Empirically Building and Evaluating a Probabilistic Model of User Affect

Cristina Conati, Heather Maclaren

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada V6T 1Z4
{conati, maclaren}@cs.ubc.ca

Abstract

We present the development of a probabilistic model of user affect designed to allow an intelligent agent to recognise multiple user emotions during the interaction with an educational computer game. Our model deals with the high level of uncertainty involved in recognizing a variety of user emotions by combining information on both the causes and effects of emotional reactions within a Dynamic Bayesian Network. In this paper we illustrate how we built our model in a series of stages of construction and direct evaluations. We started by designing the causal part of the model by relying on empirical data integrated with relevant psychological theories of emotion and personality. We then analyzed the model's limitations via empirical evaluations. Finally we used this analysis to guide the second part of the work, devoted to understanding if and how some of the student's emotional assessment could be more easily provided by the part of the model that diagnoses emotional states from their observable effects. Our results provide encouraging support for the combining of causal and diagnostic information to form a single assessment of the user's affective state.

Keywords: affective computing, biometrics, empirical data, dynamic Bayesian networks, evaluation, user modeling.

This paper (or a similar version) is not currently under review by another journal or conference, nor will it be submitted to such within the next three months.

1 Introduction

Recent years have seen a flourishing of research directed towards adding an affective component to human–computer dialogue. One key element of this endeavour is the computer's capability to recognize the user's emotional states during the interaction, which requires a model of the user's affect. Humans use different sources of information to assess a person's emotions, including causal information on both context and the person's relevant traits, and symptomatic information on the person's visible bodily reactions. However, this information is often incomplete and even contradictory, making assessment of emotion a task riddled with uncertainty.

To handle this uncertainty, Conati (2002) proposed a probabilistic framework for affective user modeling that integrates, in a Dynamic Decision Network (DDN) (Dean & Kanazawa, 1989), information on both the possible causes of the user's affective reaction and its observable effects. Leveraging any information available on the user's emotional state is crucial, because the different sources of evidence are often ambiguous, and their reliability varies significantly according to both the user and each particular interaction.

In this paper, we illustrate how we used the framework proposed by (Conati, 2002) to build a model of user affect during interaction with an educational computer game. The long–term goal is to employ this user model to guide adaptive system interventions aimed at improving the overall success of the student's educational experience with the game.

Although there is still no hard evidence that taking user affect into account can substantially improve human–computer interaction in general, there are several studies indicating that maintaining positive student affect is beneficial in educational settings. Craig et al. (2004) reported that flow and confusion were positively correlated with learning, whereas boredom was negatively correlated. Linnenbrink and Pintrich (2002) found that while most students experience some confusion when confronted with information that does not fit their current knowledge, those in a generally positive affective state will adapt their known concepts to assimilate it, whereas students in a generally negative affective state will reject the new knowledge. Cordova and Lepper (1996) found that learners exposed to motivationally embellished educational software (Lepper et al., 1993) had higher levels of intrinsic motivation. As a result, they become more deeply engaged by the interaction, and learned more in a fixed period of time.

We believe that the benefits of taking user affect into account are even stronger for educational activities that rely heavily on the student's direct involvement in the learning process, such as those supplied by exploratory learning environments and educational games. An educational game tries to increase the learner's motivation by embedding pedagogical activities in highly engaging, game-like interactions. Several studies have shown that, while educational games are usually successful in increasing student engagement, they often fail to trigger learning , (e.g., Klawe, 1998). To overcome this limitation, we are designing emotionally intelligent pedagogical agents that, as part of game playing, generate tailored interventions aimed at stimulating the student to learn better from the game (Conati & Klawe, 2002). However, in order not to interfere with the high level of engagement that is a key asset of educational games, we argue that these agents need to take into account the players' affective states in addition to their cognitive states when deciding how to act. The affective model we describe in this paper is meant to be used by our pedagogical agents, together with a model of student learning, to generate interventions that improve learning without compromising engagement.

In this paper, we illustrate how we built and evaluated our affective model in two stages. In the first stage we built the part of the model that reasons from causes to emotions (*predictive model* from now on) by relying on empirical data integrated with relevant psychological theories of emotion and personality. We then empirically evaluated the performance of this part of the model, and used this analysis to guide the second part of the work. The second stage was devoted to understanding whether and how some of the student's emotional assessment could be more easily provided by the part of the model that diagnoses emotional states from their observable

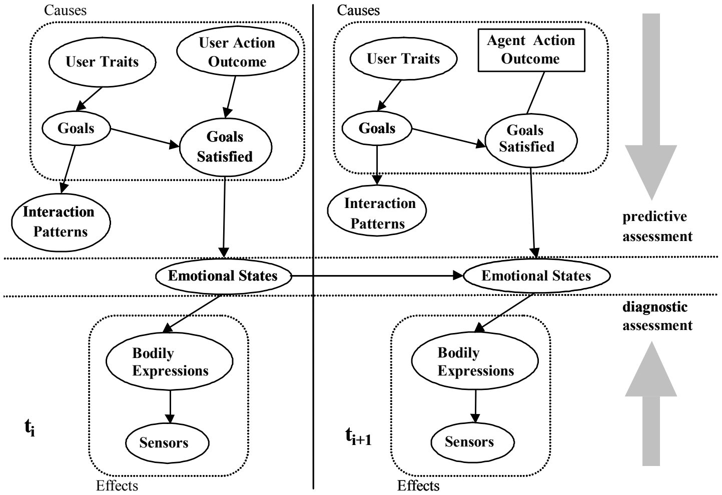


Figure 1. Two time-slices of the DDN for affective modeling

effects (*diagnostic model* from now on). In particular, we ran user studies to explore whether this diagnostic information can be provided by physiological sensors. These sensors have been extensively investigated for emotion detection in laboratory settings and controlled environments (e.g., Lang et al., 1993; Vyzas & Picard, 1998) There has also been some research on using sensors for affect detection in more realistic settings, but this research has focused mostly on either detecting one specific emotion (Healey & Picard, 2005; Kapoor & Picard, 2005), lower–level affective measures such as valence and arousal (e.g., Prendinger et al., 2005), or overall emotional predisposition over a complete interaction (Mandryk et al., 2006). Here we extend this research to the instantaneous detection of multiple, rapidly changing emotions that possibly overlap and conflict, as often experienced by students playing an educational game.

The structure of this paper is as follows. In Section 2, we describe the general framework we used to build our probabilistic model of user affect. In Section 3, we introduce Prime Climb, the educational computer game we used as a test–bed application for model development. Section 4 illustrates the predictive part of the affective model. In Section 5, we introduce our technique for model evaluation and apply it to test the predictive model. In Section 6, we present our investigation into using physiological sensors to provide information for the diagnostic part of the model. In Section 7, we discuss related work, and in Section 8 we conclude with a discussion of the research presented here as well as ideas for future work.

2 A Dynamic Decision Network for Emotion Recognition

A DDN is a graph where nodes represent either stochastic variables of interest or points where an agent needs to make deliberate decisions. Arcs in the graph capture the direct probabilistic relationships between the nodes, including temporal dependencies between the evolving values of dynamic variables. Each node has an associated probability distribution representing the conditional probability of each of its possible values, given the values of its parent nodes. As evidence on one or more network variables becomes available, *ad hoc* algorithms update the posterior probabilities of all the other variables, given the observed values.

Figure 1 shows a high-level representation of two time-slices in the DDN-based framework for affective modeling proposed in (Conati, 2002). Each time slice represents the model's variables at a particular point in time. For illustration purposes, the nodes in Figure 1 represent classes of variables instead of individual variables in the DDN. As the figure shows, the network can

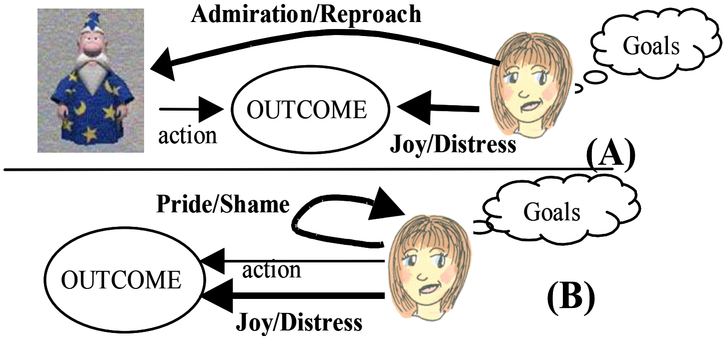


Figure 2. Example emotions in the OCC theory

combine evidence on both the causes and effects of emotional reactions, to compensate for the fact that often evidence on causes or effects alone is insufficient to accurately assess the user's emotional state.

The subnetwork above the nodes *Emotional States* is the predictive component of the framework. It represents the relations between possible causes and emotional states as described in the OCC cognitive theory of emotions (Ortony, Clore, & Collins, 1988). According to this theory, emotions derive from cognitive appraisal of the current situation, which consists of events, agents, and objects. The outcome of the appraisal depends on how the situation fits with one's goals and preferences. For instance, depending on whether the current event (e.g., the outcome of an action in Figure 2) does or does not fit with one's goals, that person will feel either *joy* or *distress* toward the event (see Figure 2, A and B). Correspondingly, if the current event is caused by a third-party agent, that person will feel *admiration* or *reproach* toward the agent (see Figure 2A); if that agent is oneself, the person will feel either *pride* or *shame* (see Figure 2B). Based on this structure, the OCC theory defines 22 different emotions, described in terms of their valence and the entity they relate to.

We adopted this particular theory of emotion for our affective modeling framework because its clear and intuitive representation of the causal nature of emotional reactions lends itself well to a computational representation. Furthermore, the fact that the OCC model includes the target of an emotion in its definition provides more fine-grained information to direct the actions of an interactive intelligent agent, compared to alternative models that define emotions in terms of their level of valence and arousal. For instance, if an interface agent can recognize that the user feels a negative emotion toward herself (*shame* by OCC definition) it can decide to provide hints aimed at making the user feel better about her performance. If the agent recognises that the negative feelings are directed toward itself (*reproach* by OCC definition) it may decide to take actions that allow it to make amends with the user.

To apply the OCC theory to emotion recognition during human-computer interaction, our DDN includes variables for goals that a user may have during the interaction with a system that includes an intelligent agent, (nodes *Goals*[1] in Figure 1). The events subject to the user's appraisal are any visible interface outcomes generated by the user's or the agent's action (nodes *User Action Outcome* and *Agent Action Outcome* in Figure 1)[2]. Agent action outcomes are

[1] We currently represent players preferences in terms of goals, as suggested in (Gratch, 2000).

[2] We explicitly model action outcomes rather than action themselves because one individual action may generate several effects at once, each of which may be appraised in relation to a different goal (we provide examples of this scenario later in the paper). We don't need to include action nodes in addition to action outcome because we assume that the visible effects of an action are deterministic. Thus, the occurrence of an action is implicitly represented by the description of its outcomes.

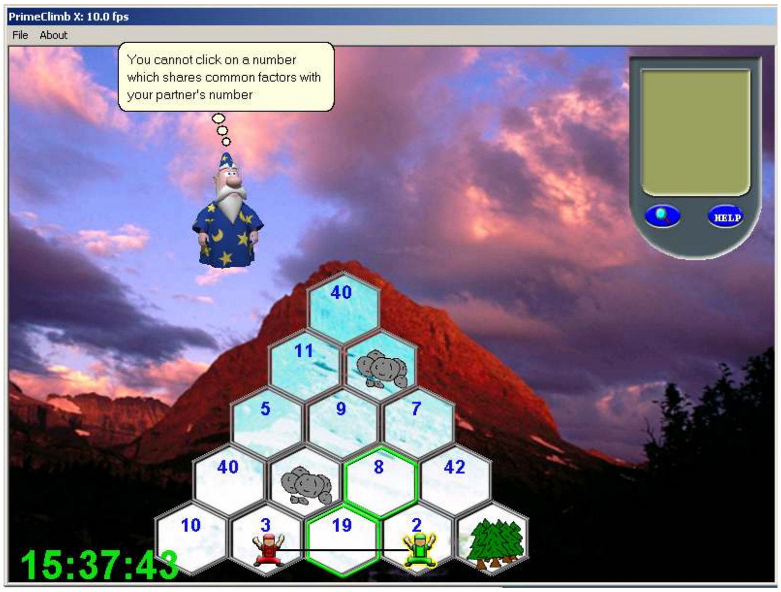
represented as decision variables in the framework, indicating points where the agent decides how to intervene in the interaction. The desirability of an event in relation to the user's goals is represented by the node class *Goals Satisfied,* which in turn influences the user's *Emotional States*.

The user's goals are a key element of the OCC model, but assessing these goals is not trivial, especially when eliciting them with queries to the user during the interaction would be too intrusive, as is the case during game playing. Thus, our DDN also includes nodes to infer user goals from indirect evidence. User goals can depend on *User Traits* such as personality (Costa & McCrae, 1992). Also, user goals can influence user *Interaction Patterns*, which in turn can be inferred by observing the outcomes of individual user actions. Thus, observations of both the relevant user traits and action outcomes can provide the DDN with indirect evidence for assessing user goals.

The sub-network below the nodes *Emotional States* is the diagnostic part of the affective modeling framework, representing the interaction between emotional states and their observable effects. *Emotional States* directly influence user *Bodily Expressions*, which in turn affect the output of *Sensors* that can detect them. Because in many situations a single sensor cannot reliably identify a specific emotional state, our framework is designed to modularly combine any available sensor information, and gracefully degrade in the presence of partial or noisy information.

In Figure 1, the links between emotion nodes in different time-slices indicate how the corresponding variables evolve over time. These links model, for example, the fact that a user is more likely to feel a given emotion at time t_{i+1} if the user felt it at time t_i. A new time-slice is added to the network whenever either the user or the agent performs an action (this happens, for instance, between every 3 and 10 seconds in the framework application we describe in the following sections); the new slice represents the state of the world just after the corresponding action occurred. In a DDN, only the time-slices that directly influence the current state need to be maintained. We currently assume that maintaining two time-slices is sufficient to capture the relevant temporal dependencies in our framework. Since at the moment the only temporal variables in the framework are the emotion variables, this assumption implies that the user's emotions at any given time depend only on the last game action and his or her emotional state in the previous slice, while the effects of earlier actions on the current emotional state are channelled through the emotional state in the previous time-slice. This assumption would be invalid in situations where a sequence of actions directly causes a particular emotional reaction, rather than influencing it via a chain of subsequent emotional states. Our framework also assumes that a user's high–level goals do not change over time, as indicated by the lack of a link between the *Goals* node at time t_i and the *Goals* node at t_{i+1} in Figure 1. Both assumptions derive from our philosophy for tackling the complexity of modeling affect: start with reasonably simplified models and increment them as limitations are uncovered by empirical evaluations.

Having described the general framework underlying our model of user affect, we will now present the educational game we used to apply and test the framework.



PrimeClimb X: 10.0 fps

You cannot click on a number which shares common factors with your partner's number

Figure 3. The Prime Climb Interface

3 The Prime Climb Educational Game

As a test-bed for the general affective modeling framework described in the previous section, we used Prime Climb, an educational game designed by the EGEMS group at the University of British Columbia to help 6th and 7th grade students practise number factorization. Figure 3 shows a screenshot of Prime Climb. Two players must cooperate to climb a series of mountains that are divided in numbered sectors. Each player should move to a number that does not share any common factors with her partner's number, otherwise she falls. Prime Climb provides two tools to help students: (i) a *magnifying glass* that the student can use to see a number's factorization (accessible by clicking on the magnifying glass icon at the bottom-left corner of the hand-held device in Figure 3); (ii) a *help box* (accessible by clicking on the *help* icon at the bottom-right corner of the hand-held device) that allows the student to ask for advice, which is provided by the pedagogical agent we are building for the game.

The pedagogical agent is an autonomous agent that provides individualized support, both on demand and unsolicited, when the student does not seem to be learning from the game (Conati & Zhao, 2004). To decide when to intervene and what hints to provide, the agent relies on a probabilistic model of the player's factorization knowledge which is continuously updated during the player's interaction with the game. When the probabilities in the model of student learning indicate that the player is missing key pieces of knowledge to learn from her current move, the pedagogical agent provides hints designed to stimulate the student to reason about the relevant domain knowledge. This can happen even after a student's correct move, if the underlying student model predicts that the successful move was based on luck rather than knowledge. When the player falls, the agent provides hints in three incremental levels of detail. At the most general level, the agent's hints include reminders to think about number factorization or to think about common factors when climbing. At a second level, the agent suggests that the player uses the magnifying glass to see a number's factorization. The hints in the last level include examples of how to factorize numbers or how to determine whether two numbers have a common factor. When the player makes a successful move, the agent attempts to stimulate reasoning about the domain knowledge by asking the player if she knows why the move she has just made was correct. The agent also occasionally attempts to encourage the student by congratulating her when she is successful.

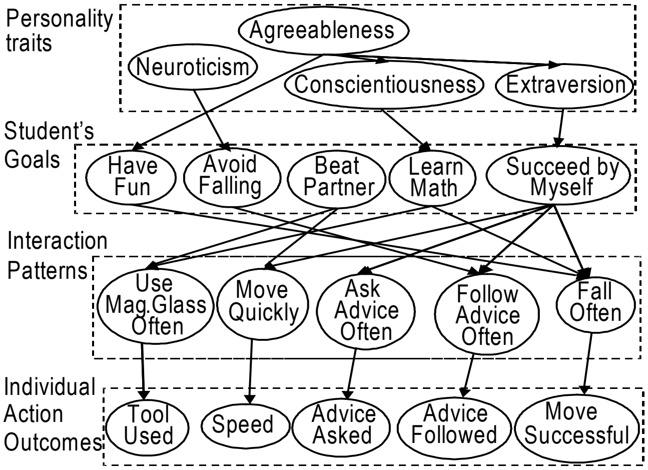


Figure 4. Sub-network for Goal Assessment

To avoid interfering with the student's level of engagement while playing the game, we used the framework described in the previous section to build an affective user model for Prime Climb that the agent can use to decide when and how to intervene. This user model produces a real-time assessment of the player's emotions during interaction with Prime Climb, and will eventually be integrated with the model of student learning to inform the agent's pedagogical decisions.

4 Building the Predictive Component of the Prime Climb Affective Model

In this section, we describe how we instantiated the predictive component of the affective modeling framework in Figure 1 to model the affective states of a Prime Climb player. We present two sub-network structures within the predictive component; the first assesses the student's goals (*goal assessment sub-network*); the second models the student's appraisal of game events in relation to those goals, to produce an assessment of the student's current affective state (*appraisal sub-network*).

4.1 Instantiation of the Goal Assessment Sub-network

Figure 4 shows the structure of the sub-network that assesses student goals. Because all of the variables in this sub-network are observable either during or after the interaction with Prime Climb, we identified relevant individual variables and built the corresponding conditional probability tables (CPTs) using data collected through a series of Wizard of Oz studies where pairs of students interacted with the game while an experimenter controlled the pedagogical agent. Here we give a high-level description of this process. For more details see (Zhou & Conati, 2003).

Information to instantiate variables representing student goals was collected via a post-game questionnaire in which students could express the goals they had while playing the game. We identified five high-level goals in our user studies, represented in the model by the following binary variables (with Boolean values indicating the probability of having/not having a given goal): *Have Fun*, *Avoid Falling*, *Beat Partner*, *Learn Math*, and *Succeed By Myself*[3]. We also

[3] The goal *Beat Partner* is inconsistent with the collaborative nature of the game, but it is not surprising given findings indicating that certain personality types tend to be competitive even during collaborative interactions.

found that students can have more than one of these goals at the same time. For this reason, the affective model represents each goal through a dedicated node rather than as one of the five mutually exclusive values on a single variable.

Because personality is known to influence one's goals and behaviours (Costa & McCrae, 1992), our model contains nodes and links representing student personality types and their relation to student goals in playing Prime Climb. We used the personality types suggested by the Five-Factor Model (Costa & McCrae, 1992), in which personality traits are structured as five domains – *neuroticism*, *extraversion*, *openness*, *agreeableness* and *conscientiousness*. Data to instantiate the prior and conditional probabilities involving these variables was collected through a personality test specifically designed for children (Graziano et al., 1997). As Figure 4 shows, our affective model currently includes variables for only four of the five domains, because our study data showed that *openness* was not directly relevant to our task. All of the personality variables are binary, with Boolean values representing the probability of belonging or not belonging to a given personality domain.

During the studies, we also collected log files of the interactions, to mine the possible relationships between student goals (assessed via the goal post-questionnaire) and interaction behaviours. Our data indicated several dependencies between student goals and playing behaviour. The interaction patterns we identified to be relevant for inferring student goals included: (1) a tendency to make moves quickly or slowly (represented by the node *Move Quickly*); (2) a tendency to use the magnifying glass often or not (node *Use Mag. Glass Often*); (3) a tendency to ask the agent for advice often or not (node *Ask Advice Often*); (4) a tendency to follow the agent's advice often or not (node *Follow Advice Often*); (5) a tendency to fall often (node *Fall Often*). All the Interaction Pattern nodes are binary, with Boolean values indicating the presence/absence of a given pattern.

The probabilistic dependencies among goals, personalities, interaction patterns and individual student actions were established through correlation analysis between the personality test results, the goal questionnaire results and student actions logged during the interactions (Zhou & Conati, 2003). Figure 4 shows the resulting sub-network, incorporating both positive and negative correlations. The bottom level specifies how interaction patterns are recognized from the relative frequency of individual action outcomes (Zhou & Conati, 2003).

We originally intended to represent different degrees of personality type and goal priority by using multiple values in the corresponding nodes. However, we did not have enough data to populate the larger CPTs that this would generate, thus all the nodes in the goal assessment sub-network are binary. This simplification has not proven to be particularly detrimental to the performance of the goal assessment sub-network, as we will see in the model evaluation section to come. However, the resulting assumption that all goals have the same priority when present, together with the assumption that goals do not change over time, do affect the accuracy of the model's assessment of student emotions, as we will also discuss in the evaluation section.

4.2 Instantiating the Appraisal Subnetwork

Figure 5 and Figure 6 show the details of the two types of time-slices used in the part of the network representing the *appraisal mechanism* (i.e., how the mapping between student goals and game states influences student emotions). Figure 5 shows the appraisal time-slice that is added to the affective model whenever the student performs an action. Figure 6 shows the time-slice added to the affective model whenever the pedagogical agent intervenes. Note that, for clarity purposes, Figure 5 and Figure 6 do not include the personality and interaction nodes used for goal assessment. The reader can refer to Figure 1 for an integrated picture of the goal assessment and appraisal sub-networks. For both types of appraisal time-slices, we specified an initial network structure based on the general OCC appraisal mechanism and our intuition, and then refined the structure by using empirical data collected from user studies designed for this task (described in

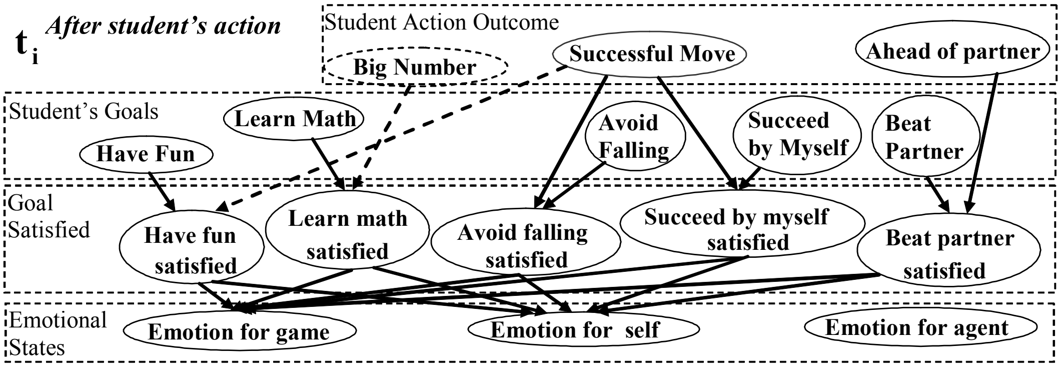


Figure 5. Sub-network time-slice for appraisal of student actions

Section 5.2). In this section, we first describe the structure of the initial network (corresponding in both figures to the solid-line nodes and links). We then describe the parts of the sub-network that were refined using empirical data (dashed-line components in both figures).

4.2.1 Initial Structure

The appraisal sub-network currently represents only 6 of the 22 emotions defined in the OCC model. They are *joy*/*distress* for the current state of the game, *pride*/*shame* of the student toward herself, and *admiration*/*reproach* toward the agent. These six particular emotions were chosen because we observed them often during pilot studies with Prime Climb, thus they seemed highly relevant for directing the actions of the Prime Climb pedagogical agent. While other emotions in the OCC model may be relevant, for instance emotions toward one's partner during game play, we decided to start with a relatively simple model and progress to more complex ones only after having ascertained the viability of our approach.

Each of the three emotion pairs included in the model is represented by a binary node — *emotion–for-game*, *emotion-for-self* and *emotion-for-agent,* respectively (see nodes in the *Emotional States* level in Figure 5 and Figure 6) — where binary values represent the probability that the student is feeling one of the two emotions in the corresponding pair. This structure was chosen because, while the two emotions in each pair are mutually exclusive and are thus best represented by a binary node, students may simultaneously feel emotions in the different pairs, requiring a separate node for each pair.

Following the OCC appraisal model, a student's emotional state depends on whether her goals are satisfied or not during game playing. In the appraisal network, goal satisfaction is explicitly represented by a *Goal Satisfied* node for each goal in the goal assessment network (see nodes in

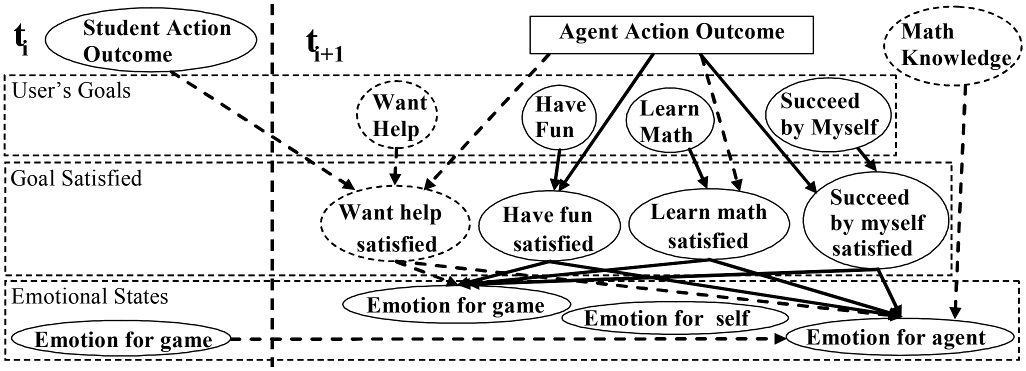


Figure 6. Sub-network time-slice for appraisal of agent interventions

the *Goal Satisfied* level in Figure 5 and Figure 6). The links between *Goal Satisfied* nodes and the emotion nodes are defined as follows. We assume that the outcome of every relevant agent or student action is subject to student appraisal. Thus, each *Goal Satisfied* node influences *emotion-for-game* in every slice. Whether a *Goal Satisfied* node influences *emotion-for-self* or *emotion-for-agent* in a given slice depends upon whether the slice was generated, respectively, by a student action (Figure 5) or an agent's action (Figure 6). Each *Goal Satisfied* node has three possible values: true, false, and neutral. The CPTs for emotion nodes are defined so that the probability of each positive emotion is proportional to the number of true *Goal Satisfied* nodes.

The probability of each *Goal Satisfied* node depends on whether the outcome of the current student or agent action matches the corresponding student's goal. In the initial appraisal sub-network, the links and CPTs between *Goal* nodes, the outcome of student or agent actions, and *Goal Satisfied* nodes were based on our intuition, and defined connections that are quite obvious. Let's start by looking at these links in the time slice for the appraisal of student action outcomes (Figure 5). Initially, this time slice included only student moves on the Prime Climb mountains as actions that trigger the appraisal mechanism, because the other two possible student game actions (using the magnifying glass and asking for help), were not seen very often during our studies and thus we did not have a clear sense of how they might influence student affect. The solid binary nodes in the *Student Action Outcome* level at the top of Figure 5 represent two different aspects of the outcome of a student move that we had observed to trigger student emotional reactions during game playing. The node *Successful Move* indicates whether the student's move was successful or not. The node *Ahead of Partner* indicates whether or not the move brought the student to be ahead of her partner on the mountain. We encoded some intuitive dependencies between these action outcomes and student goals in the initial appraisal network. For instance, if the student has the goal *Avoid Falling*, a successful move satisfies it, while a fall does not. If the student has the goal *Beat Partner*, only a move that brings the player ahead of the partner on the mountain contributes to satisfying this goal.

In the initial version of the time slice that models the appraisal of agent actions, the decision node *Agent Action Outcome*[4] was a four-valued node that represented the types of intervention that the agent could produce (see Section 3). These interventions were represented by the following decision values: (1) generate a hint at the first or second level of detail (e.g., a reminder to think about common factors when climbing or a suggestion to use the magnifying glass); (2) generate a hint with example following a fall; (3) generate a hint following a successful climb; (4) generate an encouragement. We used the same node value for the first two levels of hints following a fall because we observed in previous studies that the students tended to express similar reactions to these hints, thus we hypothesised that the students were appraising the hints the same manner. Reflecting this similarity in the decision node enabled us to reduce the complexity of that part of the network structure.

We encoded the following intuitive dependencies between agent actions and the satisfaction of student goals in this time slice. If the student has the goal *Have Fun*, providing encouragement will satisfy this goal, whereas providing any of the other more pedagogically oriented hints will not. If the student has the goal *Succeed By Myself*, providing any of the pedagogically oriented hints will not satisfy this goal (when the agent provides encouragement, goal satisfaction is neutral).

For the intuitive links described above, the conditional probabilities of the *Goal Satisfied* nodes were set by assigning: (1) a high probability that goal satisfaction is true when the student has the goal and an event that satisfies it occurs; (2) a high probability that goal satisfaction is false when the student has the goal and the opposite of a satisfying event occurs; and (3) a high

[4] We keep the label "Agent Action Outcome" for consistency with the slice for Student Action Outcome. In practice however, agent actions and their outcomes coincide because, unlike student actions, every agent action has a single outcome.

probability of goal satisfaction being neutral when the student does not have the goal.

However, we had no good intuition of how various student game actions would be appraised in relation to the goals *Have Fun* and *Learn Math*. We also had no good intuition of how agent actions would be appraised for the goal *Learn Math,* since the appraisal should reflect the student's perception of whether he/she learned math rather than whether this was what actually happened. Thus, we decided to base these appraisals of student and agent game actions on empirical data.

4.2.2 Data-based Refinement of the Appraisal Sub-network

In both Figure 5 and Figure 6, we have used dashed lines to indicate the parts of the network structure that we defined using empirical data. This data was collected through a user study in which students could explicitly express what made them have fun and learn math during their interaction with Prime Climb. This was done via two post-game questionnaires, one for each of *Learn Math* and *Have Fun*, that contained a list of statements of the type '*I learnt math/had fun when <game event>'*. Students rated each statement using a 5-point Likert scale (1=strongly disagree, 5=strongly agree). The game events considered in the questionnaires included the following:

For *Have Fun*

- *Student–generated events*: a successful move; a fall (unsuccessful move); using the magnifying glass; using the help box; reaching the top of the mountain.

For *Learn Math*

- *Student–generated events*: same as above, plus following the agent's advice, and encountering big numbers.
- *Agent–generated events*: suggestion to use the magnifying glass; reminder to think about common factors when climbing.

When considering which events to include in the questionnaires, we aimed to investigate as many game events as possible that we thought could influence the satisfaction of student goals, while keeping the questionnaire at a reasonable length (see more on this in Section 5.2). For instance, in *Agent-generated events* we included separate events for hints from the first and second levels of hint detail (as described in Section 3) rather than a single event (as specified in the *Agent Action Outcome* node described in Section 4.2.1), since data on these two levels of hint would enable us to test our initial hypothesis that these agent actions were appraised in a similar manner by the student. However, we did not include hints that provided examples of number factorization or computation of common factors, or hints generated after a successful move, because the agent version used in the study rarely generated these hints and thus most students would not be able to rate them. In addition, we restricted the number of hints included in the questionnaire to one hint per level of detail, and therefore did not include the reminder to think about number factorization.

While some of the *Student–generated events* were included in the questionnaires because they pertained to the three basic Prime Climb actions (a successful/unsuccessful move, using the magnifying glass, using the help box), other events such as *encountering big numbers* and *reaching the top of the mountain* were included based on anecdotal evidence from experimenters who had run previous user studies; we had not included them in the initial networks because the anecdotal evidence was insufficient to insert them in a meaningful way. To limit the length of the questionnaire, we did not ask students about events that already satisfied other goals within the model (e.g., the student being ahead of her partner, encouragement by the agent).

We will now describe how we generated all of the refinements to the appraisal time-slices beginning with the time-slice used to appraise the outcome of the student's action.

4.2.2.1 Appraisal of Student Action Outcomes

The students' answers to the questionnaires indicated that all of the tested student-generated events were relevant to some degree. In order to determine which events made a difference to the model's assessment, we scored all possible network structures derived from including these events, using their log marginal likelihood (Heckerman, 1999). The mutually exclusive events *successful move/fall* were represented via the binary values of the node *Successful Move* in Figure 5, while the events *using the magnifying glass*, *using the help box*, *reaching the top of the mountain* and *moving to a big number* were each represented by a new binary node in the time-slice for appraising student actions.

We found that the structure with the best score was the one encoding the following appraisal relations, in addition to those represented in the original network: (1) whether the student's move was successful or not influenced satisfaction of the goal *Have Fun*; and (2) whether the student encountered a big number influenced satisfaction of the goal *Learn Math*. The dashed components in Figure 5 show how these relations are encoded in the appraisal time-slice. The new binary node *Big Number* is linked to satisfaction of the goal *Learn Math* while the existing node *Successful Move* is linked to satisfaction of the goal *Have Fun*. We used frequencies from the questionnaire answers to set the CPTs for these new links. We based our definition of a big number on the large numbers frequently incorrectly factorized in students' pre-tests in our studies.

4.2.2.2 Appraisal of Agent Actions Outcomes

The data-based refinement of the time-slice added after an agent action consisted of two stages:

Stage 1. First, we used the students' questionnaire items related to the influence of the agent's actions on the goal *Learn Math* to test our hypothesis that hints from the first two levels of hint detail following a fall should be represented by a single value within the node *Agent Action Outcome.* To do this we created two candidate network structures, both of which represented the student's appraisal of the agent's action with regards to the goal *Learn Math*. These structures were identical except for the number of possible values in the node *Agent Action Outcome*. The first structure contained an *Agent Action Outcome* node with the four values we had initially specified using subjective heuristics. The second structure contained an *Agent Action Outcome* node where the unique value representing two levels of hint had been replaced by two separate values: *generate a hint at the first level* and *generate a hint at the second level.* We used the students' questionnaire answers to produce a log marginal likelihood score for each network structure. We found that the structure containing the *Agent Action Outcome* node with our original set of four values received the highest score and thus we retained it in the refined model.

We then refined the model by creating a link between *Agent Action Outcome* and satisfaction of the goal *Learn Math*. For the pedagogical agent actions that had been included in the questionnaire, we used the frequencies from the questionnaire answers to generate the corresponding CPT values. For the pedagogical agent actions that had not been included in the questionnaire due to their rarity (e.g. hints with examples following a fall, a hint following a successful climb) we set the CPT values for goal satisfaction to equal probability for true and false. For encouragement by the agent, a non-pedagogical action, we set satisfaction to neutral.

However, a preliminary evaluation of these changes showed that the model was underestimating students' admiration toward the agent, suggesting that the model still contained sources of inaccuracy related to appraisal of agent actions. We therefore moved to a second stage of data analysis to determine these problems.

Stage 2. During the user study, we also collected on-line self-reports on the students' feelings towards the game and towards the agent via a dialog box that would periodically pop-up during game playing (we will describe this mechanism in detail in Section 5.1), and students were asked to indicate the goals that they had during game-playing by filling in a post-game questionnaire. To better understand the relations between agent behaviours and student emotions, we analyzed the study log files to identify particular situations in the game in which students tended to report experiencing *Admiration* or *Reproach* toward the agent. Our data confirmed that encouragement by the agent generates students' admiration (45% *Admiration* reports against 7% reports of *Reproach* and 48% neutral reports), although we cannot tell whether this happens through the satisfaction of the goal *Have Fun* as we have encoded in initial time-slice for appraisal of agent actions. It also showed that students who are generally successful usually report either *Admiration* or *Neutral* feelings towards the agent, regardless of their goals (53% *Admiration* reports against 8% reports of *Reproach* and 39% neutral reports). This finding suggests that the students' positive feelings toward the game will positively influence their attitude towards the agent. We translated this finding into the model by adding a link from the student's emotion towards the game in the previous time-slice to the student's emotion towards the agent (as shown by the dashed line at the bottom of Figure 6).

Finally, we looked at situations in which students fell repeatedly and either received help or did not. Analysis of these situations revealed that approximately one–half of the students who reported *Admiration* when the agent intervened after they fell had declared the goal *Succeed By Myself*. Also, about one-half of the students who reported *Reproach* when the agent did not intervene had declared that goal. This result seems to indicate that, although some of the students may want to succeed by themselves in general, they may also want help in especially critical situations (e.g., when they fall repeatedly). That is, in these situations some students may reduce the priority of wanting to succeed by themselves in favour of wanting help. The data also revealed students who had not declared the goal *Succeed By Myself*, but when they began to fall they demonstrated annoyance when the agent intervened. That is, in these situations, the students demonstrated that they preferred to succeed by themselves rather than wanting help. These observations invalidate two of the choices previously made in the model implementation: (1) to ignore goal priority; and (2) to assume that goals are static during the interaction. Because we currently don't have enough data to model goal evolution in a principled way, we only addressed the implementation of multiple priority levels to model the relation between *Succeed By Myself* and wanting help. We changed the model as follows.

First, we added an additional goal, *Want Help*. Note that we did not represent this goal as one of the two values of the node *Succeed by Myself* because, as we discussed above, these goals are not necessarily mutually exclusive. For some students, they seem to represent a general vs. local attitude toward receiving help during game playing, and thus they may co-exist, although with different, possibly shifting priorities. The satisfaction of *Want Help* is dependent on two factors: the outcome of the student's move (i.e., a successful climb or a fall) and the agent's action. When the student falls, *Want Help* can only be satisfied if the agent provides help. If the student does not fall, then satisfaction is neutral.

Second, we tried to determine which traits influenced the students' attitudes towards receiving help during repeated falls. The only factor that seemed to play a role was students' math knowledge, a factor that we measured using pre-tests on factorization as part of our standard study design**Error! Reference source not found.**. Table 1 shows a confusion matrix comparing the students' math knowledge and whether they demonstrated that they wanted help when falling

Table 1. Confusion matrix comparing students' math knowledge with whether they wanted help.

		Math Knowledge	
		High	Low
Want Help	Yes	13	4
	No	4	9

repeatedly. We classified the students' math knowledge as 'high' if they correctly answered 50% or more of the questions on the factorization pre-test, otherwise the math knowledge was classified as 'low'. As the matrix shows, high math knowledge is associated with wanting help, whereas low math knowledge is associated with not wanting help. A Fisher test (Fisher, 1935) between the students' pre-test scores and whether they demonstrated that they wanted help after repeated falls showed a significant relationship (Fisher score = 0.025). Although this relationship seems backward, the results agree with the findings of Baker et al. (2004) that students with lower pre-test scores are more likely to want to succeed via trial and error than think about domain knowledge, whereas students with higher pre-test scores are more likely to want to learn from the resources available in the system, including provision of help. Given the above findings, a new node, representing prior math knowledge, was used to influence the priorities a student gives to the goals *Succeed By Myself* and *Want Help.*

We added a link from the new node, *Math Knowledge*, to *emotion-for-agent* (as shown in Figure 6). As we mentioned earlier in Section 4.2, the CPT for *emotion-for-agent* was defined so that the probability of the student feeling *Admiration* was proportional to the number of true *Goal Satisfied* nodes. We refined the CPT in *emotion-for-agent* so that, if the student had high math knowledge, then the influence of the node *Succeed by Myself Satisfied* on the probability of *Admiration* was lower than the influence of the other *Goal Satisfied* nodes. If the student had low math knowledge, then the influence of the node *Want Help Satisfied* on the probability of *Admiration* was lower instead.

Our third and final change to the model was to refine the decision node representing the available agent's actions so that it included the agent choosing not to intervene. All *Goal Satisfied* nodes other than *Succeed By Myself* and *Want Help* were given a neutral satisfaction for this new action. *Want Help* was discussed earlier; *Succeed By Myself* was given a small probability of satisfaction to reflect possible mild positive feelings towards the agent for not interrupting in general, rather than at specific events.

5 Evaluation of the Affective Model

Having completed the construction of the casual part of our affective model, our next step is to assess its predictive accuracy. As for other user models, affective models can be evaluated either *directly* by specifically measuring the accuracy of the model's predictions, or *indirectly* by testing the performance of an application that uses the model to adapt its behaviour.

One of the main shortcomings of indirect evaluations is that they require the user model to be embedded within a complete system, and thus usually occur at later stages of a research project. Because affective human-computer interaction is a very young field, there is little research that is mature for this type of comprehensive testing. In fact, we are aware of only two indirect empirical evaluations of affective user models, one by (Guinn & Hubal, 2003) and one by (Prendinger et al., 2005). Both of these works suffer from the second potential limitation of indirect evaluations: insights into the performance of the affective user model to be evaluated are confounded by other aspects of the system, unless one sets up carefully designed ablation studies.

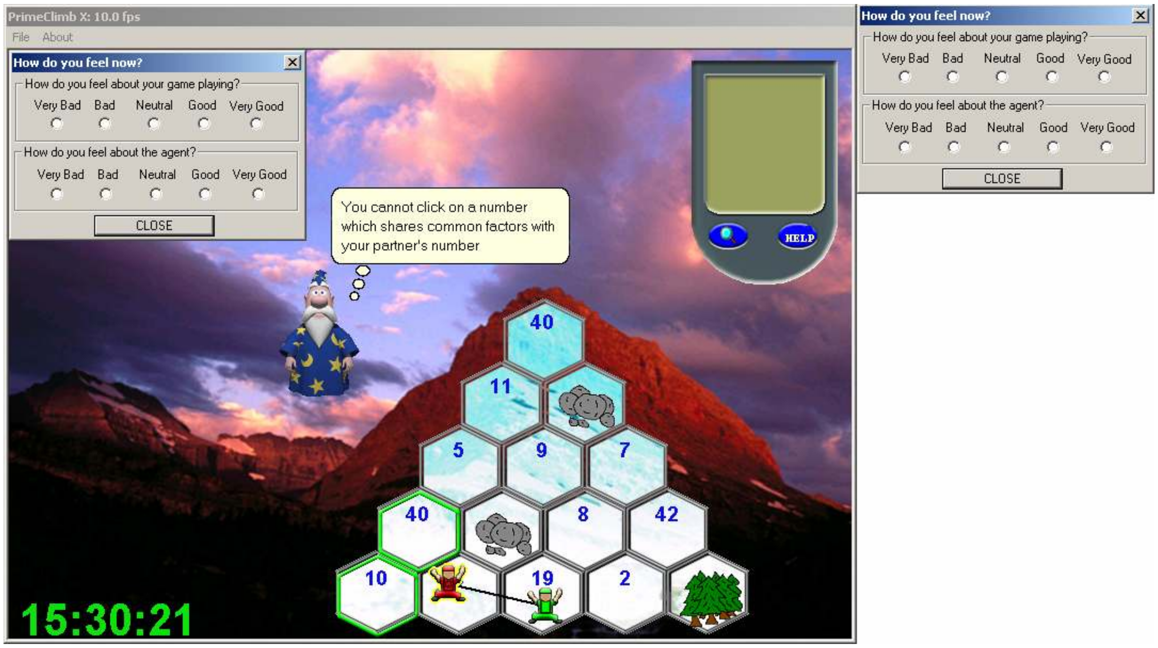
The direct approach to model validation overcomes both limitations of indirect evaluations. First, this approach does not require having a complete system built on top of the user model, since data for the evaluation can be collected either via a Wizard of Oz set up (as we did in our earlier studies (Zhou & Conati, 2003)) or by using a version of the system that does not use the model to tailor the interaction (as we did in the study we describe in Section 5.2). Second, a direct evaluation can provide a deeper understanding of the model's behaviour that is not confounded by other aspects of the application. For this reason, we used direct evaluations to test and refine different versions of our affective model. However, the main challenge of this approach is that it requires having a reliable measure of the user's affective states during the interaction for comparison with the model's assessment. Depending on the type of interaction and emotions that the model deals with, this measure can be quite hard to obtain. Thus, in this section we first describe the method for labelling student affective states during the interaction with Prime Climb that we have used in all the empirical studies on our affective model. Next, we describe the general study design of our direct evaluations (Section 5.2) and in Section 5.3 we report the results on the accuracy of the predictive model described in Section 4.

5.1 Collecting Affective Self-Reports While Playing Prime Climb

Collecting data on users' emotions with which to directly evaluate an affective user model is difficult. This is particularly true in an environment where the students' actual emotions are ephemeral and can change many times during the interaction, as we observed to often be the case with Prime Climb.

When emotions are varied and rapidly changing, it is hard for the users to describe them by using post-treatment self-reports, as was done in (Bradley & Lang, 1994; Lisetti & Nasoz, 2004; Peter & Herbon, 2006). Another commonly used method to label user emotions is to record participants using a video-camera and then ask observers to review the video to produce annotations of emotions visibly expressed during the interaction. This method has been shown to work well to measure different levels of a single emotion (e.g., level of interest (Kapoor & Picard, 2005)), or to recognize clearly separated emotions (D'Mello & Graesser, 2006). However, when we tried to use it in our research, we found that observers often had a hard time discriminating among equally-valenced feelings in our three different emotions (e.g., to discriminate between reproach toward the agent and distress toward the game). Thus, we decided to collect emotions self-reports directly from students *during* the interaction.

As far as we know, there is only one previous study that tried to use this method. In this study, a slider–based interface was used to get university students to volunteer information on their motivational state while interacting with an intelligent tutoring system (deVincente & Pain, 1999). One of the study's outcomes was that students did not volunteer information frequently (an average of 3.5 times for an interaction of about 15 minutes). Since we are dealing with much younger subjects, we were concerned that this phenomenon would be even more pronounced if we used the same approach based on volunteered student reports. Thus, we modified the method so that it can elicit self–reports from the students more frequently and provide sufficient data for model construction and evaluation. Following (deVincente & Pain, 1999), we provide an emotion–report dialog box permanently present on the side of the Prime Climb game window, for students to volunteer self–reports on their emotional states (see Figure 7 and Figure 8). However, the dialog box also pops up whenever either one of the following conditions is satisfied: (1) the student has not entered any emotion in the permanent dialog box for a period of time longer than a set threshold or (2) the underlying affective model detects a relevant change (also based on a set threshold) in what it believes to be the student's emotional state. The pop-up dialog box is necessary because a preliminary study confirmed our fears that students do not volunteer enough emotion self–reports via the permanent dialog box (Conati, 2004). The thresholds that influence the appearance of the pop-up box were adjusted through pilot studies to balance the amount of data that it allows us to collect and the level of interference that it generates during game playing (Conati, 2004). As Figure 8 shows, the emotion dialog box only elicits information on two of the three pairs of emotions targeted by our model (emotions towards the game and emotions towards the agent). We chose this design because we felt that dealing with three pairs of emotions would be too confusing for our young subjects, and because sixth and seventh grade teachers suggested that students would have more problems in reporting emotions toward themselves than toward the game or the agent.



How do you feel now?

How do you feel about your game playing?

Very Bad ○ Bad ○ Neutral ○ Good ○ Very Good ○

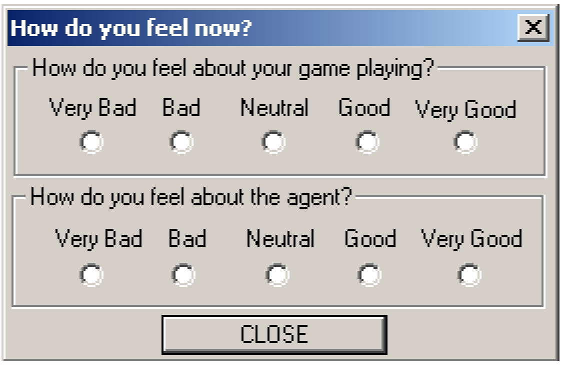
How do you feel about the agent?

Very Bad ○ Bad ○ Neutral ○ Good ○ Very Good ○

CLOSE

You cannot click on a number which shares common factors with your partner's number

Figure 7. Interface with both the permanent and pop-up emotion-reporting dialog box



How do you feel now?

How do you feel about your game playing?

Very Bad	Bad	Neutral	Good	Very Good
○	○	○	○	○

How do you feel about the agent?

Very Bad	Bad	Neutral	Good	Very Good
○	○	○	○	○

CLOSE

Figure 8. The dialog box presented to the students

Data from a post-questionnaire on interface acceptance, which 20 students filled in as part of a study to test the final version of the self-report mechanism, showed good user acceptance (Conati, 2004). For instance, the students' average ratings (on a Likert scale where 1 = strongly disagree, and 5 = strongly agree) for the statement "The popup dialog box interfered with my game playing" was 2.8 (st. dev. 1.4), while the average ratings for "It bothered me having to tell the system how I feel" was 2.1 (st. dev. 1.1). We also found that the negative emotions self-reports were only a small fraction of the self-reports generated by the students who reported annoyance with the dialog box. These results suggest that, even when subjects expressed annoyance with the dialog box, this annoyance did not necessarily translate into annoyance with the game or the agent. These findings are a quite encouraging for researchers interested in evaluating affective models, because they indicate that subjects can tolerate to some extent the interference caused by the artefacts designed to elicit their emotions.

5.2 Study Design

The primary goal of this study was to collect data from students to refine the model's event appraisal mechanism as we described in Sections 4.2.2.1 and 4.2.2.2. However, we also reused data from this study to evaluate the refined predictive model, as we will describe in Section 5.3.

Sixty-six 6th and 7th grade students from three local schools participated in our study. The study took place in the schools, and each study session had to be held during a class period (40 minutes) to avoid creating too much disruption to regular class schedules. Because of limited computer availability, we could only run two students at time. The two students were excused from the class for that period and joined the experimenters in a room provided by the school for the experiment. Each session was designed to last at most 30 minutes so that there would be sufficient time for students to get to the study room and return to their class for the next period. Students first took a pre-test on factorization knowledge. Next, they were told that they would be playing a computer game, and received a demo of Prime Climb with the emotion self-report mechanism. They were told that the game contained a computer-based agent that was trying to understand their needs and help them play the game better. Therefore, the students were encouraged to provide their feelings whenever their emotions changed so that the agent could take them into account when providing help. We did not deem it necessary to provide any further way to engage the students in the task because, from the several studies we had already run on Prime Climb, it was apparent that the mere fact of playing a computer game during school time was sufficient to greatly engage the students, at least for the short period of playing time necessary for the study. This first phase of the experiment lasted at most 10 minutes.

Next, participants played Prime Climb for about 10 minutes. Each student played with an experimenter as a climbing companion. Due to time and space constraints, we had to run two experimenter/student pairs in parallel in the same room, sitting side by side, as shown in Figure 9. We did not make students play together because we wanted to avoid the extreme emotions toward the playing partner that we often observed with that set–up, given that our affective model currently does not model these emotions. To further limit the impact of students' feelings toward their climbing companion, experimenters were instructed to play as neutrally as possible, trying to avoid making mistakes (although mistakes did happen on some of the mountains with larger numbers) and to avoid leading the climb too much. Furthermore, students did not know which of the two experimenters they were playing with. They were told that the game randomly assigned their partner, so that the partner was not necessarily the experimenter sitting across from them. This measure reduced the student's tendency to make eye contact or attempt to verbally communicate with the experimenter.

To reduce as much as possible the distraction generated by having students sitting side by side, students were reminded before the beginning of each game that they would not be playing with each other. Experimenters noted that some students did glance across to check the progress of the other student's game, but on most occasions this occurred at the end of a game level while

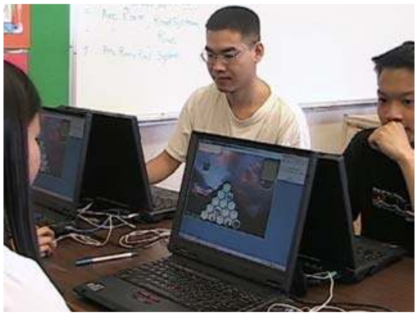


Figure 9. The study setup

they were waiting for the next level to load. However, if one student was observed to be particularly disruptive and disturbed the other student in the room, then the data from both students was discarded. In practice, this happened once or twice in each school. The discarded students were not included in the count of student participants mentioned above.

During game playing, the Prime Climb agent generated pedagogical interventions to help the student learn from the game, by relying on the model of student learning mentioned in Section 3 (Conati & Zhao, 2004). All of the agent's and student's actions were captured by the version of the affective model with the appraisal network based on a subjectively defined structure. The model was updated in real time to direct the appearance of the pop-up dialog box, as described in Section 5.1, but the pedagogical agent did not use it to direct its interventions. It should be noted that this is the reason we can re-use log files from this study to evaluate successive versions of the affective model[5].

Log files of the interaction recorded all of the events that occurred within the game, the student's reported emotions, and the corresponding model assessments. After game playing, students completed a post-test on number factorization[6], and four post-questionnaires: one to indicate the goals they had during game playing, one on interface acceptance, and the two post-questionnaires on the events that affect *Have Fun* and *Learn Math*, described in Section 4.2.2. Although this may seem quite a lot of material for young children to have to deal with, each test/questionnaire was designed and pilot-tested so that this final phase of the experiment lasted at most 10 minutes. As with the emotion self-report collection mechanism, we did our best to strike a balance between the amount of data we needed for model construction/evaluation and avoiding student fatigue that could make the data unreliable.

5.3 Evaluating the Predictive Part of the Affective Model

We evaluated the predictive part of the affective model in two stages. First, we evaluated the model's event–appraisal mechanism, independently from the performance of the sub–network for goal assessment. To do so, we assumed that the students' answers to the goals post–questionnaire were an accurate representation of the goals that they had during game playing. We used these answers to set the values of goal nodes in the appraisal network, rather than relying on the model's own assessment of student goals. Second, we tested the complete predictive network by repeating the evaluation with the model's own assessments of student goals during the interaction.

Before describing the results of each of the two evaluations, we illustrate the common evaluation method we used. We measured the performance of each model to be tested via a simulator that replays the event logs from the study described in Section 5.2 with that model. The simulator includes the execution of an additional '*no agent action'* event after each student action that was not followed by an agent intervention. This "no agent action" event had not been recorded in the original log files because its relevance was discovered through the data analysis for model refinement described in Section 4.2.2.2.

We performed cross-validation on model accuracy by using the following well-known random resampling method (Mitchell, 1997). We divided the set of students into a training set and a test set of equal size using random selection. We then used the data from the students in the training set to train the necessary CPTs in the model, and ran the event logs of the students in the test set through the simulator to produce a measure of model accuracy (computed as we describe below). We then randomly divided the original set of students into a new training set and test set, and

[5] It should also be noted that the study is not a Wizard of Oz, since the agent acts autonomously, even if it does not use the affective model, and the experimenter plays the role of a human player.

[6] Post–test data was collected for purposes not related to this research.

performed the same evaluation steps again. In total, we performed this procedure 100 times[7].

For each test set, we measured model accuracy by computing how often the model's assessment agreed with the student's reported emotions at corresponding times. To enable the comparison, we translated both the students' reports for each emotion pair (e.g., *Joy/Distress*) and the model's probability over the emotion node corresponding to that pair (e.g., the node *emotion-for-game* in Figure 5) into 2 values: positive (indicating the element with positive valence in the pair, e.g., *Joy*) and negative (indicating the element with negative valence, e.g., *Distress*). A student report was classified as positive if it was higher than 'neutral' in the dialog box, and as negative if it was lower. The model's assessment was classified as positive if the probability of the corresponding emotion node was higher than a set threshold, and negative otherwise. The threshold value of 0.65 was determined using the data from an earlier empirical evaluation (Conati & Maclaren, 2004). For each emotion pair, we report individual model accuracies in detecting the positive and the negative emotions. These correspond to standard measures of true positive rate (sensitivity) and true negative rate (specificity). We also report a combined accuracy that is the average of the two.

It should be noted that making a binary prediction from the model's assessment is guaranteed to disagree with any 'neutral' reports given by the students. The only way to fix this problem in the predictive network would be to add a third value to each emotion node that represents neutrality with respect to that emotion type. However, altering the emotion nodes' CPTs to include this additional value would not be trivial. An alternative is to catch at least some instances of neutrality in the diagnostic part of the model. We found that 65 student reports were neutral for both *emotion-for-game* and *emotion-for-agent* (63% and 58% of the neutral *emotion-for-game* and *emotion-for-agent*, respectively). Because neutrality on both emotions corresponds to a low level of emotional arousal, this state should be easily picked up by adequate physiological sensors in the diagnostic part of the model (see Figure 1). This is a clear example of a situation where the observed evidence of a student's emotional state can be combined with predictive assessment, and we will discuss our investigations in this direction in Section 6.

5.3.1 Evaluation of the Event Appraisal Sub-network

Table 2 shows the results of using the mechanism discussed in the previous section to evaluate the refined appraisal network from Section 4.2.2. As we mentioned earlier, in order eliminate possible confounding factors deriving from inaccuracies in the goal assessment network, the values of goal nodes were directly derived from the students' answers in the goal post-questionnaire. Although the goal *Want Help* was added to the model after the study and thus did not have a pre-dedicated item in the post-questionnaire, we were able to derive its value from the questionnaire item *'I wanted help when I became stuck'*, originally used together with another item to assess the goal *Succeed By Myself*.

In order to assess how well our model performed compared to a simpler approach, we calculated the baseline accuracy of predicting the emotion with the highest probability based on the frequency of emotions occurring in the students' reports. Because our data set has a much higher number of positive data-points for each emotion pair (see Table 2), the baseline model

[7] We used this method for cross-validation because we did not have enough data (especially negative data-points, as we will see in a later section) to perform a traditional N-fold cross-validation, where the N test/training pairs are non-overlapping partitions of the data and N is large enough to allow for measures of statistical significance. The drawback of random re-sampling, however, is that the test/training sets it generates are not independent and thus violate one of the assumptions required by standard tests for statistical significance. Thus, although random resampling is commonly used in machine learning to deal with limited data (Mitchell, 1997), any statistically significant results that it generates should be interpreted as significant trends.

Table 2. Emotional belief accuracy of the affective model

Emotion	Accuracy (%)		Data-points
	Mean	Std. Dev.	
Joy	64.38	4.77	170
Distress	65.63	16.20	14
Combined Joy/Distress	65.01*		
Admiration	64.76	6.44	127
Reproach	34.67	10.38	28
Combined Admiration/Reproach	49.72		

* Significantly above the baseline accuracy

always predicts *Joy* and *Admiration* and would thus have an accuracy of 100% in predicting these emotions, but 0% in predicting *Distress* and *Reproach*. Therefore, the model's combined accuracy would be 50% for emotions towards the game (*Joy/Distress*) and 50% for emotions towards the agent (*Admiration/Reproach*). The model's accuracy in predicting students' emotions towards the game is a 30% improvement over the baseline, and is significant[8]. However, the model's poor performance in predicting *Reproach* has reduced its overall accuracy in predicting feelings towards the agent and there is no significant difference from baseline accuracy (p=.62).

To understand the reasons for the model's poor performance on *Reproach* we engaged in a detailed analysis of the model's assessment in relation to the interactions simulated from the log files. This analysis showed that approximately 50% of the misclassified *Reproach* data-points, and approximately 28% of the misclassified *Admiration* data-points*,* were due to the fact that the students' declarations for the goal *Want Help* at the end of a game session did not seem to consistently match with whether they were trying to achieve this goal during the game. Five students did not declare the goal *Want Help*, but they reported *Reproach* towards the agent when they began to fall and the agent did not intervene, suggesting that they did want help in these situations. Three students declared the goal *Want Help* but then reported *Admiration* instead of *Reproach* when they fell repeatedly and the agent did not intervene, suggesting that they actually did not mind the lack of help. These findings confirm what we had already seen from the data analysis in Section 4.2.2.2, i.e., that goal priority can change during the interaction. As we discussed in that section, we currently don't have enough data to model goal evolution in a principled way, and thus our model still includes the incorrect assumption of static goals and cannot model correctly those students who have shifting goals. The influence of this assumption on the *Reproach* inaccuracy reported here is amplified by the fact that goal nodes were set to deterministic values based on evidence from student questionnaires. Deterministic values have a higher negative influence than probability distributions on model assessment when they do not actually reflect the current student's goals.

A second factor that explains an additional 25% of the misclassified *Reproach* data-points is that using only previous math knowledge to help assess the relative priority some students gave to succeeding by themselves vs. receiving help incorrectly modeled four students. In each case the students reacted as expected for the goals they had declared, for example, one student had declared the goal *Succeed By Myself*, and had subsequently reported *Reproach* when the agent intervened after a repeated fall. However, in each case, the math knowledge of the student indicated that the model should give a low priority to the goal that was not satisfied by the agent's action. Thus the negative impact of giving help (or not giving help, in some cases) was underestimated. This result indicates that there are other traits that should be taken into account to correctly model priority shifts for some individuals.

[8] All measures of statistical significance when comparing model performance to the baseline refer to a two-tailed one-sample t-test with $p < 0.05$

Table 3. Comparing the emotional belief accuracy of the affective model using goal evidence and goal population priors

Emotion	Accuracy using goal evidence (%)		Accuracy using population priors (%)		Data-points
	Mean	Std. Dev.	Mean	Std. Dev.	
Joy	64.38	4.77	64.97	4.39	170
Distress	65.63	16.20	73.09†	18.31	14
Combined J/D	65.01*		69.03†*		
Admiration	64.76	6.44	61.00†	6.00	127
Reproach	34.67	10.38	48.01†	12.10	28
Combined A/R	49.72		54.51†*		

* Significantly above the baseline accuracy † Significant increase/decrease from model with goal evidence

5.3.2 Evaluation of the Predictive Model Using Model's Assessments of Student Goals

Our assessments of the accuracy of the predictive affective model have thus far been limited to the appraisal sub-network. That is, we have used student answers to the goal post-questionnaire as evidence for setting goal nodes in the appraisal network, so as to separate it from the model's goal assessment mechanism and isolate inaccuracies in the event–appraisal mechanism. However, information on students' goals will not be available when using the model in real-time during game playing. Instead, the model's own probabilistic assessments of the students' goals will be used.

In order to determine how well the complete predictive model will perform during real-time interactions, we evaluated it by using the simulator described in Section 5.3 and allowing the model to use its own probabilistic assessments of the students' goals instead of the evidence from the students' post-questionnaires. However, we used the frequencies of goals declared by students from a previous study (Zhou & Conati, 2003) to help inform the model's goal assessments by creating population priors for each goal being assessed[9].

Table 3 compares the accuracy of the model using goal evidence and using goal assessment with population priors. As the table shows, the model's performance using goal assessments increased significantly for *Reproach* and *Distress,* although the value for *Reproach* is still below 50%[10]. The increase in accuracy for *Distress* produced a significant improvement in the model's combined accuracy for emotions toward the game, which is now 38.06% over the baseline. Although there is also a statistically significant decrease in *Admiration,* the decrease is small compared to the increase for *Reproach*. As a result, the combined accuracy of the model with goal assessment is significantly higher than the accuracy of the model with goal evidence for emotions towards the agent, and becomes also significantly higher than the baseline (9.02%).

Thus, from a practical standpoint, trends for all of the emotions except *Admiration* are in favour of the probabilistic goal assessment, and the decrease for *Admiration* is small. The good performance of the goal assessment mechanism should not come as a surprise given our previous discussion on the dynamic nature of student goals. When goal nodes are set as evidence, their values are fixed throughout the interaction, no matter what students do during game playing. When they are associated with a probability distribution, the distribution changes as evidence on student actions comes into the goal–assessment network. Although these changes cannot fully

[9] Since goal nodes are not root nodes, population priors are included by (i) adding a fictitious root node as an additional parent for each goal node in the goal–assessment network; (ii) setting the CPT of the root node to the population prior for the goal.

[10] All measures of statistical significance comparing different versions of the affective model are based on a two-tailed t-test with $p < 0.05$

reflect changes in the students' actual goals because the network models goal nodes as static, the changes in the probability distributions still approximate the students' changing goals better than the immutable evidence values. Thus, for instance, the assessment of *Reproach* improves because using probabilities for the goals *Succeed By Myself* and *Want Help* allows the model to correctly assess some of the students who declared that they did not have the goal *Want Help*, but during the interaction displayed behaviours showing the contrary. These were cases where the model's assessment was already very close to the threshold value used to classify *Admiration* vs. *Reproach*, thus the fluctuations in the goal probabilities caused by the student's interface actions were sufficient to steer the model toward the right assessment, despite the lack of a mechanism to model changing goals. The rest of the cases were still misclassified, indicating that the model still requires an ability to assess changes in student goal priorities to achieve higher accuracy.

5.4 Summary and Discussion

In Sections 4 and 5, we have described how we built and evaluated the predictive part of our probabilistic model of student affect. First, we constructed the part of the model that assesses student goals, using correlations between personality, goals, and interaction patterns found in data collected during user studies. Next, we specified an initial structure for that part of the network that represents the *appraisal mechanism* (i.e., how the mapping between student goals and game states influences student emotions). We completed the construction of the sub-network using empirical data collected from real users.

Finally, we described the method and results of a direct evaluation of the predictive affective model. We first assessed the accuracy of the appraisal mechanism using evidence on student goals generated from questionnaires to set the network's goal nodes, and showed that the model could achieve an encouraging level of accuracy. We then assessed the accuracy of the complete predictive model, including the sub–network for goal assessment, and showed that it has better accuracy than the model with goals set from evidence, because it is better able to capture the dynamic nature of some student goals. This result indicates that the model can achieve reasonable real-time accuracy during interaction with Prime Climb, when it can rely only on evidence coming from student interface actions. However, we showed that our two simplifying assumptions about student goals, i.e., that they remain the same throughout the game session and all have same priority, limit model accuracy, especially in detecting negative feelings towards the Prime Climb agent. Thus, a possible direction for future work is to investigate effective ways to remove these two assumptions, with special attention to constructing a clearer picture of how the user's goal priorities fluctuate during game sessions. However, we expect that this task will be very difficult, given that we are essentially trying to do is *plan recognition* (one of AI's notoriously difficult problems) in a highly dynamic environment.

All in all, however, we should not be surprised that the predictive part of the model by itself does not achieve top levels of accuracy. Even humans often need to integrate information on both potential causes and visible effects of the interlocutor's emotional reactions to compensate for the limited reliability of each in recognizing emotion.

Therefore, before attempting to further refine the predictive part of the model, the next stage in our research was to investigate if and how information on the Prime Climb players' emotional reactions can improve the model's assessment, i.e. how to add a diagnostic component to the model (see Figure 1). We describe this investigation in the next section.

6 Investigating Physiological Sensors as a Source of Diagnostic Affective Evidence

The predictive part of our affective model will assess whether the student is feeling a negative or a positive emotion towards the game, the agent, or the student herself (i.e., the *valence* of a student's emotions towards these entities). This component currently has two main limitations. First, as we saw in the previous section, the model's accuracy for predicting feelings of *Reproach* towards the agent is quite low. Second, by design the predictive model cannot predict the level of *arousal* of the emotions arising during game playing, which can play an important role in deciding if and how an agent should act upon the student's emotions. For instance, if the model predicts *Distress* or *Reproach* but the student is feeling calm overall because these feelings are quite mild, then it may not be as important for the agent to intervene as it would be if the student's negative feelings were strong. Similarly, an interaction aimed at calming down a happy student to improve concentration may cause confusion if the student's level of arousal is not high enough for the positive emotion to be disruptive.

Existing literature on emotion recognition suggests that integrating physiological evidence on the student's current affective state into our affective model may help overcome these two problems. We have started investigating two ways of using physiological data in our model:

1. Tension in specific facial muscles can be measured using Electromyogram (EMG) sensors, and has been shown to be correlated with affective valence (Lang et al., 1993). Heart rate (HR), which can be calculated from the measurements of the Blood Volume Pulse (BVP) sensor, has also been used as an indicator of affective valence (e.g., Papillo & Shapiro, 1990).

   Adding evidence from these physiological sensors to our model may enable it to produce an assessment of the student's overall affective valence. This would help it to discriminate which of the student's current emotions is dominant, improving overall emotion assessment, and detection of *Reproach* in particular.

2. Skin Conductance (SC) has been shown to be positively correlated with levels of affective arousal (Dawson et al, 2000; Lang et al., 1993). Evidence from this physiological signal may enable the model to assess whether the student is in a state of high or low arousal, and thus how important it is for the pedagogical agent to take into account the student's affect in its interventions.

However, most of what is known on the links between physiological signals and emotions experienced was achieved either in controlled laboratory conditions, or for modeling affective states less complex than those we are targeting. For instance, (Lang et al., 1993) used images to induce specific affective states in subjects. Similarly, in other experiments that have built on this work subjects were deliberately frustrated (Scheirer et al., 2002), or asked to express a set of specific emotions (Picard et al., 2001). Experiments in less controlled conditions focused on simpler tasks, such as detecting a single strong emotion (e.g., the level of anxiety in drivers (Healey & Picard, 2005), interest during interaction with a computer-based tutor (Kapoor & Picard, 2005)), lower-level affective measures such as valence and arousal (Prendinger et al., 2005), or overall emotional predisposition over the course of a complete interaction (Mandryk et al., 2006).

In contrast, we want to explore the performance of these sensors for the instantaneous detection of affect, in a setting where students are allowed to spontaneously experience multiple emotions, possibly conflicting, varying rapidly and that may be expressed more subtly than those induced in laboratory settings.

This section describes the results of this investigation, as well as the effect of adding data from physiological sensors to our affective user model. Our current efforts focus on understanding the value of each of the sensors as individual sources of affective information, as opposed to directly

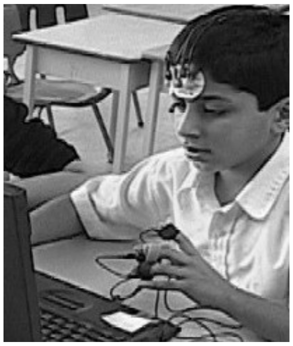


Figure 10(i). A close-up view of a student wearing the SC, BVP, and EMG sensors

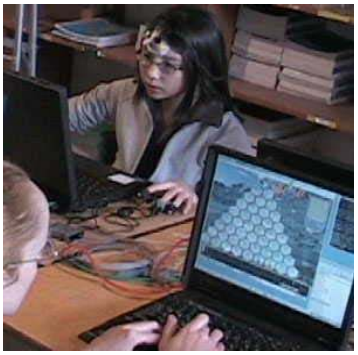


Figure 10(ii). A student interacting with Prime Climb while wearing the sensors.

combining them as has been done by other researchers (e.g., Kim & André, 2006; Vyzas & Picard, 1998). This approach allows us to determine whether the methods needed to collect information from that signal can be used within the constraints imposed by the environment and by the requirements of our model. It also allows us to develop individual nodes for our DDN-based framework, which can then be used to modularly combine evidence depending upon which signals are available/suitable to use,

The rest of this section is organized as follows. First, we describe the user study that we used to collect the physiological evidence from students interacting with Prime Climb (Section 6.1). Next, we describe our analysis of the EMG signal, including the effect of incorporating evidence from the signal into the affective model (Section 6.2). We focus on EMG analysis because it is the signal that showed the greatest potential for our purposes. In Section 6.3, we summarize the largely negative results we obtained with the other two sensors, and discuss potential reasons for these outcomes.

6.1 Study Design

In order to investigate the mapping between the affective states of Prime Climb players and evidence collected using physiological sensors, and how it can be used to improve the accuracy of our affective model, we ran a study designed to simultaneously collect both physiological evidence and accompanying affective labels.

The overall study design and materials were the same as those described in Section 5.2, with the following differences. The first is that the study participants wore four physiological sensors throughout the interaction (as shown in Figure 10 (i) & (ii)). Students wore an SC sensor and a BVP sensor on their fingers and two EMG sensors on their forehead (one on the corrugator muscle and the other on the frontalis muscle)[11].

A second difference was the content and presentation of the hints generated by the Prime Climb agent. While for general hints (e.g., reminders to use the magnifying glass, reminders to think about common factors when climbing, reminders to think about factorization when climbing) the content differences were mostly confined to wording, more specific hints underwent substantial changes. First, we added hints to define relevant factorization concepts (e.g., 'Factors are numbers that divide evenly into the number'). Second, we added more extensive examples to illustrate these concepts than what we used in the previous hints version. Because these new examples often included several lines of text, they were presented using a pop-up dialog box instead of the agent's speech bubble as was done previously.

[11] We used the sensors in the Procomp Infiniti package and Biograph software from Thought Technologies $^{(TM)}$

A third difference was with the parameters of the model of student learning used to direct the agent interventions (Manske & Conati, 2005). We introduced the differences in the agent's hinting behavior and learning model in this study because we needed to validate them and could not afford to run a separate study for this purpose. We were aware that, because the current agent behaviors were not those we used to define the appraisal part of the casual affective model, the model accuracy on this study's data may be limited. However, a decrease in the accuracy of the predictive model would not interfere with the goal of this study, i.e., investigating the potential of physiological data in modeling student affect during interaction with Prime Climb.

The fourth and last difference is that students did not fill out the post–questionnaires on goal satisfaction; instead the time was used to administer a more extensive pre-post test on factorization knowledge, necessary to evaluate of the pedagogical effectiveness of the changes we introduced in the agent behavior.[12]

Forty-one students from two local schools participated in the study. Log files of the interaction included the student's and the experimenter's actions, the agent's interventions, the student's reported emotions and the physiological signals. In the next section, we describe our data analysis of these log files with respect to the EMG signals.

6.2 Analysis of Electromyogram (EMG) Signals for Valence Indicators

Based on numerous other works (e.g., Healey & Picard, 2005; Lang et al., 1993; Mandryk et al., 2006), our analysis of the EMG signals looked for possible indicators of affective valence.

Following (Scheirer et al., 1999), we originally tried to use two EMG sensors, one on the corrugator muscle to detect frowns and one on the frontalis muscle to help detect eye-brow raises. We felt it was important to distinguish between these two expressions because frowning tends to be an indicator of negative valence (Lang et al., 1993), whereas raised eyebrows do not. While activity on the frontalis muscle is not viewed as a strong indicator of positive valence (Cacioppo et al., 1993), we were hoping to gain a consensus of positive valence by combining evidence from this signal with additional indicators of positive valence from signals such as the students' heart rates (Papillo & Shapiro, 1990).

Unfortunately, however, we discovered that several of our subjects did not have sufficient forehead space to reliably accommodate the two fairly large EMG triodes available to us. With these students, we could not attach the EMG on the frontalis muscle firmly enough to obtain a reliable signal. Thus, we decided to limit our analysis to the EMG signal measured on the corrugator muscle, hoping to at least partially reproduce results from previous laboratory studies showing that it can be a reliable indicator of negative affect (Lang et al., 1993). In the remainder of this section, we first describe how we created the mapping between this EMG signal and student self-reports. Next, we discuss the mapping results in terms of signal reliability to predict the valence of students' affective states while playing Prime Climb. Finally, we describe how we added signal evidence to our affective model and the effect of this refinement on the model's accuracy.

6.2.1 Analysis Method

The goal of our analysis was to create a set of data-points of the form <*affective valence, signal prediction*> for each game event currently included in our affective model. These data-points could then be used to assess the reliability of the EMG signal in predicting the valence of students' affective reactions to the relevant game events. The events we used were: a student's successful climb, fall, use of the magnifying glass, and agent interventions. We did not use the

[12] The lack of post-questionnaires on goal satisfaction is the reason why we could not re-train the appraisal component of the causal model with the new agent behaviors.

Table 4. Valence labels for students' affective self-reports

Affective Valence	Type of self-report
Positive	Both answers to the emotion questions were positive, or one was strongly positive ('very good') and the other was neutral.
Negative	Both answers to the emotion questions were negative, or one was strongly negative ('very bad') and the other was neutral.
Unknown	The report was neither strongly positive nor strongly negative.

events *encountering big numbers* or *being ahead of a partner* because they are side-effects that co-occur with the aforementioned events.

The first step in our analysis was to obtain the data on affective valence from the students' self-reports of emotions towards the game and towards the agent. To do so, we used the scheme shown in Table 4. In this step, we focused on the self-reports that could clearly be translated to positive or to negative valence (67 self-reports out of the 180 collected during the study, see Table 4 for the definition), in order to obtain a more reliable mapping between the EMG signal and the overall valence of the student's affective state. Of the remaining 113 self-reports (labelled as "unknown" in Table 4), 99 will be used later to determine how our EMG signal analysis translates to more difficult assessments involving feelings that have ambiguous valence. That is, feelings with weak or conflicting valence. The final 14 reports received neutral answers for both of the emotion questions, and are therefore guaranteed to be misclassified in terms of valence. Thus, we did not include them in this analysis.

Next, in order to assign appropriate values for *affective valence* and *signal prediction* for each data-point corresponding to a game event, we needed to address the following issues:

1. To assign the *affective valence* value, we needed to find which game events contributed to generating a particular self-report. This issue exists because students are not required to generate self-reports after each game action. The reader should recall that the pop-up affective dialog box appears either when the existing affective model detects a change in the student's affective state, or when a predefined amount of time has gone by. Because several events could happen between two consecutive appearances of the pop-up dialog box, it is not trivial to discern which events have influenced a specific student self-report and can thus be labelled with the valence label from that self-report.
2. For each of the relevant game events, how to isolate the EMG signal that it generated, avoiding as much as possible the overlap with influence from other game events or external factors.

To address (1), we chose to consider only the final game event before each given self-report. That is, the last game event before a self-report generated a data-point with valence derived from that self-report as described earlier, and signal-prediction computed as we describe shortly. Although all the events that occurred before a given self-report could have influenced it, we chose this narrow focus because only the last event in the sequence has a clear connection with the reported feelings, not obscured by reactions to subsequent events.

To address issue (2) above, for each relevant game event we considered only the EMG signal in the four-second interval immediately after the event occurred. The period of four seconds was chosen based on the work by (Lang et al., 1993), because it balances the amount of time required to detect a response in each signal against avoiding recording the student's reaction to the next game event or to any other event outside the game[13].

[13] It should be noted that, although we focused on the events that are currently included within the causal model (i.e., climbs, falls, using a tool, and agent interventions) we did not include the occasions on which the agent did not

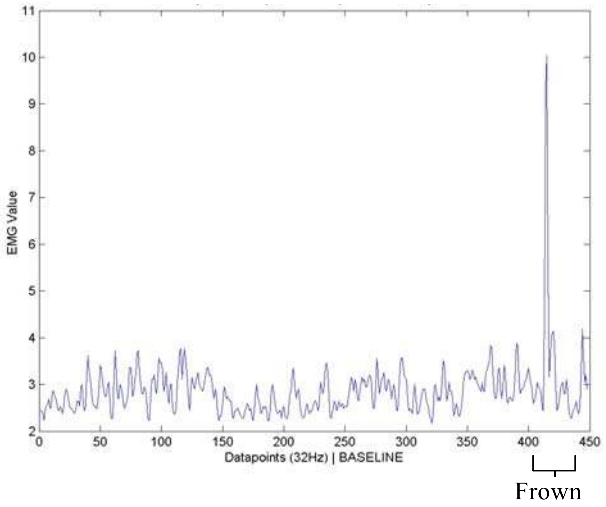


Figure 11. Example of EMG signal recorded for rest followed by frowning

To summarize, in our analysis the last game event that happens before a student self-report generates one data-point of the form <*affective valence, signal prediction*>, where the value of *affective valence* is derived from that self-report and *signal prediction* is computed by analyzing the EMG signal in the four seconds following the event. We now proceed to describe how we performed the actual signal analysis.

6.2.2 Creating Predictions from the EMG Signal

Electromyography (EMG) measures muscle activity by detecting surface voltages that occur when a muscle is contracted. The corrugator muscle, located on the forehead between the eyebrows, is used to generate facial expressions such as frowning. Both (Lang et al., 1993) and (Cacioppo et al., 1993) report that greater EMG activity in this area tends to be associated with expressions that have negative affective valence. Figure 11 shows an example of an EMG signal recorded over the corrugator muscle during a period of rest followed by the subject frowning.

When selecting which features of the EMG signal to analyze, our aim was to build on existing results by exploring whether these results could be transferred to our environment. Therefore our selection of features was driven by whether the feature had a previously established mapping to labels of affective valence.

Recent work on exploring the link between EMG and affective expressions in environments where subjects are interacting with computers has used features such as the mean and standard deviation of the raw signal (Bosma & André, 2004; Mandryk et al., 2006; Partala & Surakka, 2004; Picard et al., 2001), but only the mean has been shown to hold a reliable mapping with affective valence. Some investigations explored more complex measures, such as the gradient and the change in gradient of both the raw and normalized signals (Vyzas & Picard, 1998), but failed to establish a link between these features and valence. Thus, we focused on the mean of the of raw EMG signal for our analysis.

intervene immediately after a student action. This is because the four-second period following this event overlaps with the previous event, and making it impossible to isolate the EMG signal that it generates.

For each data-point, <*affective valence, signal prediction*>, *signal prediction* was set by comparing the value of mean EMG to a simple threshold that was calculated using the mean of the raw signal for the student's entire interaction (as shown in Eqn 1).

$$\textit{signal prediction}(e,s) = \begin{cases} \textit{positive} \text{ if } \textit{mean}(EMG\_e) < \textit{mean}(EMG\_s) \\ \textit{negative}\text{, otherwise.} \end{cases} \qquad (1)$$

where *s* is the current student

e is the last game event before the self-report used to create the value for *affective valence*

EMG_s is the set of values recorded for the EMG signal during the entire interaction for student *s (overall signal mean)*

EMG_e is the set of values recorded for the EMG signal during the 4-second period following event *e*

If the mean EMG value was below the threshold, then *signal prediction* was positive valence, if not then *signal prediction* was negative valence. Our choice of using the overall signal mean as a threshold is based on the experimenters' observations that most students experienced both positive and negative affect at some point during the interaction, thus the overall EMG mean would be higher than the signal mean in those intervals where the student did not experience negative affect.

An alternative method for generating valence predictions from EMG is to compare the EMG signal over the interval of interest against a baseline signal recorded during a resting period before the experiment (Cacioppo et al., 1993). Unfortunately, due to limitations on time with the students, in our study we could not set up an idle "resting time" that we could use as a baseline. Nor could we use their signals just before starting the interaction with the game as a baseline because of their initial feelings of excitement at participating in an activity other than class work.

Having completed the construction of our data-points of the form <*affective valence, signal prediction*>, we proceeded to assess the reliability of using the signal to predict the affective valence of our Prime Climb players. Table 5 shows a confusion matrix comparing the values for *affective valence* and *signal prediction* for each data-point. The high number of true negative predictions by the signal compared to false positives (where *affective valence* is negative but *signal prediction* is positive) indicates that the mean of the EMG signal is a reliable feature for assessing feelings with negative valence (89% accuracy in detecting negative valence). This result is encouraging because it indicates that evidence from the mean of the EMG signal may help us to achieve our goal of improving the model's assessment of *Reproach*. However, the number of true positive predictions by the signal compared to the number of false negatives (where *affective valence* is positive but *signal prediction* is negative), indicates that this feature is not reliable when assessing feelings with positive valence. Both of these findings agree with the results of other work (e.g., Cacioppo et al., 1993; Lang et al., 1993).

The high number of false negatives may be due to phenomena associated with other cognitive processes involved in the task the student is concentrating on, e.g., squinting, or frowning in concentration. In addition, 12 of the 41 students who took part in the study only generated self-

Table 5. Confusion matrix comparing the valence predictions made by the mean EMG signal with the valence labels produced from the students' self-reports.

		Valence Label	
		Positive	Negative
Signal Prediction	Positive	33	1
	Negative	41	8

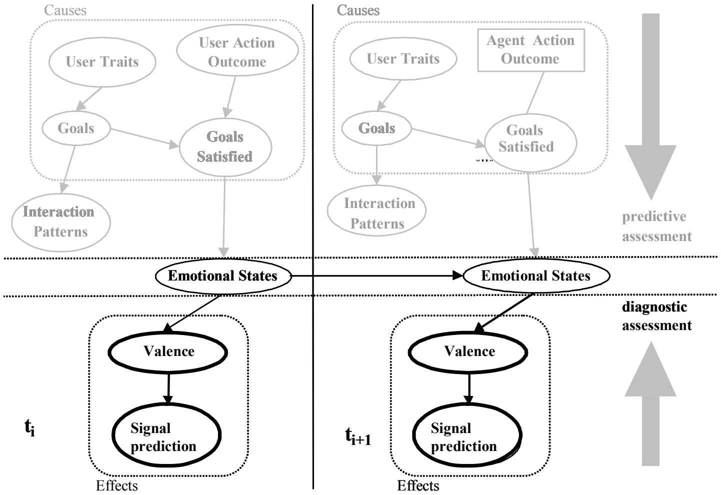


Figure 12. Two time-slices of the affective model, with proposed nodes highlighted.

reports that were translated to positive valence. If these students really experienced only positive affect during the interaction, our baseline threshold would be too low for them. That is, their overall signal mean would not be necessarily higher than the mean over segments when they experienced positive affect. Of the 41 data-points that received false negative predictions, 20 came from these students, suggesting that acquiring a more reliable baseline can indeed improve the performance of our EMG signal as predictor of positive valence.

Despite the high number of false negative predictions, a Fisher's exact test using the results in Table 5 showed that there is a marginally significant relation between the affective valence of the students' reports and the signal's valence predictions (p=.075, 2-tailed). Therefore, we decided that it would be worth exploring the addition of the mean EMG to our affective model as evidence of affective valence. We describe our approach to this task in the next section.

6.2.3 Adding the Mean of the EMG Signal to the Affective Model as Evidence of Valence

We begin by describing how we incorporated evidence from the EMG signal into the model. We then evaluate the resulting model to determine whether assessing affective valence via EMG evidence can improve the model's overall accuracy in predicting specific emotions.

In the predictive part of the affective model, a new time-slice is added in response to events that occur during game play. Therefore, the most logical approach to including EMG evidence of students' reactions to game events in the affective model is to add two binary nodes to each time-slice (shown in bold in Figure 12): (i) a *Valence* node that represents the model's overall prediction on the valence of the student's affective state; (ii) a *Signal prediction* node encoding the valence predicted by the EMG signal recorded in the four-second following the most recent event. Both nodes have values *positive* and *negative*. The *Valence* node has the *Emotional States* nodes as parents, and its CPT represents the dependency between the overall valence of the student's current affective state and her feelings towards herself or the agent, and the game. The CPT values are defined so that the probability that valence is positive/negative is proportional to the number of positive/negative emotion nodes.

The node *Signal prediction* has the *Valence* node as parent, and the link between the two nodes represents the probability of observing an expression of positive or negative valence in the mean of the EMG signal, given the overall valence of the student's affective state. The CPT for the *Signal prediction* node is computed using the frequencies from our dataset of <*affective valence, signal prediction*>.

Table 6. The CPT for the node S*ignal prediction*

Signal Prediction	Valence	
	Positive	Negative
Positive	0.48	0.17
Negative	0.52	0.83

Whenever a new time-slice is added to the model, the node *Signal prediction* is set to a value of either *positive* or *negative*, depending on the outcome of the EMG signal analysis described in the previous section[14]. Bayesian propagation then integrates this evidence with evidence on the current state of the game set in the diagnostic part of the model, and generates the model's combined assessment over the student's individual emotions.

Evaluating the Combined Model

Our aim when evaluating the affective model was to answer the following questions:

1. How does the new combined model perform on *clear-valence* data-points, i.e., those 67 data-points that have a clear indication of positive or negative valence (see Table 4), and that we used in our analysis of the EMG signal?
2. How does the new combined model perform on *ambiguous-valence* data-points, i.e., on those 99 data-points corresponding to students' self-reports containing emotions with mixed valence or only mild positive/negative valence (see Table 4). Recall that we excluded these self-reports from the analysis of the EMG signal in Section 6.2.1.

For both questions, we ran the simulator described in Section 5.3 with the new combined model and the log files from the study described in Section 6.1. However, when answering the first question we used only *clear-valence* data-points to train the CPT for the *Signal prediction* node and to test the overall model.

To compute model accuracy, we used the same measures and 100-fold cross validation with random resampling described in Section 5.3. In each fold, we used the data from the students in the training set to train the *Signal prediction* node CPT. A sample CPT created from one of the training sets is shown in Table 6. For each data-point in the test set, we compared the student's self-reports of her emotions towards the game and towards the agent with the model's prediction for these emotions. Table 7 shows the overall mean and standard deviation of the collected results. For comparison, the table also reports the accuracy of the predictive part of the model on this reduced data-set.

As the table shows, the combined model performs significantly better than the predictive model on *Joy* and *Reproach,* while the two models perform roughly the same for *Admiration* and *Distress* (p>.40). While the substantial increase in *Reproach* results in a significant increase of the combined accuracy of the *Emotion for agent* pair, the smaller increase for *Joy* results in an increase of the combine accuracy of the *Emotion for game* that only approaches significance (p = 0.067).

The fact that the increase in the accuracy for *Joy* is not as substantial as that for *Reproach* is consistent with the fact that EMG measured on the corrugator muscle was shown to be a mediocre predictor of positive valence, both in our analysis (see Section 6.2.2) and in previous research. Still, the fact that adding evidence from the EMG signal to our affective model improved its combined accuracy over both our emotion pairs provides encouraging evidence that

[14] Because we cannot define precisely when the student will react to the agent *not* intervening, we do not add evidence to the node *Signal prediction* in time-slices generated to represent a lack of agent intervention.

Table 7. Accuracy comparison between the predictive model and the combined model with EMG data on *clear-valence* data-points

Emotion	Accuracy (%)				Total number of Data-points
	Predictive Model		Combined Model (EMG)		
	Mean	Stdev	Mean	Stdev	
Joy	74.80	7.75	79.10†	7.57	74
Distress	53.48	29.83	56.70	26.18	5
J/D Combined	64.14*		67.90*		
Admiration	83.49	4.45	81.18	5.29	67
Reproach	39.11	16.75	63.02†	16.55	9
A/R Combined	61.30*		73.10†*		

* Significantly above the baseline accuracy † Significant increase/decrease compared to the diagnostic model

the EMG signal can help assess clearly valenced emotions during the students' interaction with Prime Climb.

To answer our second evaluation question, we re-ran the simulator on the combined model, still using the *clear-valence* data-points to train the CPT for the *Signal prediction* node, but testing the resulting model on the *ambiguous-valence* data-points, i.e. the 99 data-points set aside earlier.

The results of this evaluation (see Table 8), show a statistically significant increase in accuracy for *Admiration*, but also a significant decrease for *Joy* and *Distress* and no relevant change for *Reproach*. Thus, our data indicates that evidence from the EMG signal is not as valuable in helping to recognize multiple emotions with mixed or mild valence. However, we should remain aware that these results have been produced by using a single source of evidence that is known to be unreliable for positive affect and subject to inaccuracies in detecting negative affect due to the its inability to distinguish between frowns and eye-brow raises. It is not surprising that these inaccuracies are more prominent in the presence of affective states that are not strongly valenced, since these states likely generate more subtle facial expressions, difficult to discriminate by only monitoring the movements of the corrugator muscle. Adding information from other EMG sensors (e.g., measuring activity of the frontalis muscle, or the zygomatic major muscle), and other sensors linked with affective valence, may help produce more reliable information on the student's affective valence.

6.3 Discussion of Skin Conductance (SC) and Heart Rate (HR) as Affective Indicators

Following our analysis of the EMG signal, we focused on the two other signals that we had recorded during the study. Unfortunately, for a number of reasons that we will describe in this

Table 8. Accuracy comparison between the predictive model and the combined model with EMG data on ambiguous-valence data-points.

	Accuracy (%)				Total number of Data-points
	Diagnostic Model		Combined Model (EMG)		
	Mean	Stdev	Mean	Stdev	
Joy	83.66	6.11	74.15†	7.79	51
Distress	43.82	15.21	38.72†	14.08	15
J/D Combined	63.74*		56.44†*		
Admiration	58.58	8.84	71.70†	11.15	28
Reproach	25.36	7.62	25.11	12.85	33
A/R Combined	42.11		48.41†		

* Significantly above the baseline accuracy † Significant increase/decrease compared to the diagnostic model

section, we did not succeed in confirming the results of other work with regards to using features of these signals as indicators of the student's affective state.

For each signal, we will briefly describe the steps we took to analyze the signal, discuss the possible reasons why we did not achieve our expected results, and then describe how we intend to address these reasons as part of our future work.

6.3.1 Analysis Method for Heart Rate as an Indicator of Affective Valence

Heart rate can be calculated from the measurements of several different sensors, including the Blood Volume Pulse (BVP) sensor. We chose to use heart rate based on the results of other works that demonstrated links between heart rate and affective valence (e.g., Bosma & André, 2004; Papillo & Shapiro, 1990).

However, a preliminary inspection of the heart-rate measurements for each student showed that 33 of the 41 students had over 50% noise in their data. In previous work (Conati et al., 2003) we had already observed high levels of noise in the BVP signal due to the apparent sensitivity of the sensor to movement, but decided to use it for a second time to confirm our observations. It is possible that other researchers had better results with this physiological signal because their subjects and experimental set-up did not create the same amount of movement generated by kids freely playing a computer game. We still intend to look for ways to include information on heart rate as evidence within the model because, combined with EMG evidence, it may generate more reliable evidence on players' affective valence. In particular, we plan to investigate the usage of Beats per Minute (BMP) sensors that are advertised to be especially suitable for usage with children[15].

6.3.2 Analysis Method for SC as an Indicator of Affective Arousal

Skin conductance has been found to be linearly correlated to the level of arousal of emotional response (Lang et al., 1993) and it has been frequently used to measure subjects' arousal in situations that elicit some form of anxiety, e.g., stress (Healey & Picard, 2005) or frustration (Scheirer et al., 2002). The aim of our analysis was to try and reproduce these previous results and then attempt to add evidence from the SC signal to our model, which currently has no way to assess a player's level of arousal. However, our initial results did not produce the expected relation between SC and arousal levels. In this section, we will briefly describe our analysis, discuss the reasons for this result and how we intend to address them.

For SC analysis, we created a set of data-points <*signal prediction, affective arousal*> using the same methods we used for EMG analysis. That is, for each student self-report, we selected the last game event that occurred before the self-report. We then generated a data-point with an arousal label derived from that self-report, and a signal prediction computed using a threshold similar to what we used in EMG analysis. For each game event, we considered only the SC signal in the four-second interval immediately after the event occurred.

Since we did not have explicit self-reports for arousal, we attempted to derive the arousal label from the emotion self-reports, using the scheme shown in Table 9. We were aware that the Likert scale in the emotion self-report measures emotion intensity and that intensity does not directly translate into measures of arousal. However, we decided to try and see if we could still extract some information on the mapping between arousal and SC from our data by considering only self-reports clearly indicating the presence of emotions (labelled as High Arousal in Table 9), and fully neutral self-reports (labelled as Low Arousal in Table 9).

We chose to investigate two features as potential indicators: the amplitude of peaks detected within the signal over the 4 seconds following a target game event and the mean of the SC signal

[15] See for instance, http://www.dataharvest.com/Products/easysense/sensors/heart.htm

Table 9. Arousal classifications for students' online affective reports

Classification	Description
High Arousal	Both answers to the emotion questions were non-neutral, or one answer was neutral and the other was strongly positive or strongly negative.
Low Arousal	Both answers to the emotion questions were neutral.
Unknown	Neither high arousal nor low arousal classification was appropriate.

over the same interval. We chose these features because previous work (e.g., (Healey & Picard, 2005; Lang et al., 1993)) had already linked them to labels of affective arousal. We created a separate data set for each feature, but none generated a clear mapping with our arousal labels. We see two potential reasons for this result.

The first reason is that our interpretation of the students' self-reports may indeed be inadequate to generate labels of arousal. When we designed our study, we did not ask the students to report their level of arousal because we did not want to increase the level of disruption generated by the emotion dialog boxes. However, the results of this analysis suggest that in order to collect reliable data on students' affective arousal, we will need to run a separate study that uses a formal method for arousal detection. One possibility is the Self-Assessment Manikin (SAM) (Bradley & Lang, 1994), a commonly used tool that uses pictorial representations to help subjects understand the nature of the affective self-report they are being asked to give. Another possibility is to resort to external judges. While in our research this method proved to be inadequate to label specific emotions, the Prime Climb players may provide sufficient behavioural evidence for the judges to recognize different levels of arousal.

The second reason may be the inadequacy of our chosen baseline. As for EMG, we could not set up a resting period before the start of the interaction to obtain a true baseline measure for SC. Nor could we use as a baseline the students' signals just before starting the interaction with the game, because of their initial feelings of excitement at participating in an activity other than class work. (Mandryk et al., 2006) also mention this difficulty, commenting that in their first experiment often the baseline values measured were higher than some of the values during the rest of the study. Thus, as with our EMG analysis, we chose as a baseline the mean of the SC signal over the entire game session, with the assumption that the student would experience some periods of low arousal and some periods of high arousal. However, given the situation in which the students were interacting (taken out of class to play a computer game while wearing physiological sensors) it is possible that they did not have low levels of arousal at all during the interaction. The only way to tell is to collect a true baseline signal, by finding ways to extend the length of our study sessions so that we can set up adequate resting states.

7 Related Work

Probabilistic approaches based on Bayesian Networks have become quite popular in modeling user affect. Ball & Breese (1999) were the first to advocate this approach, proposing a Bayesian network that used diagnostic information from the user's linguistic behaviour, vocal expression, posture and facial expressions, to assess valence and arousal of user affect during interaction with an embodied conversational agent.

Since the initial proposal of a probabilistic model that combines predictive and diagnostic inference to form a single affective assessment (Conati, 2002), several models have followed this approach. Like our model, the Bayesian network produced by Bosma & André (2004) is intended for use by a pedagogical agent within an educational game. Following our framework, the network combines contextual information on the current state of the game with diagnostic information including user's eyebrows position, heart rate and evidence collected from skin conductance. The goal is to produce an assessment of general arousal and valence, and then use it to disambiguate the utterances students generate when playing the game. The authors report significant correlations between the physiological signals used and valence/arousal, but do not evaluate the accuracy of their complete model. Li & Ji (2003) produced a Dynamic Bayesian Network (DBN) for use in intelligent user assistance systems (e.g. monitoring car drivers to detect potentially dangerous conditions such as fatigue). The model combines contextual information with evidence in the form of head gestures, hand gestures, and eye-movements to produce an assessment of affective states that include 'fatigue', 'confused', and 'frustration'. The authors do not test the accuracy of their model with real users. Instead, they evaluate its efficiency by measuring the number of time-slices required by the DBN to identify an affective state that was continuously expressed using simulated sensor data. The probabilistic decision network developed by Prendinger et al., (2005) combines contextual information with information from SC and EMG sensors to assess the user's current levels of valence and arousal. The model is used by an agent designed to help a user cope with the negative affective states that arise during a simulated job interview scenario (i.e., states with high arousal and negative valence). The model was evaluated indirectly with users interacting with the agent while responding to questions from a specific interview script. The experimenters were unable to show that the presence of their empathic agent resulted in an overall positive effect on the users' interactions. However, they showed that the empathic agent's interventions had an effect on the way users perceive questions in terms of lower levels of arousal (or stress).

Hudlicka & McNeese (2002) also propose a framework that integrates diagnostic and causal information to model user affect, but their framework relies on fuzzy heuristic rules rather than a probabilistic approach. The fuzzy rules specify how to combine various diagnostic and predictive factors to assess the anxiety experienced by combat pilots during a mission. The predictive factors include general properties of the mission at hand, events that happen during the mission, and pilot's traits (such as personality, experience and expertise). The only diagnostic factor used is the pilot's heart rate. A very preliminary evaluation of the framework was conducted on a sample set of simulated users, i.e. made-up users with scripted behaviours desirable for testing.

In addition to the work of Prendinger et al., (2005), we are aware of only one other attempt to evaluate an affective user model indirectly, via the evaluation of a user-computer interaction directed by the model. In this work, Guinn & Hubal (2003) devised a technique to detect affect from speech, and embedded it in Avatalk, a system that trains people who must be able to convey specific affective states through speech as part of their job. Guinn & Hubal (2003) ran two field studies with Avatalk, designed to test system acceptance rather than training effectiveness. While the studies generated fairly positive results, the authors acknowledge that they cannot tell how much of the obtained results is due to the affective model, because of the presence of so many confounding variables introduced by the other components of the system.

While there are still very few indirect evaluations of affective models embedded in complete systems, increasingly more and more researchers are using direct evaluations to test proposed models or potential modeling techniques. Most of this work focuses on the assessment of valence/arousal or of a single affective states, rather then targeting multiple specific emotions. Kapoor & Picard (2005) propose a unified Bayesian approach based on a mixture of Gaussian Process classifiers to detect levels of interest in children interacting with an educational game. Their approach is designed to generate an assessment of *high interest*, *low interest*, or *'taking a break'* from facial recognition, posture recognition and information on the state of the game. When evaluated on the simpler task of classifying states of *interested* vs *uninterested* from data labelled by experts, this approach reached an excellent accuracy of 86.55%. Qu & Johnson (2005) propose a model that assesses learner motivation during interaction with a computer-based learning environment, given information on the user's attention patterns and possible plans. The model uses data from the keyboard, the mouse, and a camera focused on the student's face, to infer three motivation-related measures: confidence, confusion, and effort. In a preliminary evaluation the model achieved accuracies of 70.7%, 75.6%, and 73.2% for the three measures when evaluated using data that had been classified by a human tutor. Litman & Forbes-Riley (2006) evaluate acoustic–prosodic features from student speech, and lexical items from the transcribed or recognized speech as data sources to assess the affective valence (positive, negative and neutral), of students engaged in tutorial dialogues. They compare results of machine learning experiments using these features alone, in combination, and with student and task dependent features, showing significant improvements in prediction accuracy over relevant baselines.

One notable exception of research that, like ours, directly evaluates with user data a model designed to recognize multiple emotions is the work by D'Mello et al., (2006). The authors first coded mixed-initiative dialogues from students' interaction with an intelligent tutoring system, based on relevant conversation patterns. Next, they showed that dialogue features can be reliable predictors of three affective states relevant for learning: eureka, frustration and confusion. Finally, they tested the performance of six well-known machine learning methods for the automatic detection of the three affective states from conversational features. The overall conclusion was that the selected machine learning methods produced reasonable accuracy (the best classifier reached an accuracy of 59% for confusion, 72% for eureka, and 58% for frustration). This paves the way for using more sophisticated machine learning methods in future.

There have been various preliminary attempts to use the OCC theory in predictive models of user affect. One example is the work by Streit et al., (2004). They propose using the OCC theory for an affective user model embedded within the multi-modal dialog system SmartKom, which recommends products and services based on the user's goals, likes, dislikes, and standards. The system is implemented using logical rules, and uses abduction to infer user goals from the user's reactions to the system's generated dialog. Since the proposed affective model was still quite preliminary, the authors do not report any evidence of the effectiveness of their proposed approach. Chalfoun et al., (2006) propose using the OCC theory to model student affective reactions upon receiving the results of completing a web-based quiz. They assume that students have either one of two goals: (1) to achieve an expected mark in a post-treatment quiz; or (2) to achieve a mark above the passing mark in that quiz. However, they bypass the problem of goal assessment, and instead learn a decision tree to predict affective reactions from data on student sex, personality and test score. The authors report a prediction accuracy of 84% for their approach, although they do not provide details on how this accuracy was computed.

Investigating potential sources of affective data for diagnostic assessment has been the focus of several research groups, including assessing the effectiveness of using combined features from multiple physiological sensors. Vyzas & Picard (1998) used a combination of feature selection, Fisher projection, and a day-matrix designed to account for individual and day-to-day fluctuations, to produce an online recognition system that can distinguish between 8 deliberately expressed emotional states with an accuracy of 81.25%. By combining physiological and vocal information to detect affective valence and arousal, Kim & André (2006) produced results ranging from 69% to 92% (depending on the subject) if their classifier was trained on data from individual subjects. This accuracy reduced to a mean of 55% over all subjects if the classifier was trained on population data. The data for this analysis was collected in an environment where emotions were elicited by the experimenters following a worked script designed to evoke situations that led to a certain emotional response. Healey & Picard (2005) collected measurements from five physiological sensors (including electromyogram and skin conductance), three video-cameras and a microphone to predict levels of anxiety for subjects who experienced a sequence of different driving conditions. Using Sequential Forward Floating Selection (Jain & Zongker, 1997) to select the most appropriate combination of physiological features, they were able to distinguished four different levels of stress with 89% accuracy. Sensors have also been used to assess an emotional disposition to over an entire interaction rather than reactions to specific events. Mandryk et al., (2006) used skin conductance, heart rate variability, electromyogram (to measure jaw-clenching), and respiration sensors to measure the overall emotional dispositions of players interacting with a video-game. The authors identified correlations between signal features calculated using the entire recorded signal for each episode of game-play (lasting 5 minutes) and the players' post-episode subjective ratings of default experience labels such as 'fun', 'challenge', 'boredom', and 'frustration'.

Although, as we mentioned above, there have been various attempts to use the OCC theory in affective user modeling, psychological theories have more often been used in models that direct the affective behaviour of agents such as virtual humans (e.g., ALMA (Gebhard, 2005), Émile (Gratch, 2000), and the model produced by Dias & Paiva (2005)). Some of these computational models of affect have the potential to transfer to modeling the affect of users. The authors of FLAME (Seif El-Nasr et al., 2000) say that their affective framework could support a user model once it incorporates additional factors such as individual differences. In addition, (Gratch & Marsella, 2004) consider extending their computational framework of appraisal & coping, EMA, by modeling the link from appraisal to bodily expression as future work. Finally, (Elliot et al., 1999) discuss how the Affective Reasoner, a rule-based framework to build agents that respond emotionally, could also be used to model users' affect.

8 Discussion and Conclusions

In this paper, we presented and evaluated an affective user model designed to detect multiple individual emotions of players interacting with Prime Climb, an educational game for number factorization. The model is to be used by an intelligent pedagogical agent that attempts to improve how students learn from the game while still maintaining the high level of positive emotional engagement that is one of the key assets of game-based education. The model relies on a general framework for affective modeling that tackles the high level of uncertainty in emotion recognition by probabilistically combining information on both causes and effects of users' emotional reactions (Conati, 2002). While approaches to combining diagnostic and predictive inference have received substantial attention from researchers interested in affective user modeling (e.g., Bosma & André, 2004; Hudlicka & McNeese, 2002; Li & Ji, 2003; Prendinger et al., 2005), to our knowledge ours is the first attempt to provide a detailed evaluation of this technique. Furthermore, ours is one of the few affective models targeting the recognition of multiple emotions. Most existing models focus on assessing measures of affective valence and arousal, or individual emotions such as anxiety, frustration, interest or stress. While these restricted modeling tasks are appropriate in certain circumstances and applications, environments like educational games tend to trigger multiple, possibly overlapping and rapidly changing emotions. We argue that recognizing these emotions can improve the effectiveness of a pedagogical agent for game-based learning, because it can improve the precision of the agent's interventions.

In the paper, we illustrated how we incrementally built the model via repeated cycles of design and evaluation. The model's foundations lie in a well known emotion theory, the OCC model of cognitive appraisal (Ortony et al., 1988). The details of the implementation have been based as much as possible on data from real users. Because of the model's complexity, collecting reliable data for model construction was an extremely laborious process requiring several user studies. In the paper, we have summarized some of these studies, along with their outcomes and limitations, to give the reader a sense of the scope of the work and the challenges it entailed. We have not always been able to overcome these challenges at best, and thus our resulting model has shortcomings that affects its accuracy. Still, we believe that our results are both very promising and informative for the future development of this and other research in affective user modeling.

We evaluated our model on four of the six emotions that it can assess: joy or distress toward the game; admiration or reproach towards the agent. We showed that the predictive part of the affective model alone can already achieve good accuracy on the mutually exclusive emotions towards the game. Accuracy for *Joy* was 65%, for *Distress* was 73%, and the combined accuracy of 69% was statistically significantly better than the baseline accuracy of 50% achieved by always predicting the most likely emotion (*Joy*). Combined accuracy for emotions toward the agent is still significantly better than the baseline, but practically very close to it (54%). This is mostly due to the model's problems in detecting *Reproach* (48% accuracy), while *Admiration* reaches an accuracy of 61%. It should be noted that our accuracies for *Joy*, *Distress*, and *Admiration* are comparable to those achieved by (D'Mello et al., 2006) in recognizing the multiple emotions *confusion*, *eureka* and *frustration*.

An important aspect of our results is that they have been achieved via our complete model that includes an assessment of user goals, crucial for performing predictive inference based on the OCC theory. We have shown that model performance with goal assessment is comparable, if not better, to model performance when data on user goals is given to the model as evidence. Goal recognition is one of the hard problems in AI and thus one of the bottlenecks for a more widespread use of the OCC theory for affective user modeling. Ours is the only work that has shown with hard data the feasibility of this approach. Still, and not surprisingly, goal recognition is one of the limiting factors of our model's accuracy. In the paper, we discussed how the poor performance on *Reproach* is largely due to two goal-related model shortcomings: its inability to assess goals that dynamically change during the interaction, and the fact that we don't properly model goal priority in the presence of multiple goals.

In order to refine the model so as to remove these assumptions, we would first need to collect empirical data to understand why and how student goal priorities may change during game playing. Data on goal priorities could be recorded either via a self-report mechanism similar to the one we use to collect emotion self-reports, or by post-session annotations by experts. However both of these options have inherent difficulties. Asking students to identify their own goal priorities, even if asking about a reduced set of at most two goals, is likely to cause confusion as to what is being asked. For experts, it is likely that attempting to annotate student goals during the interaction would be rather difficult and laborious, given the novelty of the interaction and the fact that goals are often related to a non-trivial combination of factors including student personality, goals and interaction patterns. Thus, although this direction of investigation is possible, it also contains some very difficult challenges.

Instead, we explored adding diagnostic evidence for physiological sensors to our model, to overcome the limitations of its predictive component. We were able to confirm the link reported by others (e.g., Lang et al., 1993; Scheirer et al., 1999) between the mean of the EMG signal measured on the corrugator muscle and Prime Climb player emotions that could be clearly labelled with a negative valence. We were also able to show that this information could be used to significantly improve the model's predictive accuracy in cases where the students' affective state had a clear valence. Thus, our results provide initial support for the hypothesis that a model that combines information on both causes and effects of emotional reaction can compensate for the fact that often evidence on causes or effects alone is insufficient to accurately assess the student's emotional state. Although we found that our results did not immediately transfer to the assessment of ambiguously valenced feelings, this finding is not in contradiction with previous work showing the effectiveness of EMG for predicting valence. This work always investigated the mapping between EMG and clearly valenced emotions, while we deal with an environment in which students often experience multiple, possibly conflicting feelings, possibly expressed more subtly than the emotions induced in controlled laboratory settings. To accurately model these more difficult cases, it will be necessary to combine multiple sources of valence information, such as sensors for measuring heart rate, alternate EMG sensors or software to capture facial expressions. Our framework is already set up to support flexible combinations of sensors given the modularity of its diagnostic component (Conati, 2002).

In addition to EMG, we also explored the use of BVP to help assess affective valence, and SC to assess affective arousal. Unfortunately we obtained inconclusive results. The failure with the BVP signal was mostly due to lacking a sensor suitable for use with highly active children. Thus, we plan to repeat investigations on BVP by using a sensor less sensitive to movement. The failure with SC is most likely due to the restrictions imposed on our experimental protocol by the fact that we had limited time with our subjects. Time limitations prevented us from collecting a proper baseline measure for signal processing, and from eliciting proper self-reports of arousal from the students. Thus, our next step towards being able to use information from this sensor will be to design a study exclusively aimed at collecting data on affective arousal during interaction with Prime Climb.

Other future work relates to better addressing two other major challenges we encountered during model construction.

The first challenge relates to reliably recording students' affective states during game playing. As we mentioned in a previous section, in our research using judges to produce affective labels from video-recordings of the interaction is very difficult because of the requirement to distinguish separate feelings towards the game and towards the agent. We could not try to ask students to recall their feelings by viewing a replay of the interaction after game playing because of time constraints. Thus, we introduced the mechanism for obtaining emotions self-reports during game playing. While we have evidence that this mechanism is not overly intrusive on average, it does introduce an extraneous element in the interaction that may have unwarranted side effects for some students. Furthermore, it does not allow us to obtain data simultaneously on all of the emotion pairs we aim to assess and also on affective arousal. In the future, we plan to pilot test asking students to generate the self-reports after game playing, and, if they prove able to deal with the task, we will explore ways to extend our study sessions to include this alternative method for collecting affective labels.

The second challenge relates to collecting data-points for negative emotions. Throughout our studies, students have generated far fewer reports of negative emotions (*Distress* or *Reproach*) than positive emotions. One suggested reason for the small number of negative reports is the nature of our test-bed application, i.e. Prime Climb does not generate negative emotions very often, at least in the short playing time involved in our studies. We could induce these emotions on purpose during game-playing, but we argue that the students' negative feelings in this case would not be indicative of the real emotions that they would experience during real interactions. Possible solutions that we are planning to explore to overcome this challenge include: (i) find ways to have students interact longer with the game; (ii) have students play with each other. This second solution has the double advantage of giving us twice as much data for each playing session, and being likely to generate more and stronger emotional episodes, given what we have seen when students play together. However, it requires that we add to the model the capability of assessing emotions toward a partner, which is one of the next steps of this research

But the most important challenge that we need to address in the development of this research is to prove that having a sophisticated model for the assessment for players' individual emotions is worth the effort. That is, we need to show that the Prime Climb pedagogical agent can indeed benefit from having detailed information on user affect. Although our model currently underestimates the student's feelings of *Reproach*, its accuracy in assessing *Joy, Distress* and *Admiration* is high enough for us to consider an indirect evaluation to determine whether the model as it is would contribute to the pedagogical effectiveness of Prime Climb. We have begun to investigate ways to combine the assessment of the affective model and the model of student knowledge (Manske & Conati, 2005) into a decision theoretic framework that will allow the Prime Climb pedagogical agent to decide how to intervene so as to maximize the trade-off between student learning and engagement. Once we have completed this task, we can compare the overall effectiveness of the pedagogical agent with and without affective assessments. Once we have improved the accuracy for identifying feelings of *Reproach*, we can also run ablation studies to test our assertion that the more detailed information the agent has on the student affect, the better it can help the student interacting effectively with Prime Climb.

References

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z.: 2004, Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System"*, ACM CHI'04: Conference on Human Factors in Computing systems*, Vienna, Austria.

Ball, G., & Breese, J.: 1999, Modeling the Emotional state of Computer Users*, Workshop on 'Attitude, Personality and Emotions in User Adapted Interaction', UM'99*, Banff, Canada.

Bosma, W., & André, E. (2004). Exploiting Emotions to Disambiguate Dialogue Acts. *IUI'04, International Conference on Intelligent User Interfaces*.

Bradley, M. M., & Lang, P. J.: 1994, The Self-assessment Manikin and the Semantic Differential, *Journal of Behavior Therapy and Experimental Psychiatry,* **25**, 49-59.

Cacioppo, J. T., Klein, D. J., Berntson, G. G., & Hatfield, E.:1993, The Psychophysiology of Emotion, In R. Lewis & J. Havliand (Eds.), *The Handbook of Emotions*. New York: Guildford Press, pp. 119-142.

Chalfoun, P., Haffar, S., & Frasson, C.: 2006, Predicting the Emotional Reaction of the Learner with a Machine Learning Technique*, Workshop on Motivaional and Affective Issues in ITS, ITS'06, International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan.

Conati, C.: 2002, Probabilistic Assessment of Users' Emotions in Educational Games, *Journal of Applied Artificial Intelligence, special issue on "Merging Cognition and Affect in HCI",* **16**(7-8), 555-575.

Conati, C.: 2004, How to Evaluate Models of User Affect?*, ADS'04, Tutorial and Research Workshop on Affective Dialog Systems*, Kloster Irsee, Germany.

Conati, C., Chabbal, R., & Maclaren, H.: 2003, A Study on Using Biometric Sensors for Detecting User Emotions in Educational Games*, Workshop on "Assessing and Adapting to User Attitude and Affects: Why, When, and How?" in conjunction with UM'03, 9th International Conference on User Modeling*, Pittsburgh, USA.

Conati, C., & Klawe, M.:2002, Socially Intelligent Agents in Educational Games, In K. Dautenhahn, A. Bond, D. Canamero & B. Edmonds (Eds.), *Socially Intelligent Agents - Creating Relationships with Computers and Robots*: Kluwer Academic Publishers, pp. 213-220.

Conati, C., & Maclaren, H.: 2004, Evaluating a Probabilistic Model of Student Affect*, ITS'04, International Conference on Intelligent Tutoring Systems*, Maceio, Brazil.

Conati, C., & Zhao, X.: 2004, Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectivenesss of an Education Game*, IUI'04, International Conference on Intelligent User Interfaces*, Funchal, Madeira, Portugal.

Cordova, D., & Lepper, M.: 1996, Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice, *Journal of Educational Psychology,* **88**, 715-730.

Costa, P. T., & McCrae, R. R.: 1992, Four Ways Five Factors are Basic, *Personality and Individual Differences,* **13**, 653-665.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B.: 2004, Affect and Learning: An Exploratory Look into the Role of Affect in Learning with AutoTutor, *Journal of Educational Media,* **29**, 241-250.

Dawson, M. E., Schell, A. M., & Filion, D. L.:2000, The Electrodermal System, In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of Psychophysiology* (2 ed.). Cambridge, UK: Cambridge University Press, pp. 200-223.

Dean, T., & Kanazawa, K.: 1989, A Model for Reasoning about Persistance and Causation, *Computational Intelligence,* **5**(3), 142-150.

deVincente, A., & Pain, H.: 1999, Motivation Self-Report in ITS*, AIED'99, 9th International Conference on Artificial Intelligence in Education*, Le Mans, France.

Dias, J., & Paiva, A.: 2005, Feeling and Reasoning: A Computational Model for Emotional Characters*, Progress in Artificial Intelligence, 12th Portuguese Conference on Artificial Intelligence, EPIA 2005*, Covilhã, Portugal.

D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C.: 2006, Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialoguee, *International Journal of Artificial Intelligence in Education,* **16**, 3-28.

D'Mello, S. K., & Graesser, A. C.: 2006, Affect Detection from Human-Computer Dialogue with an Intelligent Tutoring System*, IVA 2006, 6th International Conference on Intelligent Virtual Agents*, Marina del Rey, CA, USA.

Elliot, C., Rickel, J., & Lester, J. C.:1999, Lifelike Pedagogical Agents and Affective Computing: An Exploratory Synthesis, In M. Wooldridge & M. Veloso, M. (Eds.), *Artificial Intelligence Today*: Springer, pp. 195-211.

Fisher, R. A.: 1935, The Design of Experiment, New York: Hafner.

Gebhard, P.: 2005, ALMA - A Layered Model of Affect*, 4th International Joint Conference of Autonomous Agents & Multi-Agent Systems (AAMAS'05)*, Utrecht, The Netherlands.

Gratch, J.: 2000, Emile: Marshalling Passions in Training and Education*, 4th International Conference on Autonomous Agents*, Barcelona, Spain.

Gratch, J., & Marsella, S.: 2004, A Domain Independent Framework for Modeling Emotion, *Journal of Cognitive Systems Research,* **5**(4), 269-306.

Graziano, W. G., Jensen-Campbell, L. A., & Finch, J. F.: 1997, The Self as a Mediator Between Personality and Adjustment, *Journal of Personality and Social Psychology,* **73**, 392-404.

Guinn, C., & Hubal, R.: 2003, Extracting Emotional Information from the Text of Spoken Dialogue*, Workshop on "Assessing and Adapting to User Attitude and Affects: Why, When, and How?" in conjunction with UM'03, 9th International Conference on User Modeling*, Pittsburgh, PA.

Healey, J. A., & Picard, R. W.: 2005, Detecting Stress During Real-World Driving Tasks Using Physiological Sensors, *IEEE Transactions on Intelligent Transportation Systems,* **6**(2), 156-166.

Heckerman, D.:1999, A Tutorial on Learning with Bayesian Networks, In M. Jordan (Ed.), *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Hudlicka, E., & McNeese, M.: 2002, Assessment of User Affective and Belief States for Interface Adaptation: Application to an Air Force Pilot, *User Modeling and User Adapted Interaction,* **12**(1), 1-47.

Jain, A., & Zongker, D.: 1997, Feature Selection: Evaluation, Application, and Small Sample Performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **19**(2), 153-158.

Kapoor, A., & Picard, R. W.: 2005, Multimodal Affect Recognition in Learning Environments*, 13th Annual ACM International Conference on Multimedia*, Singapore.

Kim, J., & André, E.: 2006, Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation*, PIT'06, Perception and Interactive Techologies*, Kloster Irsee, Germany.

Klawe, M.: 1998, When Does The Use Of Computer Games And Other Interactive Multimedia Software Help Students Learn Mathematics?*, Technology and NCTM Standards 2000 Conference*, Arlington, VA.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O.: 1993, Look at Pictures: Affective, Facial, Visceral, and Behavioral Reactions, *Psychophysiology,* **30**, 261-273.

Lepper, M., Woolverton, M., Mumme, D., & Gurtner, J.-L.:1993, Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-based Tutors, In S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools*: Lawrence Erlbaum Associates, NJ.

Li, X., & Ji, Q.: 2003, Active Affective State Detection and User Assistance*, Workshop on "Assessing and Adapting to User Attitude and Affects: Why, When, and How?" in conjunction with UM'03, 9th International Conference on User Modeling*, Pittsburgh, PA.

Linnenbrink, E. A., & Pintrich, P. R.:2002, The Role of Motivational Beliefs in Conceptual Change, In M. Limon & L. Mason (Eds.), *Reconsidering Conceptual Change: Issues in Theory and Practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 115-135.

Lisetti, C. L., & Nasoz, F.: 2004, Using Non-invasive Wearable Computers to Recognize Human Emotions from Physiological Signals, *EURASIP Journal on Applied Signal Processing,* **11**, 1672-1687.

Litman, D., & Forbes-Riley, K.: 2006, Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogues with both Human and Computer Tutors, *Speech Communication,* **48**(5), 559-590.

Mandryk, R. L., Inkpen, K. M., & Calvert, T. W.: 2006, Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologes, *Journal of Behavior and Information Technology (Special Issue on User Experience),* **25**, 141-158.

Manske, M., & Conati, C.: 2005, Modeling Learning in Educational Games*, AIED'05, 12th International Conference on AI in Education*, Amsterdam.

Mitchell, T.: 1997, Machine Learning: McGraw Hill.

Ortony, A., Clore, G. L., & Collins, A.: 1988, The Cognitive Stucture of Emotions: Cambridge University Press.

Papillo, J., & Shapiro, D.:1990, The Cardiovascular System, In L. G. Tassinary & J. T. Cacioppo (Eds.), *Principles of Psychophysiology: Physical, Social, and Inferential Elements*: Cambridge University Press, pp. 456-512.

Partala, T., & Surakka, V.: 2004, The Effects of Affective Interactions in Human-Computer Interaction, *Interacting with Computers,* **16**(2), 295-309.

Peter, C., & Herbon, A.: 2006, Emotion Representation and Physiology Assignments in Digital Systems, *Interacting with Computers,* **18**(2), 139-170.

Picard, R. W., Vyzas, E., & Healey, J. A.: 2001, Toward Machine Emotional Intelligence: Analysis of Affective Physiological State, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **23**(10), 1175-1191.

Prendinger, H., Mori, J., & Ishizuka, M.: 2005, Recognizing, Modeling, and Responding to Users' Affective States*, UM'05 10th International Conference on User Modeling*, Edinburgh.

Qu, L., & Johnson, L.: 2005, Detecting the Learner's Motivational States in an Interactive Learning Environment*, AIED'05 12th International Conference on Artificial Intelligence in Education.*, Amsterdam.

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W.: 2002, Frustrating the User on Purpose: A Step Toward Building an Affective Computer, *Interacting with Computers,* **14**(2), 93-118.

Scheirer, J., Fernandez, R., & Picard, R. W.: 1999, Expression Glasses: A Wearable Device for Facial Expression Recognition*, CHI'99, Human Factors in Computer Systems*, Pittsburgh, PA.

Seif El-Nasr, M., Yen, J., & Ioerger, T. R.: 2000, FLAME - Fuzzy Logic Adaptive Model of Emotions, *Autonomous Agents and Multi-Agent Systems,* **3**, 219-257.

Streit, M., Batliner, A., & Portele, T.: 2004, Cognitive-Model-Based Interpretation of Emotions in a Multi-Model Dialog System*, ADS'04 Tutorial and Research Workshop on Affective Dialogue Systems*, Kloster Irsee, Germany.

Vyzas, E., & Picard, R. W.: 1998, Affective Pattern Classification, *AAAI 1998 Fall Symposium, Emotional and Intelligent: The Tangled Knot of Cognition*, 176-182.

Zhou, X., & Conati, C.: 2003, Inferring User Goals from Personality and Behavior in a Causal Model of User Affect*, IUI'03, International Conference on Intelligent User Interfaces*, Miami, FL.