dimension ranges from sleepiness to high arousal, while the valence dimension ranges from unpleasant feelings to pleasant feelings. After completing each lecture, participants indicated their affective state by marking an X at the appropriate location on the grid.

The *Post-Lecture Engagement Questionnaire* required participants to self-report their engagement levels after each lecture. There were three questions which asked participants to rate their engagement at the *beginning*, *middle*, and *end* of each lecture. For example, the question, "How engaged were you during the *start* of this session?" was used to track engagement at the start of the lecture (participants were instructed that session pertained to the lecture they just completed). Participants indicated their ratings on a six-point scale ranging from (1) *very bored* to (6) *very engaged*.

3.3.3. Subjective impressions measures

The *Subjective-Impressions Questionnaire* asked participants to evaluate the tutorial session on measures of perceived performance, user satisfaction, and task difficulty. Each participant completed the questionnaire once after each lecture. The questionnaire required participants to answer the following five questions: (1) *Importance*: "How important was it for you to understand the material?", (2) *Usefulness*: "How useful did you find the material covered?", (3) *Interest*: "How interesting did you find the material covered?", (4) *Challenge*: "How challenging did you find the material?", (5) *Effort*. "How much effort did you put into this section?". Participants responded to each question via a six-point scale. For example, the scale for importance ranged from (1) *very unimportant* to (6) *very important* while the scale for effort ranged from (1) *very easy* to (6) *very difficult*.

3.3.4. Individual difference measures

There were three individual difference measures. First, participants' scores on the pretest were used to track *prior knowledge* pertaining to the biology topics covered in the tutorial session. Second, self-reported American College Testing (ACT) or Scholastic Aptitude Test (SAT) scores served as a measure of *aptitude*. The ACT and SAT are standardized tests required for college admissions in the United States. SAT scores were converted to ACT scores using the ACT–SAT concordance chart (ACT–SAT Concordance Chart, 2009). Self-reported ACT and SAT scores have been found to strongly correlate with actual test scores (Cole and Gonyea, 2010), so we have some confidence in this measure.

Participants also completed a locally created *Perceptions of Learning Biology Questionnaire* (PLB). The PLB consisted of three questions that were designed to gage participants' *interest* in learning biology, their perceived *usefulness* of learning biology, and their *confidence* that they could learn biology from a computer tutor. Participants used a six-point scale to provide their responses to these questions.

3.4. Apparatus (eye tracker)

Eye movements were recorded with a Tobii T60^{TM} eye tracker. The eye tracking infrared sensors and cameras are integrated into a 17-in computer monitor so there was no need for any special mounts or head gear. The participants were calibrated before they started the tutorial session. The calibration process consisted of the participant tracking a moving circle around the screen. The calibration process typically took between 15 s and a minute depending on individual differences. Eye tracking data was recorded at 60 Hz (approximately once every 17 ms).

3.5. Data streams recorded

Five streams of information were recorded during the tutorial session. First, participants' gaze-patterns were

recorded with the Tobii T60™ eye tracker. Second, videos of their faces were recorded with a camera that was integrated with the eye tracker system. Third, videos of their computer screens were also recorded with the eye tracker software, Tobii Studio™. The screen video also included the audio generated by the animated pedagogical agent (see Fig. 2). Fourth, log files consisting of the tutor's responses, students' responses, response times, and other interaction parameters were recorded for offline analysis. Fifth, participants' body movements (not relevant to the present paper) were recorded with a custom body posture measurement system (Olney and D'Mello, 2010). Data from the eye tracker, log files, and body movements were synchronized with each other and with the face and screen videos.

3.6. Procedure

Participants were individually tested in a 1.5-h session. Participants first signed an informed consent form and underwent the eye tracking calibration procedure.[3] They then completed a demographics questionnaire which assessed their age, sex, and ethnicity. Next, they self-reported their ACT or SAT scores, completed the Perceptions of Learning Biology Questionnaire (PLB), and answered the multiple-choice pretest.

The tutorial session commenced after the pretesting and calibration procedures. It consisted of three phases for each of the four topics. In Phase 1, participants interacted with the tutor for one of the four biology topics. They immediately evaluated this lecture with the Subjective Impressions Questionnaire (Phase 2). Next, they self-reported their engagement levels with the Post-Lecture Engagement Questionnaire and the Affect Grid (Phase 3). Phases 1–3 were repeated for each of the four topics.

As mentioned in Section 3.2, each participant alternated between gaze-reactive and non-gaze reactive versions of the tutor with equal exposure to both. On average, the participants took 6.75 min (SD=.965) to complete one topic in the non-gaze-reactive condition and 6.95 min (SD=1.29) in the gaze-reactive condition. The difference was not statistically significant (p=.385).

Participants completed the multiple-choice post-test after the tutorial session. They were subsequently debriefed.

4. Results and discussion

Our analyses focused on five questions pertaining to the impact of gaze-reactivity on learning, motivation, and engagement. First, were the gaze-sensitive interventions successful in reorienting students' attention towards the tutor? Second, was the gaze-reactive tutor more effective in promoting learning gains compared to the non-gaze-reactive tutor? Third, did participants evaluate the tutorial sessions more favorably when interacting with a tutor that was sensitive to their gaze-patterns, or did they find the gaze-sensitivity intrusive and possibly annoying? Fourth, did gaze-reactivity increase student engagement? Fifth, did individual differences in prior knowledge, aptitude, and perceptions towards learning biology moderate the effects of gaze-reactivity on learning, motivation, and engagement?

4.1. Gaze-patterns associated with gaze-sensitive dialogs

An analysis of the dialog moves delivered by the gaze-sensitive tutor indicated that 16 out of the 48 students never received a gaze-sensitive move. This suggests that a third of the participants never zoned out sufficiently to warrant corrective action via a gaze-reactive statement. It should be noted that the lack of gaze-reactive statements associated with these participants cannot be attributed to eye tracking failures because we verified that their gaze was being accurately tracked. Since our primary goal was to compare the effects of gaze-reactive dialogs, the subsequent analysis focuses on the remaining 32 students who received at least one gaze-reactive statement.

On average, these students received 5.31 (SD=5.75) gaze-reactive statements. The distribution was highly skewed with the number of statements ranging from 1 to 20. Nineteen of the students (59.4%) received less than four statements, while the remaining 13 students received four or more statements, so there were some individual differences with respect to the efficacy of the gaze-reactive tutor in reorienting students' attention.

There is the important question of how students adapt their gaze patterns in response to the gaze-reactive statements. We addressed this question by contrasting students' gaze patterns before and after receiving gaze-reactive statements. We identified the start (t_s) and end (t_e) of gaze-reactive utterances from the stream of tutor dialog moves extracted from the log files. Turning to the stream of student gaze patterns (i.e., focus on one of the five zones as described in Section 2.3); we examined the probability distribution of gaze events w seconds before (t_{s-w}) and after (t_{e+w}) the gaze-reactive utterances. This process was repeated for window sizes ranging from 7.5 s to 30 s with increments of 7.5 s. Separate probability distributions were computed for each of the 32 students who received at least one gaze-reactive statement. If a student received more than one gaze-reactive statement, then the probability distribution for that student was the average of the distributions associated with each gaze-reactive statement.

The mean probability distributions (averaged across students) across all windows is presented in Fig. 4. Gaze patterns associated with the text box and blank areas of the screen are not included in Fig. 4 because students rarely focused on these zones.

The actual text of a gaze-reactive dialog move is randomly selected, so the lengths of these utterances vary. Hence, for simplicity, the time interval associated with the

[3]The experimenter verified that the participant's gaze could be reliably tracked. Data from participants with tracking difficulties were discarded and not counted in the sample of 48 students.

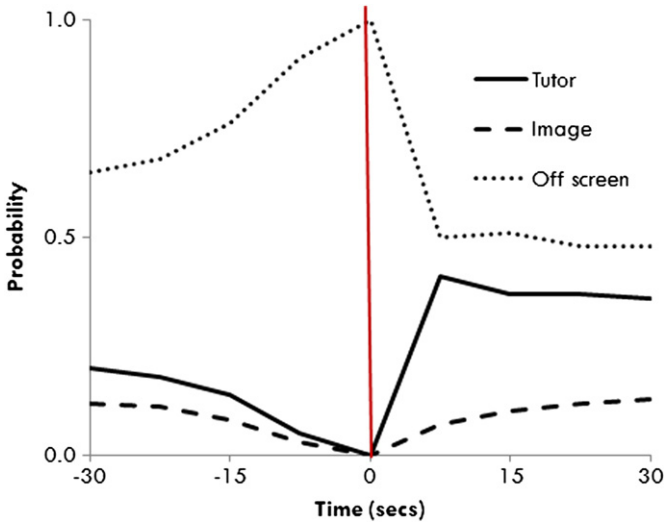

Fig. 4. Probability distribution of gaze-patterns before and after gaze-reactive statements.

synthesis of the gaze-reactive statement is not factored into Fig. 4 (i.e., $t_s = t_e$). This is a valid simplification because we are primarily interested in the change in gaze-patterns *after* the delivery of gaze-reactive statements, instead of gaze-patterns *during* the delivery of these statements.

As Fig. 4 illustrates, prior to receiving a gaze-reactive statement, the probability that students are off-screen steadily increases until it peaks at $t=0$. Focus on the tutor and image steadily decreases until the gaze-reactive statement is launched. A drastic change in gaze-patterns follows the incidence of the gaze-reactive statement. Now off-screen gaze behaviors rapidly decrease, while focus on the tutor steadily increase, until these peak at $t=7.5$ s.

It is important to mention three additional insights that can be gleaned from Fig. 4. First, it takes some time (approximately 7.5 s following the end of the gaze-reactive utterance) for students to drastically reorient their attention towards the interface. Second, although off-screen gaze behaviors were reduced, they did not completely dissipate upon receipt of a gaze-reactive statement. Third, and more importantly, attentional reorientation after the gaze-reactive statement was primarily directed to the tutor (i.e., the source) rather than randomly scattered across the interface (note the drastic increase in tutor-oriented gazes compared to the increase in image-oriented gazes after the delivery of the gaze-statement; the difference is considerably more subtle before the gaze-statement).

As a complement to the descriptive analyses described above, we also performed a series of paired-samples *t*-tests (at the subject level) comparing gaze patterns before and after the gaze-reactive statements. There were statistically significant patterns ($p < .001$) associated with an increased focus on the tutor and a corresponding decrease in off-screen gaze behaviors. The mean effect sizes (across windows) associated with these significant patterns were 1.31 and -1.29 sigma for tutor-focused and off-screen gaze-shifts, respectively. There were no statistical differences associated with the other zones, thereby confirming our initial claim that the gaze statements explicitly reorient attention towards the tutor.

Finally, we investigated whether the number of gaze-reactive statements received was related to students' attentional reorientation patterns. It might be the case that students rapidly redirect their attention towards the tutor when the first few gaze-reactive statements are received, but their attentional reorientation responses might be slower as more statements are encountered. They might also eventually begin to tune these messages out and not reorient attention at all. We investigated this issue by correlating the number of gaze-statements received to the change in probability that students attended to the key areas (tutor, image, off-screen) after receiving a gaze-statement compared to before receiving the statement ($t_{e+w} - t_{s-w}$). Data from the four participants who received more than 10 gaze-statements were identified as outliers and were discarded from the analyses.

The results for the 7.5 s window indicated that the number of gaze-statements was positively correlated with off-screen behaviors ($r=.403$, $p=.033$), negatively correlated with focus on the tutor ($r=-.351$, $p=.067$), and not correlated with focus on the image ($r=-.123$, $p=.533$). Number of gaze-statements marginally[4] correlated with off-screen behaviors ($r=.319$, $p=.098$) and focus on the tutor ($r=-.329$, $p=.088$), but not with focus on the image ($r=.045$, $p=.820$), for the 15-s window. There were no significant or marginally significant correlations for the 22.5 and 30 s windows. The patterns of these correlations indicate that students were *slower* to reorient attention on the tutor as the number of gaze-reactive message increases. However, they do not completely tune out these messages as reorientation patterns were not correlated with the number of gaze-statements for the longer window sizes.

4.2. Learning gains

The gaze-reactive statements did have their desired effect of directing students' attention towards the tutor, but did this attentional reorientation have an impact on learning? We answered this question by examining students' responses to the multiple-choice knowledge tests that were administered before and after the tutorial session. The pretest and post-test were scored for the proportion of questions that students answered correctly. The measure of learning consisted of *proportional learning gains*, computed as: (post-test scores–pretest scores)/(1 − pretest scores). Proportional learning gains represent the degree of improvement at post-test above and beyond pretest performance.

Proportional learning gains scores were separately computed for prompt questions, assertion questions, and deep reasoning questions (see Section 3.3). There was also an overall proportional learning gains score, which made no

[4]It should be noted that these marginally significant effects are likely to be significant with a larger sample.

distinction for question type. Proportional learning gains for the three question types were not significantly correlated ($p > .10$) for either condition, thereby indicating that the different question types were assessing different levels of understanding.

Paired-samples *t*-tests comparing learning gains across conditions resulted in a significant difference for the assertion and deep reasoning questions; these differences were consistent with medium sized effects (see Table 2 and Fig. 5). There was an interesting interaction between learning gains for these two question types. Learning gains for assertion questions, which tap into knowledge of surface level facts, were higher with the non-gaze-reactive tutor. In contrast, students provided more accurate responses to deep reasoning questions when they interacted with the gaze-reactive tutor. These questions are the gold standard for learning and comprehension, so this pattern highlights the benefits of gaze-reactivity.

There are three additional points associated with the patterns in learning gains that are worth mentioning. First, the lack of an effect for prompt questions might be attributed to the fact that the tutor's prompts explicitly command students' attention. Hence, gaze-reactivity does not play a substantial role when students' are tested on content covered with a prompt.

Table 2
Descriptive statistics and paired-samples *t*-tests on proportional learning gains.

Learning measure	Descriptives		Paired-samples *t*-test		
	Non-GR *M* (SD)	GR *M* (SD)	*t*(31)	*p*	*d*
Prompt	.547 (.481)	.484 (.466)	−.626	.268	−.13
Assertion	.531 (.457)	.281 (.621)	−1.72	.048	−.46**
Deep	.000 (.741)	.313 (.632)	1.88	.035	.45**
Overall	.185 (.840)	.369 (.534)	−.939	.178	.26

Note: *GR*=Gaze-reactive.
**$p < .05$ on a one-tailed test.

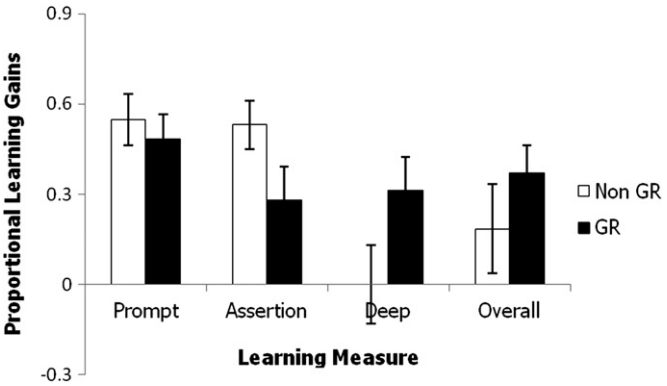

Fig. 5. Proportional learning gains with standard error.

Second, although there was no significant difference associated with overall learning gains, there was a small effect (d=.26 sigma) in favor of the gaze-reactive tutor. A post-hoc power analysis indicated that our sample size of 32 participants yielded a power value of .42 for a one-tailed paired-samples *t*-test with an alpha value of .05. This achieved power is substantially less than the recommended power of .8 (Cohen, 1992), so this difference might be significant with a larger sample.

Third, although the zero mean for deep reasoning questions in the non-gaze-reactive condition might appear to be odd, there was considerable variability associated with this mean. This implies that some students demonstrated some learning in this condition, while other students showed negative learning gains. There is the question of identifying the individual differences associated with these two groups of students, an issue that we address in Section 4.5.

4.3. Subjective impressions of the session

We addressed the question of whether gaze-reactivity influenced students' perceptions of the tutor by examining their responses to the five questions on the Subjective-Impressions Questionnaire (importance, usefulness, interest, challenge, effort; see Section 3.3). An analysis of these five measures yielded some interesting patterns of correlations among the measures. In particular, importance was correlated with usefulness (r=.372), interest (r=.493), effort (r=.254), but not with challenge (r=.020). In general, importance, interest, effort, and usefulness were correlated with one another, but were not correlated with challenge. This pattern suggests that the individuals' measures can be reduced to two underlying components. It is advantageous to extract and focus on these components instead of examining each measure independently because this reduces the number of subsequent statistical tests, thereby alleviating the discovery of spurious effects. Furthermore, it might be the case that the underlying components might have more explanatory power than the individual measures because they provide a composite measure of the underlying constructs.

One simple possibility to extract the components is to consider challenge as one measure and take the average of the importance, usefulness, interest, and effort scores as the second measure. However, this solution does not model the pattern of correlations among these variables. Therefore, latent components were extracted with an exploratory factor analysis (see Appendix B for a brief overview of factor analyses). We used a principal components analysis with varimax rotation and Kaiser normalization.

Several indicators of factorability indicated that the data were in fact factorable (i.e., the assumptions of the factor analysis were satisfied). In particular: (a) 4 out of the 5 included items had a correlation of at least .3 with at least one other item, suggesting reasonable factorability, (b) the Kaiser–Meyer–Olkin measure of sampling adequacy was

.673, which exceeds the recommended value of .6, (c) Bartlett's test of sphericity was significant, $\chi^2(10)=127$, $p<.001$, (d) the diagonals of the anti-image correlation matrix were above .5 for four of the items, and above .27 for the one remaining item (this supports the inclusion of each item in the factor analysis), and (e) the commonalities were all above .4, indicating that each item shared a degree of common variance with the other items.

The analysis yielded two components with eigenvalues greater than 1. These collectively accounted for 69.2% of the variance. Component 1, which accounted for 48.1% of the variance, consisted of four out of the five measures (interest, useful, effort, and importance). Loadings of items onto components are presented in Table 3. One interpretation of these loadings is that Component 1 aligns with *state motivation*. The fifth measure (*perceived challenge*) loaded on Component 2, which explained 21.0% of the variance.

We considered Component 1 to be an assessment of motivation because it involves: (a) expressed interest in the material, (b) perceived importance and usefulness of the material, and (c) an evaluation of effort exerted towards learning the material. It is considered to be a measure of *state* motivation because the items on the questionnaire were specific to learning the material covered in the session that was just completed. This measure is expected to be unstable and malleable compared to *trait motivation*, which is a more stable and rigid motivation to learn biology in general.

Descriptive statistics on the component scores associated with the two conditions are presented in Table 4 (the engagement row is described in Section 4.4). It should be noted that the scores can be negative as they are standardized scores. Paired-samples *t*-tests did not yield any significant ($p>.05$) differences in perceptions of either tutor, although there was a small effect of −.21 sigma, which is in favor of the non-gaze-reactive tutor.

Table 3
Coefficients (factor loadings) of items from the Subjective-Impressions Questionnaire.

Measure	Component 1 (State motivation)	Component 2 (Perceived challenge)
Importance	**.284**	.117
Usefulness	**.335**	−.082
Interest	**.361**	−.143
Effort	**.305**	.180
Challenge	.054	**.942**

Table 4
Descriptive statistics and paired-samples *t*-tests on subjective measures.

Measure	Descriptives		Paired-samples *t*-test		
	Non-GR	GR			
	M (SD)	*M* (SD)	*t*(31)	*p*	*d*
State motivation	−.215 (.990)	−.402 (.909)	−1.39	.173	−.21
Perceived challenge	.002 (1.04)	.001 (.956)	−.009	.993	.00
Engagement	−.264 (.861)	−.474 (.930)	−1.35	.187	−.19

Note: *GR*=Gaze-reactive.

4.4. Engagement

Gaze-reactivity had minimal impact on students' state motivation and perceptions of challenge, but did it facilitate or hinder their engagement levels? This question was addressed by examining students' responses to the Post-Lecture Engagement Questionnaire and the Affect Grid (see Section 3.3). The measures included valence, arousal, and self-reported engagement at the beginning, middle, and end of each lecture. There were strong correlations among these measures, so we proceeded by conducting an exploratory factor analysis on the five engagement measures. All requirements of factorability were satisfied, so we proceeded by examining components with eigenvalues greater than one. There was one such component, which explained a robust 72.0% of the variance. Since all engagement related variables loaded onto this component,[5] we subsequently refer to this component as *engagement*.

Paired-samples *t*-tests comparing engagement across tutors did not yield a significant ($p>.05$) difference (see Table 4), although there was a small effect suggesting that students reported more engagement with the non-gaze-reactive tutor. Taken together, the results associated with learning, state motivation, incremental challenges, and engagement suggest that gaze-reactivity has a positive impact on objective measures of learning (deep learning gains), but does not appear to have an impact on the subjective measures.

4.5. Individual differences

We investigated if individual differences in prior knowledge, aptitude, interest in learning biology, perceived usefulness of learning biology, and confidence in learning biology from a computer tutor influenced the impact of gaze-reactivity on learning, state motivation, perceived challenge, and engagement. There were interesting patterns of correlation among these measures so our analyses proceeded by conducting an exploratory factor analysis on the five individual difference measures. All requirements of factorability were satisfied, so we proceeded by examining the two components with eigenvalues greater than one. These components explained 71.1% of the variance.

The coefficients of these two individual difference components are presented in Table 5. Component 1 (*trait motivation*), which explained 48.1% of the variance, is consistent with motivated students who considered

[5]Engagement Component=.202 Valence+.211 Arousal+.240 Beginning Engagement+.253 Middle Engagement+.256 End Engagement. All variables are standardized prior to computing the component scores.

Table 5
Coefficients (factor loadings) of individual difference items.

Measure	Component 1 (Trait motivation)	Component 2 (Aptitude)
Useful	**.500**	−.267
Interest	**.404**	.000
Aptitude	−.219	**.713**
Confidence	.143	**.349**
Prior knowledge	.192	**.255**

learning biology to be motivating and useful. This component is referred to as trait motivation because it represents more stable and dispositional attitude towards learning biology rather than unstable and situational state motivation. Aptitude negatively loaded onto this component, but was the major factor for Component 2. This component accounted for 23.0% of the variance. The coefficients of confidence and prior knowledge were in the same direction for both components, but the magnitude of these coefficients was higher for the aptitude component (Component 2).

Our analysis proceeded by identifying whether the individual differences associated with trait motivation and aptitude *moderated* the effects of gaze-reactivity on our dependent variables. These included overall learning gains, state motivation, perceived challenge, and engagement. We focused on overall learning gains in order to reduce the number of analyses. A moderation analysis investigates whether a moderating variable alters the strength of the causal influence of an independent variable on a dependent variable (Aguinis, 2004). For the present analyses, the two individual difference measures were the moderator variables, an indicator variable for gaze-reactivity (non-gaze-reactive=0, gaze-reactive=1) was the independent variable, and the four measures listed above were the dependent variables.

The effect of moderator M on the relationship between independent variable X and dependent variable Y is obtained from the interaction term $(X \times M)$ of a linear regression model. In particular, we would obtain a moderation effect if coefficient c in the following regression model was statistically significant: $Y = aX + bM + c\ (XM) + e$.

We tested for individual difference moderation effects by performing eight multiple regression analyses (4 dependent variables × 2 moderator variables). All interaction terms were mean-centered prior to computing the analyses (Cohen et al., 2002). Of primary interest is the interaction term, which was not statistically significant ($p > .05$) in any of the analyses with trait motivation as the moderating variable. Trait motivation apparently did not moderate the influence of gaze-reactivity on learning, state motivation, perceived challenge, and engagement.

One out of the four models that tested the influence of aptitude as a moderator yielded a significant ($p = .012$) interaction term, as well as an overall significant model $F(2, 60) = 2.78$, $p = .049$. The dependent variable in this model was overall learning gains, which is a key variable of interest. The model explained 12.2% of the variance, with an impressive 9.9% of the variance being explained by the interaction term.

We performed a simple slopes analysis in order to identify how aptitude moderates the effect of gaze-reactivity on overall learning. In this context, the simple slopes analysis investigates the nature of the relationship between gaze-reactivity and overall learning gains for different values of aptitude (typically one standard deviation above and below the mean).

Fig. 6 presents the relationship between gaze-reactivity and overall learning one standard deviation below, above, and at the aptitude mean. We note that there is a (small) positive but non-significant slope for mean aptitude ($B = .18$, $p = .281$). This implies that gaze-reactive dialogs have a very small impact on overall learning for the average student. More importantly, these dialogs were very successful at facilitating learning for students with aptitudes one standard deviation above the mean ($B = .63$, $p = .011$). The impact is even more pronounced for the gifted students with aptitude levels two standard deviations above the mean ($B = 1.07$, $p = .007$; not shown in Fig. 6).

On the other hand, there is a negative relationship between gaze-reactivity and overall learning gains for the less gifted students (i.e., lower aptitude). The simple slope one standard deviation below the aptitude mean failed to reach significance ($B = -.26$, $p = .284$), but the slope two standard deviations below was marginally significant ($B = -.70$, $p = .070$).

These patterns indicate that the positive relationship between gaze-reactivity and learning observed for students with high aptitude is stronger than the negative relationship for the less gifted students (simple slopes were .63 and −.26 +1 and −1 SD below the aptitude mean, respectively).

Finally, we investigated whether individual differences in trait motivation and aptitude interacted with each other to predict the seven dependent variables of interest (i.e., state

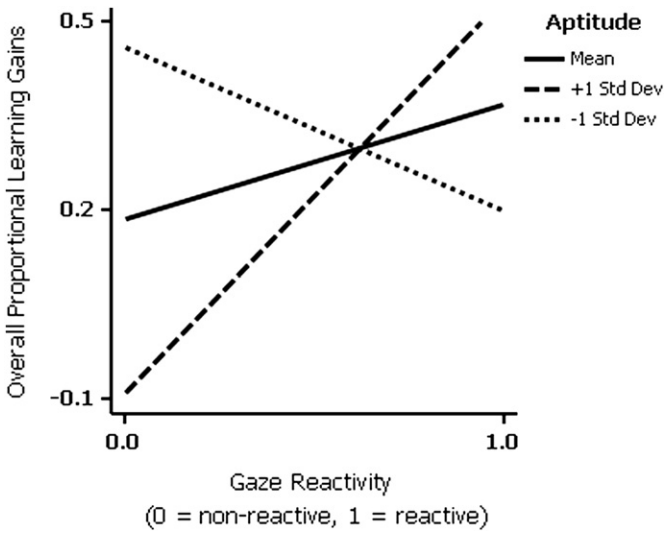

Fig. 6. Aptitude × gaze-reactivity interaction.

motivation, perceived challenge, engagement, and the four measures of learning). There were, however, no significant interactions, so these analyses are not reported here.

5. General discussion

The present paper described the design and evaluation of a novel ITS which dynamically monitored and responded to student disengagement via eye tracking. The results support a number of conclusions pertaining to the efficacy of gaze-reactivity in promoting learning, motivation, and engagement. The subsequent discussion focuses on some of the most significant findings, followed by an analysis of some of the limitations, possible avenues for future work, and some concluding remarks.

5.1. Overview of major findings

The results support three major conclusions pertaining to the gaze-reactivity with respect to (a) attentional reorientation patterns, (b) learning gains, state motivation, perceived challenge, and engagement, and (c) individual differences. These are addressed below.

5.1.1. Attentional reorientation patterns

There was some uncertainty as to the exact effect of the gaze-sensitive statements prior to testing its effects on students. Would students pay attention to the gaze-sensitive statement and productively change their behavior? Or would they simply consider the statement to be an intrusive annoyance and continue to disengage? If behavior was changed, and students did in fact pay more attention to the tutor, was this a lasting change? Or would they quickly revert to their previous state of disengagement? Another possibility is that students could have zoned out to a point where they have essentially tuned the tutor out and did not even actively comprehend the gaze-reactive statement. Yet another possibility is that students consciously looked away after receiving the gaze-reactive statement as an act of defiance.

The results from Section 4.1 indicated that in general just-in-time gaze-reactivity was quite effective in reorienting students' attention towards the tutor. However, off-screen gaze-behaviors still persisted even after students were explicitly instructed to pay attention to the tutor. There were also individual differences in the efficacy of the gaze-reactive dialogs. Some students rapidly corrected their behaviors with a single cue, others required more than one cue, and a few never adapted their behavior. The task of identifying the individual differences that are associated with these different classes of students warrants further research.

5.1.2. Learning gains, state motivation, perceived challenge, and engagement

Our second important finding was that gaze-reactivity positively influenced learning gains, particularly at deeper levels of comprehension. This discovery of a medium-sized deep-learning effect with gaze-reactive dialogs represents a major advantage to this form of *direct* disengagement repair. Although this finding warrants replication in other learning environments and with different student populations, there is the important question of why simply directing students to reorient their attention had such a significant effect?

One possibility is that students are presumably not expecting a computer tutor to monitor their gaze-patterns and explicitly instruct them to focus on the material. This unexpected level of intelligence from a computer might have motivated them to focus on the tutor and process the material more deeply. If this is the case, then increased attention on the tutor should positively predict learning. This prediction was confirmed in a follow-up analysis that yielded a significant correlation between proportional focus on the tutor and deep learning gains ($r=.362$, $p=.042$) as well as overall learning gains ($r=.465$, $p=.007$). Focus on the image and text box were not statistically related to learning gains. Taken together, the data suggest that just-in-time gaze-reactivity directs students' attention towards the tutor and this attentional reorientation is what correlates with learning.

While the discussion above seems to extol the virtues of the gaze-reactive tutor, there are some caveats to direct gaze-reactive statements as well. In particular, students' performance on assertion questions was lower with the gaze-reactive tutor. There was also a small non-significant trend towards students reporting more state motivation and engagement after interacting with the non-gaze-reactive tutor. Although, state motivation and engagement did not correlate with deep learning gains, it appears that some students might not be amenable to the directness of the gaze-sensitive dialogs. Perhaps an alternate strategy that politely encourages students to focus on the tutor might be warranted. Indeed, politeness has been found to be effective in facilitating learning (Brown and Levinson, 1987; Wang et al., 2008), however, it is an open question as to whether polite alternatives will be as effective as the more direct gaze-sensitive statements.

5.1.3. Individual differences

Our third important finding pertained to the discovery of a significant aptitude × treatment interaction with respect to the impact of gaze-reactivity on overall learning gains. This interaction suggests that while gaze-reactivity was associated with a small improvement in overall learning for students with average aptitude, learning gains were statistically significant and substantially higher for students with high aptitude, and slightly lower for their counterparts, although this effect was marginally significant.

It is important to note that aptitude did not moderate the influence of gaze-reactivity on any of the other variables. Therefore, it would be inappropriate to claim that students with high aptitude learned more from the gaze-reactive tutor because they preferred this tutor or

experienced higher engagement with this tutor. Aptitude was also not correlated with the number of gaze statements ($r = .001$, $p = .996$). Since gaze-reactive statements were provided in response to disengagement behaviors, aptitude does not appear to explain students' tendencies to disengage from the tutor.

There is also the possibility that aptitude might influence how students process the gaze-reactive statement. That is, are there aptitude differences associated with gaze patterns following gaze-reactive statements? We addressed this question with a follow-up-analysis which first involved assigning students to a low or high aptitude group (based on a median split on aptitude scores). Independent-samples *t*-tests were used to test for aptitude differences in gaze patterns 7.5 s after the gaze-reactive statement. There were no aptitude related differences pertaining to focus on the tutor, the text box, the blank area, and gazing off-screen.

There was, however, a marginally significant effect for attention on the image. It appears that the high aptitude students were more likely to focus on the image than the low aptitude students soon after receiving a gaze-reactive statement ($M = .123$, $SD = .199$ for high aptitude and $M = .032$, $SD = .050$ for low aptitude; $t(30) = -1.88$, $p = .069$, $d = .63$). Integrating the spoken content of the lecture with the image on the screen is an essential element of learning biology. The fact that gaze-reactivity was more effective in promoting speech-content (text-diagram) integration for the high aptitude students is one potential explanation of this aptitude × treatment effect.

5.2. Limitations, unanswered questions, and future work

There are four primary limitations with the present research. These, along with potential solutions, are discussed here.

5.2.1. Indirect inference of boredom

The first limitation pertains to the fact that boredom, disengagement, and zoning out were not directly measured but were inferred from students' gaze-patterns. More specifically, we assumed that students were not engaged in the learning session if they were looking away from the screen for an extended and contiguous block of time. We considered this to be an appropriate assumption because most would agree that a noticeable lack of eye contact with a conversational partner (computer tutor in this case) is a defensible sign of disengagement unless one is trying to be covert or deceptive. Nevertheless, future research should explicitly verify this assumption with measures of boredom, engagement, interest, and other relevant emotions.

This can be accomplished in a number of ways. One option is for the tutor to simply ask students if they are bored before providing gaze-reactive statements. This approach, however, relies on the honesty of the students, has the potential of interrupting the primary task of learning, and cues students to the fact that the tutor is explicitly monitoring their boredom levels. Alternatively, sensors that track facial features, physiology, reaction time, and other informational channels could be used to corroborate the gaze-based diagnosis of boredom (Beck, 2005; Cocea and Weibelzahl, 2009; D'Mello and Graesser, 2010b; Drummond and Litman, 2010; Jacobs et al., 2009).

Perhaps a simpler approach would involve a post-hoc verification of the assumption that persistent off-screen gaze-patterns correlate with disengagement. This could be achieved via a retrospective affect judgment protocol (D'Mello and Graesser, 2010b), where videos of students' faces captured *during* the learning session are replayed *after* the session. Students make boredom judgments over the course of viewing these videos. These *offline* boredom judgments can then be compared to the tutor's *online* gaze-based predictions of boredom.

5.2.2. Within-subjects experimental design

The second limitation with this research can be linked to the within-subjects methodology used to test the causal link between gaze-reactivity and learning. The value of the repeated measures (within-subjects) design is that it controls for participant variability and allows us to assess how particular individuals differ with respect to treatment and control. However, practice effects and carry-over effects are common disadvantages of within-subjects designs. These problems might be applicable to the current experiment as well.

A carry-over effect might have occurred for students who first interacted with the gaze-sensitive tutor (Session 1) followed by the non-gaze-sensitive tutor (Session 2). Since students were blind to condition (i.e., they did not know that there were two tutors until the debriefing), a gaze-sensitive statement during Session 1 might have impacted behavior and performance in Session 2.

Similarly, a practice effect might have caused performance to improve from Session 1 to Session 2. Performance might have improved as knowledge of biology accrued, but it might also have worsened due to fatigue and burnout.

Although the use of a counterbalancing scheme to determine the ordering of tutors (see Section 3.2) somewhat reduces the impact of these effects, we performed a follow-up analysis to explicitly assess if there were any carry-over or practice-effects. Each participant was assigned to either a gaze-first or gaze-second group, with respect to whether they first interacted with the gaze-reactive tutor followed by the non-gaze-reactive tutor, or vice versa. An independent-samples *t*-test yielded a significant group-difference associated with tutor-directed gaze patterns. In particular, students in the gaze-first group were more likely to focus on the tutor during both conditions; this suggests that there were some carry-over effects.

Fortunately, tests for group-differences (gaze-first vs. gaze-second) across the seven dependent variables did not yield any significant effects ($p > .05$), so we have some confidence that our major findings cannot be simply attributed to methodological complications. Nevertheless,

it would be desirable if the major patterns were replicated in an experiment that implemented a between-subjects design, where students are randomly assigned to a tutor that is gaze-reactive or to one that is not.

5.2.3. Scalability concerns

One disadvantage of commercially available eye trackers is that they are expensive, need expensive hardware and software, and require some expertise to correctly operate. This raises some practical concerns for those who want to extend this program of research into classrooms.

Fortunately, recent advances in cost-effective eye tracking address this concern in a significant way. Opengazer (Zieliński, 2010) and TrackEye (Zafer, 2010) are two freely available software programs for eye tracking. These systems perform lower-precision eye tracking by utilizing inexpensive commercially available web cameras, which are integrated into most laptops. The lower-precision is not a major concern for the gaze-sensitive tutor because it only needs to infer whether the student is gazing at large regions of the screen or is looking elsewhere. Developing a version of the tutor that uses one of these freely available eye trackers is the next step forward.

5.2.4. Lack of sensitivity to all students

One limitation of the experiment is that 16 of the 48 participants (33.3%) were excluded from the analyses because they did not receive a single gaze-reactive statement. This could have occurred because these participants were never sufficiently bored enough to trigger a gaze-reactive statement. Alternatively, they could have been bored, but their disengagement behaviors in terms of gaze patterns were not captured by the gaze-sensitive system. Although further research is needed to decide among these two alternatives, the fact that a third of the participants had to be excluded from the analyses is an important limitation of the experiment. It might also be a limitation of the current version of the gaze-reactive system, because it implies that though there is evidence in favor of the system, the system itself might not be applicable to all students.

5.2.5. Unanswered questions and new opportunities

In addition to yielding some important insights into the feasibility of gaze-reactivity as a mechanism to diagnose and alleviate boredom, the present study also generated some important questions that warrant further research. One question pertains to understanding why performance on the assertion questions was higher when students interacted with version of the tutor that was not sensitive to their gaze patterns?

There is also the issue of identifying specific areas of the content that triggered boredom and the resultant gaze-reactive dialogs. It would also be informative to identify how students' attentional reorientation behaviors at these critical junctures were related to learning. Unfortunately, the small number of questions on the knowledge tests (3 for each topic) makes such an analysis unfeasible with the currently available data. Indeed, a more content-focused fine-grained analysis of the tutorial session with respect to disengagement tendencies, gaze-reactivity, attentional reorientation, and learning is an important item for future work.

The small non-significant state motivation effect also warrants further consideration. One possibility is that the items included on the questionnaire did not tap into the relevant dimensions of their impressions of the session. One potential way to alleviate this concern is to include free response questions as well as expanding the scope of questions on the Subjective-Impressions Questionnaire. It might also be useful to implement think-aloud protocols (Ericsson and Simon, 1993) in order to tap into students' moment-by-moment cognitive processes while they interact with the gaze-reactive tutor.

Our results also indicated that the gaze-reactive dialogs were more effective for high aptitude students. This effect might be attributed to the fact that the gaze-reactive statements only instructed students to pay attention but did not provide any instructions on how to focus attentional resources. High ability students might have been able to interpret this general instruction more effectively because they might be more skilled in allocating attentional resources. Low ability students might need greater individual adaptation, perhaps in the form of explicit instructions on what specifically to focus on. For example, the tutor might have said: "Please pay attention. It's important that you understand how chromatids work in mitosis. I'm going to tell you about how they form."

It is also possible that gaze sensitivity might have to be tailored to address differences in motivation as well as interactions between aptitude and motivation. For example, it is unlikely that the same set of gaze-reactive dialogs that are effective for high aptitude students who lack intrinsic motivation (i.e., gifted but lazy students) might also be effective for motivated by low aptitude students. Hence, further research is needed to address the general question of how to adapt gaze-reactive statements so they are tailored to individual student's abilities and needs.

5.3. Concluding remarks

As most people in the field of education will attest, the task of keeping students engaged in educational activities is extremely challenging. Establishing and maintaining student engagement is especially critical with ITSs and other computer-based learning environments because students can end the session at will when they are outside of the laboratory and are no longer under the watchful eye of an experimenter. The engagement problem is undoubtedly more severe in situations where the computer tutor does most of the talking as is the case when collaborative lectures are delivered to remedial students. Keeping students engaged so they become actively involved in their own learning instead of being passive information receivers is an important challenge

for next-generation ITSs that aspire to impact motivation and emotion in addition to cognitive states.

The research community is taking heed of this challenge by developing affect-sensitive ITSs that detect and respond to the negative emotions that are inextricably bound to learning (Afzal and Robinson, 2009; Burleson and Picard, 2007; Calvo and D'Mello, 2011; Conati and Maclaren, 2009; D'Mello and Graesser, 2010b; D'Mello et al., 2010b; Forbes-Riley et al., 2008; Robison et al., 2009; Woolf et al., 2010). Our gaze-sensitive ITS complements the emerging research in this area by providing one possible solution to the disengagement problem. The next step is to improve the system so that students' perceptions of the session improve in conjunction with advances in learning gains. This is a critical step because although state motivation may not be linked to learning in the short-term, it will undoubtedly influence long-term use and acceptance of the tutor, and consequently impact learning as well.

Finally, it is important to emphasize that while the present research tested one potential intervention to increase engagement, there are several alternate strategies that can be implemented. In lieu of the rather direct gaze-reactive dialog, the tutor could have used less direct and more polite statements (Brown and Levinson, 1987; Wang et al., 2008). While the direct statements were effective for the high aptitude students, perhaps polite statements might be more effective for students with lower scholastic aptitude scores. Indeed, there is some evidence that students with low prior knowledge or students who make the most errors learn best from tutors that use polite compared to direct language (McLaren et al., 2011a, 2011b).

Importantly, disengagement-repair interventions do not have to be restricted to gaze-reactive statements. The tutor could have highlighted relevant areas in the image, asked the student a question, provided a domain-relevant puzzle or challenge, or launched an alternate task that would increase cognitive arousal such as an engaging animation or a simulation (Dickey, 2005; Gee, 2003). On a somewhat different front, another strategy is to provide students with just-in-time training on emotion regulation strategies (e.g., cognitive reappraisal) to help them manage their boredom (Gross, 2008; Strain and D'Mello, 2011).

Some of the emerging theories on emotions and learning also provide some useful recommendations. According to the control-value theory of academic emotions, engagement is positively influenced by students' perceived control of and value in the learning activity (Pekrun, 2006, 2010). Interventions that increase control, such as freedom of choice on the learning tasks (Cordova and Lepper, 1996), and increase perceived value such as aligning topics with interests to increase intrinsic motivation (Hidi and Renninger, 2006; Hulleman et al., 2008) might also be effective in promoting engagement.

According to flow theory (Csikszentmihalyi, 1975, 1990), a balance between challenge and skill is the key to keeping students engaged. More specifically, engagement is high when challenge slightly exceeds skills but boredom is prominent when skill greatly exceeds challenge. Therefore, dynamically selecting learning tasks that are sensitive to individual students zone of proximal development is yet another strategy to enhance engagement (Brown et al., 1998; Vygotsky, 1986).

In summary, there appear to be a number of strategies that attempt to *proactively* increase engagement as well as interventions that *reactively* respond to boredom when it inevitably occurs. Although future research is needed to comparatively evaluate the efficacy of these strategies, it is likely that a combination of both proactive as well as reactive interventions might be needed to have a lasting impact on both short- and long-term engagement.

Acknowledgments

We would like to thank the Associate Editor Carolyn Penstein Rosé and the anonymous reviewers whose comments and suggestions significantly improved this paper.

This research was supported by the National Science Foundation (NSF) (HCC 0834847 and DRL 1108845) and Institute of Education Sciences (IES), US Department of Education (DoE), through Grant R305A080594. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF, IES, or DoE.

Appendix A. Sample test questions from Mitosis Lecture (correct answers are bolded)

1. (*Prompt question*) What kind of hormone would be responsible for telling a cell to grow?
   a. An enzymatic hormone
   b. **A growth hormone**
   c. A receptor hormone
   d. A signal hormone
2. (*Assertion question*) Mitosis is the splitting of a cell's:
   a. Cytoplasm
   b. Organelles
   c. **Nucleus**
   d. Membrane
3. (*Deep question*) Which of the following does not occur in mitosis?
   a. Condensation of the chromosomes
   b. **Replication of the DNA**
   c. Separation of sister chromatids
   d. Separation of the centrosomes

Appendix B. Brief description of factor analysis

A factor analysis is a widely used statistical method to identify a small number of latent unobserved variables that model the variability associated with a larger number of observed variables. For example, a researcher interested in

identifying an individual's mood state (i.e., positive or negative) might ask the person to rate the extent to which they feel delighted, cheerful, excited, blue, downhearted, and lonely. If these adjectives are selected correctly, then ratings for delighted, cheerful, and excited (adjectives describing positive affect) are expected to moderately to strongly correlate with each other but not correlate with ratings for blue, downhearted, and lonely (adjective describing negative affect). Much like coefficients of a linear regression, the latent factor for positive affect will be modeled as a linear combination of the observed variables (i.e., large positive coefficients for cheerful, excited, and delighted but negative or negligible coefficients for blue, downhearted, and lonely). A reverse pattern of coefficients will be obtained for the factor describing negative effect. Thus, the set of six observed emotional adjectives has been reduced to two latent factors (or components).

There are seven main steps involved in a factor analysis. These are listed below:

1. *Collect data*: Collect measures for factor analysis.
2. *Obtain correlational matrix*: This step involves computing a correlation matrix between the measures that will be included in the factor analysis.
3. *Assess factorability of data*: This step involves ensuring that the data does not violate any of the requirements of a factor analysis. For example, there needs to be a modest correlation among variables for a factor analysis to be successful.
4. *Determine number of factors*: Specify the number of factors to be extracted using some criteria (e.g., eigenvalues of extracted factors should be greater than 1).
5. *Extract initial factors*: Extract factors using one of several methods such as principal components analysis or maximum likelihood extraction.
6. *Rotate factors*: This step is important to help with the interpretation of the factors. It involves rotating the extracted factors using one or more rotation methods to yield a final solution. Examples of rotation methods include varimax for orthogonal factors and direct quartimin for oblique rotations (i.e., correlated factors).
7. *Interpret factors*: The measures used to generate the factor analysis are linearly related to the extracted factors. Examining the patterns in the relationships between measures and factors is used to interpret the factors.

References

ACT–SAT Concordance Chart, 2009. Retrieved April 26, 1010, from ACT 〈http://www.act.org/aap/concordance/〉.

Afzal, S., Robinson, P., 2009. Natural affect data—collection & annotation in a learning Context. Paper presented at the Proceedings of 2009 International Conference on Affective Computing & Intelligent Interaction, Amsterdam.

Aguinis, H., 2004. Moderated Regression. Guilford, New York.

Asteriadis, S., Karpouzis, K., Kollias, S., 2009a. Feature extraction and selection for inferring user engagement in an HCI environment. In: Jacko, J.A. (Ed.), Human–Computer Interaction, Part I, vol. 5610. Springer-Verlag, Berlin, pp. 22–29.

Asteriadis, S., Tzouveli, P., Karpouzis, K., Kollias, S., 2009b. Estimation of behavioral user state based on eye gaze and head pose-application in an e-learning environment. Multimedia Tools and Applications 41 (3), 469–493.

Baker, R., D'Mello, S., Rodrigo, M., Graesser, A., 2010. Better to be frustrated than bored: the incidence and persistence of affect during interactions with three different computer-based learning environments. International Journal of Human–Computer Studies 68 (4), 223–241.

Barrett, L.F., 2009. Variety is the spice of life: a psychological construction approach to understanding variability in emotion. Cognition & Emotion 23 (7), 1284–1306.

Beck, J.E., 2005. Engagement tracing: using response times to model student disengagement. In: Looi, C.K., McCalla, G., Bredeweg, B., Breuker, J. (Eds.), Artificial Intelligence in Education—Supporting Learning through Intelligent and Socially Informed Technology, vol. 125. IOS Press, Amsterdam, pp. 88–95.

Berlyne, D., 1978. Curiosity in learning. Motivation and Emotion 2, 97–175.

Bickmore, T., Mauer, D., Crespo, F., Brown, T., 2007. Persuasion, task interruption and health regimen adherence. In: Kort, Y.D., IJsselsteijn, W., Midden, C., Eggen, B., Fogg, B.J. (Eds.), Proceedings of the Second International Conference on Persuasive Technology, vol. 4744. Springer-Verlag, Berlin, Heidelber, pp. 1–11.

Bickmore, T., Schulman, D., Yin, L., 2010. Maintaining engagement in long-term interventions with relational agents. Applied Artificial Intelligence 24 (6), 648–666.

Bickmore, T.W., Picard, R.W., 2005. Establishing and maintaining long-term human–computer relationships. ACM Transactions on Computer–Human Interaction 12 (2), 293–327.

Brown, A., Ellery, S., Campione, J., 1998. Creating zones of proximal development electronically in thinking practices in mathematics and science learning. In: Greeno, J., Goldman, S. (Eds.), Mahawah. Lawrence Erlbaum, NJ, pp. 341–368.

Brown, P., Levinson, S., 1987. Politeness: Some Universals in Language Usage. Cambridge University Press, Cambridge.

Burleson, W., Picard, R., 2007. Evidence for gender specific approaches to the development of emotionally intelligent learning companions. IEEE Intelligent Systems 22 (4), 62–69.

Cade, W., Copeland, J., Person, N., D'Mello, S., 2008. Dialogue modes in expert tutoring. In: Woolf, B., Aimeur, E., Nkambou, R., Lajoie, S. (Eds.), Proceedings of the Ninth International Conference on Intelligent Tutoring Systems. Springer-Verlag, Berlin, Heidelberg, pp. 470–479.

Calvo, R., D'Mello, S. (Eds.), 2011. Springer, New York.

Calvo, R.A., D'Mello, S.K., 2010. Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Transactions on Affective Computing 1 (1), 18–37.

Chaffar, S., Derbali, L., Frasson, C., 2009. Inducing positive emotional state in intelligent tutoring systems. In: Dimitrova, V., Mizoguchi, R., Du Boulay, B., Graesser, A. (Eds.), Proceedings of 14th International Conference on Artificial Intelligence in Education. IOS Press, Amsterdam, pp. 716–718.

Chi, M., Roy, M., Hausmann, R., 2008. Observing tutorial dialogues collaboratively: insights about human tutoring effectiveness from vicarious learning. Cognitive Science 32 (2), 301–341.

Cocea, M., Hershkovitz, A., Baker, R.S.J. d., 2009. Off-task and gaming behaviors on learning: immediate or aggregate? In: Dimitrova, V., Mizoguchi, R., Boulay, B.D., Graesser, A. (Eds.), Proceedings of the 14th International Conference on Artificial Intelligence in Education. IOS Press, Amsterdam, pp. 507–514.

Cocea, M., Weibelzahl, S., 2009. Log file analysis for disengagement detection in e-Learning environments. User Modeling and User-Adapted Interaction 19 (4), 341–385.

Cohen, J., 1992. A power primer. Psychological Bulletin 112 (1), 155–159.

Cohen, P., Cohen, J., West, S., Aiken, L., 2002. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences 3rd ed. Taylor & Francis, Inc..

Cohen, P., Kulik, J., Kulik, C., 1982. Educational outcomes of tutoring: a meta-analysis of findings. American Educational Research Journal 19 (2), 237–248.

Cole, J.S., Gonyea, R.M., 2010. Accuracy of self-reported SAT and ACT test scores: implications for research. Research in Higher Education 51 (4), 305–319.

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., Merten, C., 2007. Eye-tracking for user modeling in exploratory learning environments: an empirical evaluation. Knowledge-Based Systems 20 (6), 557–574.

Corbett, A., 2001. Cognitive computer tutors: solving the two-sigma problem. In: Bauer, M., Gmytrasiewicz, P., Vassileva, J. (Eds.), Proceedings of Eighth International Conference on User Modeling. Springer, Berlin/Heidelberg, pp. 137–147.

Cordova, D.I., Lepper, M.R., 1996. Intrinsic motivation and the process of learning: beneficial effects of contextualization, personalization, and choice. Journal of Educational Psychology 88 (4), 715–730.

Craig, S., Graesser, A., Sullins, J., Gholson, J., 2004. Affect and learning: an exploratory look into the role of affect in learning. Journal of Educational Media 29, 241–250.

Csikszentmihalyi, M., 1975. Beyond Boredom and Anxiety. Jossey-Bass, San Francisco, CA.

Csikszentmihalyi, M., 1990. Flow: The Psychology of Optimal Experience. Harper and Row, New York.

D'Mello, S., Graesser, A., 2010a. Modeling cognitive-affective dynamics with Hidden Markov Models. In: Catrambone, R., Ohlsson, S. (Eds.), Proceedings of the 32nd Annual Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 2721–2726.

D'Mello, S., Graesser, A., 2010b. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-adapted Interaction 20 (2), 147–187.

D'Mello, S., Hays, P., Williams, C., Cade, W., Brown, J., Olney, A., 2010a. Collaborative lecturing by human and computer tutors. In: Kay, J., Aleven, V. (Eds.), Proceedings of 10th International Conference on Intelligent Tutoring Systems. Springer, Berlin/Heidelberg, pp. 609–618.

D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., 2010b. A time for emoting: when affect-sensitivity is and isn't effective at promoting deep learning. In: Kay, J., Aleven, V. (Eds.), Proceedings of 10th International Conference on Intelligent Tutoring Systems. Springer, Berlin/Heidelberg, pp. 245–254.

D'Mello, S., Olney, A., Person, N., 2010c. Mining collaborative patterns in tutorial dialogues. Journal of Educational Data Mining 2 (1), 1–37.

D'Mello, S.K., A meta-analysis on the incidence of emotions during complex learning, in review.

D'Mello, S.K., Lehman, B., Person, N., 2010d. Expert tutors feedback is immediate, direct, and discriminating. In: Murray, C., Guesgen, H. (Eds.), Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference. AAAI Press, Menlo Park, California, pp. 595–560.

D'Mello, S., Chipman, P., Graesser, A., 2007. Posture as a predictor of learner's affective engagement. In: McNamara, D., Trafton, G. (Eds.), Proceedings of the 29th Annual Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 905–991.

D'Mello, S., Graesser, A., 2011. The half-life of cognitive-affective states during complex learning. Cognition & Emotion 25 (7), 1299–1308.

D'Mello, S., Graesser, A., 2012. Emotions during learning with AutoTutor. In: Durlach, P., Lesgold, A. (Eds.), Adaptive Technologies for Training and Education Cambridge University Press, New York NY, pp. 117–139.

de Koning, B.B., Tabbers, H.K., Rikers, R., Paas, F., 2010. Attention guidance in learning from a complex animation: seeing is understanding?. Learning and Instruction 20 (2), 111–122.

de Melo, C., Carnevale, P., Gratch, J., 2010. The influence of emotions in embodied agents on human decision-making. In: Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A. (Eds.), Proceedings of the 10th International Conference on Intelligent Virtual Agents. Springer, Berlin, Heidelberg, pp. 357–370.

del Soldato, T., du Boulay, B., 1995. Implementation of motivational tactics in tutoring systems. International Journal of Intelligence in Education 6, 337–378.

Dickey, M.D., 2005. Engaging by design: how engagement strategies in popular computer and video games can inform instructional design. Educational Technology Research and Development 53, 67–93.

Dodds, P., Fletcher, J., 2004. Opportunities for new "smart" learning environments enabled by next-generation web capabilities. Journal of Educational Multimedia and Hypermedia 13 (4), 391–404.

Drummond, J., Litman, D., 2010. In the zone: towards detecting student zoning out using supervised machine learning. In: Aleven, V., Kay, J., Mostow, J. (Eds.), Intelligent Tutoring Systems, Part II, vol. 6095. Springer-Verlag, Berlin/Heidelberg, pp. 306–308.

Ekman, P., 1992. An argument for basic emotions. Cognition & Emotion 6 (3–4), 169–200.

Ericsson, K., Simon, H., 1993. Protocol Analysis: Verbal Reports as Data (Rev. ed.). The MIT Press, Cambridge, MA.

Fisher, C.D., 1993. Boredom at work—a neglected concept. Human Relations 46 (3), 395–417.

Fogelman, K., 1976. Bored 11-year-olds. British Journal of Social Work 6 (2), 201–211.

Forbes-Riley, K., Litman, D., 2011. When does disengagement correlate with learning in spoken dialog computer tutoring? In: Bull, S., Biswas, G. (Eds.), Proceedings of the 15th International Conference on Artificial Intelligence in Education. Springer, Berlin/Heidelberg, pp. 81–89.

Forbes-Riley, K., Rotaru, M., Litman, D., 2008. The relative impact of student affect on performance models in a spoken dialogue tutoring system. User Modeling and User-Adapted Interaction 18 (1–2), 11–43.

Gee, J.P., 2003. What video Games Have to Teach us About Learning and Literacy. Palgrave Macmillan, New York.

Graesser, A., Lu, S., Olde, B., Cooper-Pye, E., Whitten, S., 2005. Question asking and eye tracking during cognitive disequilibrium: comprehending illustrated texts on devices when the devices break down. Memory and Cognition 33, 1235–1247.

Graesser, A.C., Conley, M.W., Olney, A.M., Intelligent tutoring systems. In: Harris, K.R., Graham, S., Urdan, T. (Eds.), The APA Educational Psychology Handbook. American Psychological Association, Washington, DC, in press.

Gross, J., 2008. Emotion regulation. In: Lewis, M., Haviland-Jones, J., Barrett, L. (Eds.), Handbook of Emotions 3rd ed. Guilford, New York, NY, pp. 497–512.

Hegarty, M., Just, M., 1993. Constructing mental models of machines from text and diagrams. Journal of Memory and Language 32 (6), 717–742.

Hembree, R., 1988. Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research 58 (1), 47–77.

Hidi, S., Renninger, K.A., 2006. The four-phase model of interest development. Educational Psychologist 41 (2), 111–127.

Holsanova, J., Holmberg, N., Holmqvist, K., 2009. Reading information graphics: the role of spatial contiguity and dual attentional guidance. Applied Cognitive Psychology 23 (9), 1215–1226.

Hulleman, C.S., Durik, A.M., Schweigert, S.A., Harackiewicz, J.M., 2008. Task values, achievement goals, and interest: an integrative analysis. Journal of Educational Psychology 100 (2), 398–416.

Hyrskykari, A., 2006. Utilizing eye movements: overcoming inaccuracy while tracking the focus of attention during reading. Computers in Human Behavior 22 (4), 657–671.

Jacobs, A., Fransen, B., McCurry, J.M., Heckel, F., Wagner, A., Trafton, J.G., 2009. A preliminary system for recognizing boredom.

Paper presented at the Proceedings of the Fourth ACM/IEEE International Conference on Human Robot Interaction, La Jolla, California, USA.

Koedinger, K., Anderson, J., Hadley, W., Mark, M., 1997. Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education 8, 30–43.

Larson, R.W., Richards, M.H., 1991. Boredom in the middle school years—blaming schools versus blaming students. American Journal of Education 99 (4), 418–443.

Lehman, B., Matthews, M., D'Mello, S., Person, N., 2008. What are you feeling? Investigating student affective states during expert human tutoring sessions. In: Woolf, B., Aimeur, E., Nkambou, R., Lajoie, S. (Eds.), Proceedings of the Ninth International Conference on Intelligent Tutoring Systems. Springer, Berlin, Heidelberg, pp. 50–59.

Linnenbrink, E., Pintrich, P., 2002. The role of motivational beliefs in conceptual change. In: Limon, M., Mason, L. (Eds.), Reconsidering Conceptual Change: Issues in Theory and Practice. Kluwer Academic Publishers, Dordretch, Netherlands, pp. 115–135.

Mandler, G., 1984. Mind and Body: Psychology of Emotion and Stress. W.W. Norton & Company, New York.

Mann, S., Robinson, A., 2009. Boredom in the lecture theatre: an investigation into the contributors, moderators and outcomes of boredom amongst university students. British Educational Research Journal 35 (2), 243–258.

McGiboney, G.W., Carter, C., 1988. Boredom proneness and adolescents personalities. Psychological Reports 63 (3), 741–742.

McLaren, B.M., DeLeeuw, K.E., Mayer, R.E., 2011a. Polite web-based intelligent tutors: can they improve learning in classrooms? Computers & Education 56 (3), 574–584.

McLaren, B.M., DeLeeuw, K.E., Mayer, R.E., 2011b. A politeness effect in learning with web-based intelligent tutors. International Journal of Human-Computer Studies 69 (1–2), 70–79.

Mehan, H., 1979. Learning Lessons: Social Organization in the Classroom. Harvard University Press, Cambridge.

Moss, J., Schunn, C.D., VanLehn, K., Schneider, W., McNamara, D.S., Jarbo, K., 2008. They were trained, but they did not all learn: individual differences in uptake of learning strategy training. In: Love, B.C., McRae, K., Sloutsky, V.M. (Eds.), Proceedings of the 30th Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Austin, TX, pp. 1389–1395.

Mota, S., Picard, R., 2003. Automated posture analysis for detecting learner's interest level. Paper presented at the Computer Vision and Pattern Recognition Workshop.

Olney, A., 2010. Extraction of concept maps from textbooks for domain modeling. In: Kay, J., Aleven, V. (Eds.), Proceedings of 10th International Conference on Intelligent Tutoring Systems. Springer, Berlin/Heidelberg, pp. 390–392.

Olney, A., D'Mello, S.K., 2010. Interactive event: a DIY pressure sensitive chair for intelligent tutoring systems. In: Kay, J., Aleven, V. (Eds.), Proceedings of 10th International Conference on Intelligent Tutoring Systems. Springer, Berlin/Heidelberg, pp. 456.

Pekrun, R., 2006. The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review 18 (4), 315–341.

Pekrun, R., 2010. Academic emotions. In: Urdan, T. (Ed.), APA Educational Psychology Handbook, vol. 2. American Psychological Association, Washington, DC.

Pekrun, R., Elliot, A., Maier, M., 2006. Achievement goals and discrete achievement emotions: a theoretical model and prospective test. Journal of Educational Psychology 98 (3), 583–597.

Pekrun, R., Goetz, T., Daniels, L., Stupnisky, R.H., Raymond, P., 2010. Boredom in achievement settings: exploring control-value antecedents and performance outcomes of a neglected emotion. Journal of Educational Psychology 102 (3), 531–549.

Perkins, R.E., Hill, A.B., 1985. Cognitive and affective aspects of boredom. British Journal of Psychology 76 (May), 221–234.

Person, N., Kreuz, R., Zwaan, R., Graesser, A., 1995. Pragmatics and pedagogy—conversational rules and politeness strategies may inhibit effective tutoring. Cognition and Instruction 13 (2), 161–188.

Psotka, J., Massey, D., Mutter, S., 1988. Intelligent Tutoring Systems: Lessons Learned. Lawrence Erlbaum Associates.

Rayner, K., 1998. Eye movements in reading and information processing: 20 years of research. Psychological Bulletin 124 (3), 372–422.

Rayner, K., Fischer, M.H., 1996. Mindless reading revisited: eye movements during reading and scanning are different. Attention, Perception, & Psychophysics 58 (5), 734–747.

Reichle, E.D., Reineberg, A.E., Schooler, J.W., 2010. Eye movements during mindless reading. Psychological Science 21 (9), 1300.

Robinson, W.P., 1975. Boredom at school. British Journal of Educational Psychology 45, 141–152.

Robison, J., McQuiggan, S., Lester, J., 2009. Evaluating the consequences of affective feedback in intelligent tutoring systems. Paper presented at the International Conference on Affective Computing & Intelligent Interaction, Amsterdam.

Roda, C., Thomas, J., 2006. Attention aware systems: theories, applications, and research agenda. Computers in Human Behavior 22 (4), 557–587.

Russell, J.A., Weiss, A., Mendelsohn, G.A., 1989. Affect grid—a single-item scale of pleasure and arousal. Journal of Personality and Social Psychology 57 (3), 493–502.

Schmidt-Weigand, F., Kohnert, A., Glowalla, U., 2010. A closer look at split visual attention in system- and self-paced instruction in multimedia learning. Learning and Instruction 20 (2), 100–110.

Schutz, P., Pekrun, R. (Eds.), 2007. Emotion in Education. Academic Press, San Diego, CA.

Sleeman, D., Brown, J. (Eds.), 1982. Intelligent Tutoring Systems. Academic Press, New York.

Smilek, D., Carriere, J.S.A., Cheyne, J.A., 2010. Out of mind, out of sight: eye blinking as indicator and embodiment of mind wandering. Psychological Science 21 (6), 786–789.

Strain, A., D'Mello, S., 2011. Emotion regulation during learning. In: Bull, S., Biswas, G. (Eds.), Proceedings of the 15th International Conference on Artificial Intelligence in Education. Springer, New York/Heidelberg, pp. 566–568.

Thackray, R.I., 1981. The stress of boredom and monotony—a consideration of the evidence. Psychosomatic Medicine 43 (2), 165–176.

Toet, A., 2006. Gaze directed displays as an enabling technology for attention aware systems. Computers in Human Behavior 22 (4), 615–647.

van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., Paas, F., 2009. Attention guidance during example study via the model's eye movements. Computers in Human Behavior 25 (3), 785–791.

van Gog, T., Scheiter, K., 2010. Eye tracking as a tool to study and enhance multimedia learning. Learning and Instruction 20 (2), 95–99.

VanLehn, K., 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 46 (4), 197–221.

VanLehn, K., Graesser, A., Jackson, G., Jordan, P., Olney, A., Rose, C.P., 2007. When are tutorial dialogues more effective than reading? Cognitive Science 31 (1), 3–62.

Vygotsky, L., 1986. Thought and Language. MIT Press, Cambridge, MA.

Wang, H., Chignell, M., Ishizuka, M., 2006. Empathic tutoring software agents using real-time eye tracking. Paper presented at the Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, San Diego, California.

Wang, N., Johnson, W.L., Mayer, R.E., Rizzo, P., Shaw, E., Collins, H., 2008. The politeness effect: pedagogical agents and learning outcomes. International Journal of Human–Computer Studies 66 (2), 98–112.

Wasson, A.S., 1981. Susceptibility to boredom and deviant-behavior at school. Psychological Reports 48 (3), 901–902.

Wisher, R., Fletcher, J., 2004. The case for advanced distributed learning. Information & Security: An International Journal 14, 17–25.

Woolf, B., 2009. Building Intelligent Interactive Tutors. Morgan Kaufmann Publishers, Burlington, MA.

Woolf, B., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., 2010. The effect of motivational learning companions on low achieving students and students with disabilities. In: Kay, J., Aleven, V. (Eds.), Proceedings of 10th International Conference on Intelligent Tutoring Systems. Springer, Berlin/Heidelberg, pp. 327–337.

Zafer, 2010. TrackEye: Real-Time Tracking of Human Eyes Using a Webcam. Retrieved from 〈http://www.codeproject.com/KB/cpp/TrackEye.aspx〉.

Zeidner, M., 2007. Test anxiety in educational contexts: concepts, findings, and future directions. In: Schutz, P., Pekrun, R. (Eds.), Emotions in Education. Academic Press, San Diego, CA, pp. 165–184.

Zieliński, P., 2010. Opengazer: open-source gaze tracker for ordinary webcams. Retrieved from 〈http://www.inference.phy.cam.ac.uk/opengazer/〉.