How to Evaluate Models of User Affect?

Cristina Conati

Computer Science Department, University of British Columbia, 2366 Main Mall
Vancouver, BC, V6T 1Z4, BC, Canada
conati@cs.ubc.ca

Abstract. Modeling user affect is becoming increasingly important for intelligent interfaces and agents that aim to establish believable interactions with their users. However, evaluating the accuracy and effectiveness of affective user models is extremely challenging because of the many unknowns in affect comprehension. In this paper, we overview existing approaches related to the validation of affective user models, and we describe our own experience with an approach for direct model evaluation that we have used in a recent study.

1 Introduction

Recent years have seen a flourishing of research directed to add an affective component to human-computer dialogue. A key element of this endeavor is the capability to recognize user emotional states during the interaction, i.e., to build a model of the user's affect. Building such models can be extremely challenging because it requires formalizing and applying strategies for emotion recognition that even human beings sometimes cannot generate or apply successfully. But validating the models can also be extremely hard because, in addition to the challenges common to any user model evaluation, validating affective user models suffers from the difficulty of obtaining reliable measures of user affect against which to compare the model predictions. Furthermore, because the research field is rather new, there are very few complete applications that include an affective user model and that can be used to test the model indirectly through evaluation of the application itself.

In this paper, we address the problem of how to validate affective user models for improving human-computer dialogue. We start with an overview of the available techniques. We then describe our experience in using one of these techniques to evaluate an affective user model for the interaction with an educational game.

2 Overview on Techniques for Validating Affective User Models

Most of the empirical work in affective user modeling has so far been directed to assessing possible sources of affective data, *before* building a complete user model that can use these sources. In particular, researchers have been trying to identify reliable ways to recognize symptoms of emotional reactions, ranging from observable changes in facial expressions, posture and intonation, to variations in lower level measures of emotional arousal, such as skin conductance and heartbeat.

The standard technique applied in this line of work is *emotion induction*. Stimuli are devised to induce specific emotions in a set of subjects. Then, sensors are used to detect changes in behavioral expressions that are known to be influenced by these emotions. Finally, the sensors' reliability in diagnosing emotional states from these measures of behavioral changes is evaluated. Techniques to obtain the desired emotions from the subjects include: (1) using a professional actor as a subject, who expresses the relevant emotions on demand [14]; (2) using standard stimuli for emotion elicitation that are not necessarily related to the task the affective user model will eventually be designed for, but that are known to be very reliable in generating the desired emotional states (e.g. movie clips [12]); (3) designing Wizard of Oz studies to elicit specific emotions in the context of the interaction that the affective user model will eventually support. So far, this approach has been mainly used to test frustration detection (e.g., [2],[11]).

The main advantage of these emotion eliciting techniques is that they provide a reliable base line against which to test sources of emotion data. The main disadvantage is that it may be difficult to generalize the reliability of data sources tested with these techniques to real interactions, because the user's affective reactions may not be as intense, well defined and isolated as they are during the elicitation studies.

In contrast with substantial research on validating sources of affective data, there has been little work on evaluating complete affective user models. 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 behavior to a user affect. To our knowledge, so far there have been only two informal *indirect* evaluations of affective user models, because there are very few complete applications that include an affective user model. The first evaluation used a sample set of simulated users and scenarios as a preliminary validation of a proof-of-concept prototype that adapts to anxiety in compact pilots [10]. The second evaluation included two field studies of the Avatalk architecture, designed to detect affect in speech. In both studies, Avatalk was used as a training aid for users that had to learn how to convey specific affective states through speech as part of their job. The studies focused on system acceptance, and provided no results on Avatalk effectiveness as a training tool [8]. Furthermore, the authors recognize that these types of *macro studies* do not allow assessing the contribution of the affective user model to system performance, because of the confounding variables introduced by the other components that define the system's interactive dialogue (e.g., usage of synthetic characters to deliver instruction, character decision making).

The *direct* approach to model validation overcomes two of the main shortcomings of the indirect approach. First, this approach does not require having a complete system, as the interaction can be carried out via a Wizard of Oz set up. Second, a direct evaluation can provide a deeper understanding of the model behavior that is not confounded by other aspects of the application. However, the main challenge of direct evaluation 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. In the rest of the paper, we describe our experience in using this approach to evaluate a model of user affect during the interaction with an educational game, Prime Climb. To our knowledge, this is the first direct evaluation of an affective user model with real users. Gratch and Marsella [7] discuss a direct evaluation where an appraisal model is run over an

evolving situation taken from a psychological instrument and then compared to subjects responses to the same instrument. However, because data was aggregated across subjects, this study does not assess the model ability to model individual differences.

3 Validating a Model of User Affect in Educational Games

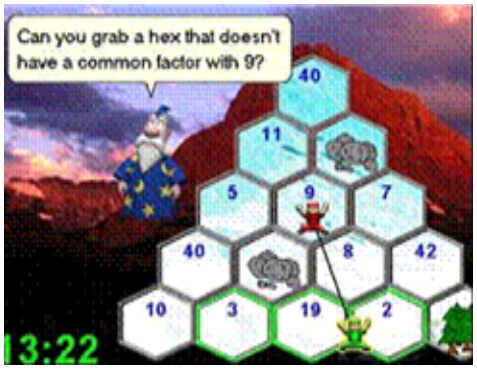

Figure 1: The Prime Climb Interface

The Prime Climb educational game has been designed by the EGEMS group at the University of British Columbia to help 6th and 7th grade students practice number factorization. It is a two-player game in which the players need to cooperate to climb to the top of a series of mountains, divided in numbered sectors (see Figure 1). Each player can only move to a sector that does not share any factors with the sector occupied by her partner, otherwise she falls. As for several other educational games, Prime Crime has been shown to be very engaging, but to have limited pedagogical effectiveness because many students do not have the learning skills necessary to benefit from this highly unstructured and easily distracting pedagogical interaction.

The long term research goal is to have a model of both student affect and knowledge that an intelligent pedagogical agent can use to balance student learning and engagement during the interaction with the game. Because of the complexity of the modeling task, we are building the affective and learning models separately, to pin down the factors that independently contribute to each assessment before proceeding to model the relevant synergies. We started to evaluate the model of student learning with the *indirect approach*, after building a pedagogical agent that uses the model to provide hints aimed at improving student learning [5]. However, we felt that we could not use the same approach to evaluate the affective user model. Because we still do not have a good understanding of how to build an agent that uses the affective user model to maintain student engagement in Prime Climb, there are too many aspects extraneous to the model that would cloud the interpretation of the indirect evaluation results in terms of model accuracy. Thus, we decided to try and evaluate the model directly, *before* building the agent that could use it. Before describing the evaluation methodology, we briefly illustrate our affective user model. More details can be found in [3] and [15].

3.1 The Affective User Model

In contrast with other affective user models that assess one specific emotion (e.g., [10]), or measures of valence and arousal (e.g., [1],[2]) our model assesses multiple specific emotions that we observed to influence the interaction with Prime Climb during pilot studies on the game. These are six of the 22 affective states described in the OCC

cognitive theory of emotions [13]: *joy*/*distress* for the current state of the game (i.e., the outcome of a student or agent action), *pride*/*shame* of the student toward herself, and *admiration*/*reproach* toward the agent.

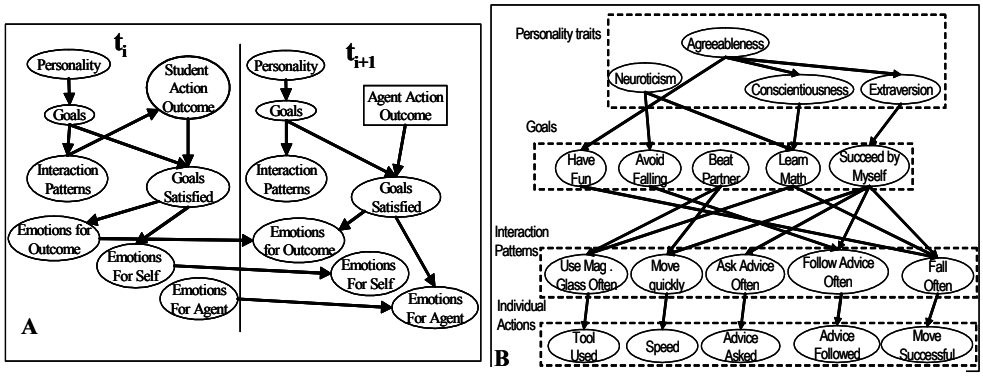

Figure 2: Two time slices of the affective model (A); sub network to assess goals (B).

The model relies on a Dynamic Bayesian Network (DBN) to probabilistically assess user emotional states from possible causes of emotional arousal, as described in the OCC theory of emotions. In this theory, emotions arise as a result of one's *appraisal* of the current situation in relation to one's goals. Thus, our DBN (see two sample slices in Figure 2A) includes variables for possible user's *Goals* when playing with Prime Climb, and for situations consisting of the outcome of any user or agent action (nodes *User Action Outcome* and *Agent Action Outcome* in Figure 2A). The desirability of an action outcome in relation to the user's goals is represented by the node class *Goals Satisfied.* This in turn influences the user's emotional states, represented in the DBN by three binary nodes (*Emotions for Outcome*, *Emotions for Self*, *Emotions for Agent*) that model the six emotions mentioned above.

User goals are a key element of the OCC model, but assessing these goals is non trivial, especially when asking the user directly is not an option (as is the case in educational games). Thus, our DBN also includes nodes to infer user goals from indirect evidence coming from both user's *Personality* [6] and *Interaction Patterns*. Because all the variables in this sub network are observable, we identified the variables and built the corresponding conditional probability tables (CPTs) using data collected through a Wizard of Oz study during which an experimenter guided the pedagogical agent. In these studies, students took a personality test based on the Five Factor personality theory [6]. After game playing, students filled a questionnaire on what goals they had during the interaction. The probabilistic dependencies among goals, personality, interaction patterns and student actions were established through correlation analysis between the tests results, the questionnaire results and student actions in the log files recorded during the interactions [15]. Figure 2B shows the resulting sub-network.

In the sub-network that represents the appraisal mechanism, the links and CPTs between *Goal* nodes, the outcome of student or agent actions and the *Goal Satisfied* nodes, are currently based on our subjective judgment. Some of these links are quite obvious, i.e., if the student has the goal *Avoid Falling*, a move resulting in a fall will lower the probability that the goal is achieved. Other links (e.g., those modeling which actions cause a student to have fun or learn math) are less obvious, and could be built only

through student interviews that we could not include in our studies. When we did not have good heuristics to create these links, we did not include them in the model.

The links between *Goal Satisfied* and emotion nodes are defined as follows. Because the outcome of every agent or student action is subject to student appraisal, every *Goal Satisfied* node influences *Emotions for Outcome* in any given slice (see Figure 2A, both slices). Whether a *Goal Satisfied* node influences *Emotions for Self* or *Emotions for Agent* in a given slice depends upon whether it was the student (slice t_i in figure 2A) or the agent (slice t_{i+1} in figure 2A) who caused the current game state. The CPTs of emotion nodes given goal-satisfied nodes are defined so that the probability of the positive emotion is proportional to the number of *Goal Satisfied* nodes that are *true.*

3.2 Model Evaluation

A direct evaluation of our affective user model requires ascertaining the actual emotions that students experienced during the interaction with Prime Climb. However, because these emotions tend to be ephemeral and can change multiple times during the interaction, it is unrealistic to expect that, after game playing, subjects remember the affective states they went through. A technique that is often used to help subjects recollect volatile states is to show them a video of their interaction. However, we could not use this approach both because we could not keep our subjects for the required additional time and, more importantly, because it is highly unlikely that our 10 and 11 year old students would be willing or able to undergo this procedure. Having an experimenter code the subjects' emotional states from a video of the interaction was also unlikely to yield reliable results. When we tried to use this technique in another Prime Climb study to test biometric sensors for emotion detection [4], our video recordings showed that users' visible bodily expressions often did not give enough indication of their specific emotions, although they were more reliable for detecting valence and arousal.

Given the above factors, we decided to devise a strategy to obtain the information on their emotions directly from our subjects *during* the interaction. However, this approach is chancy because, if not done properly, it can significantly interfere with the very emotional states we want to assess. Furthermore, subjects' self-reports can be unreliable both because of a well known tendency that subjects have to give artificially positive answers out of politeness, and because some subjects may not be able to assess their emotional states. Both these phenomena will have to be taken into account when using self-reports in empirical studies. Nonetheless, for our specific type of application, they seem to be the least noisy source of information.

In the rest of the paper, we first describe our approach to design an interface that can elicit emotion self-reports during Prime Climb playing as unobtrusively as possible, given the constraints imposed by this type of interaction. We then discuss some general methodological suggestions to deal with the potential unreliability of emotions self-reports in testing affective user models.

4 Pilot Study

To find an unobtrusive way to obtain emotion self-reports during the interaction with Prime Climb, we ran a pilot study with 6th grade students in a Vancouver school. The study tested two alternative interfaces. One interface had an emotion-report dialog box permanently present on the side of the game window (see Figures 3 and 4). This interface is quite unobtrusive but may not generate much data because it relies on the student's willingness to volunteer self-reports. The second interface included both the permanent dialog box, as well as the same dialogue box that would pop up whenever either one of the following conditions were satisfied: (1) the student had not submitted an emotion self-report for a period of time longer than a set threshold or (2) the underlying affective model was detecting a relevant change (also based on a set threshold) in its belief of the student's emotional state. Students are required to submit a self-report when the pop-up dialogue box appears.

Figure 3: Emotion-reporting dialogue box

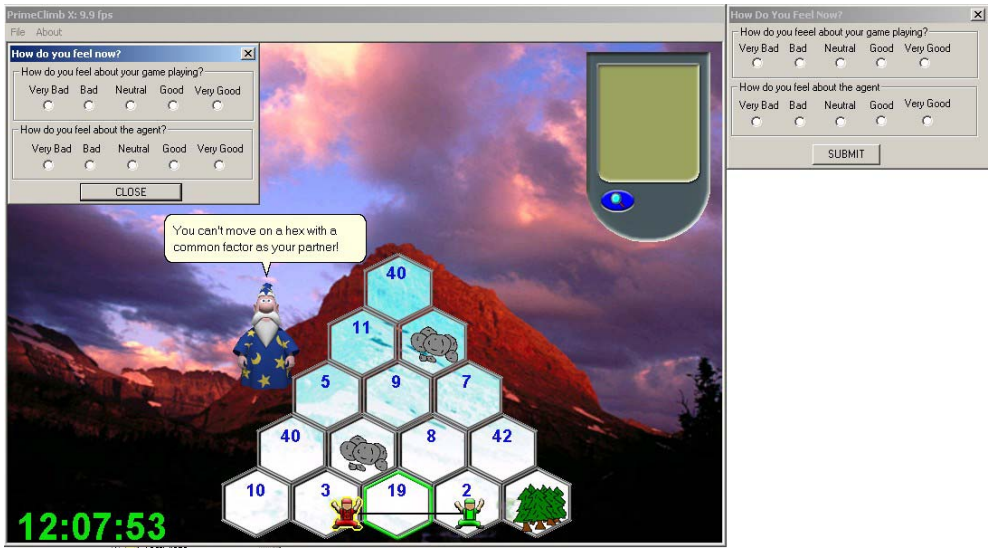

Figure 4: Interface with both the permanent and pop-up dialogue box

The questions that we were trying to explore with this pilot study were the following:

1. Do students volunteer self-reports in the permanent dialogue box frequently enough to provide sufficient data for model evaluation?
2. If not, and we must resort to the pop-up dialogue box, how do students tolerate it?
3. If we need the pop-up box, is it worth keeping the permanent box for those students who still want to volunteer affective information?

As Figure 3 shows, the emotion dialogue box only elicited information on two of the three sets of emotions targeted by our model, both because it was felt that dealing with

three different emotional states would be too confusing for our subjects, and because teachers suggested that students would have more problems in reporting emotions toward themselves than toward the game or the agent. Not having explicit information of *pride/shame* was not a serious limitation because we could still derive information on these emotions from the information obtained on the other two pairs.

We are aware that directly asking the students about their feelings may feel unnatural and perhaps too pushy. A more discreet approach that is often used in emotional psychology is to ask about factors that are antecedents of emotions, such as goals and expectations. However, getting a reliable assessment of students emotions through this indirect approach would require too many questions to be acceptable in a real time interaction as fast paced as that with Prime Climb.

The study set up was as follows: students were told that they would be playing a game with a computer-based agent that was trying to understand their needs and giving help accordingly. Therefore, the students were encouraged to provide their feelings whenever there was change in their emotions so that the agent could adapt its behavior. In reality, the agent was directed by an experimenter through a Wizard of Oz interface. A key difference between our study and previous studies that used the Wizard of Oz paradigm (e.g., [2],[11]) is that our experimenter did not try to elicit specific, intense emotions through extreme behavior, because we wanted the results of the study to be generalizable to normal interactions in which these extreme behaviors are hopefully the exception and not the rule. Thus, the experimenter was instructed to provide help anytime the student showed difficulty with the climbing task and factorization knowledge, to resemble the behavior of the pedagogical agent for student learning that at the time of the study was still under development. Help could be provided at different levels of detail (from a general suggestion to think about the factorization of the numbers involved in an incorrect move, to telling the student where to move). In the condition with the pop-up dialogue box, all the experimenter's hints and student's actions were captured by the affective user model, and the game would pop-up the additional dialogue box following the criterion described earlier. For logistic constraints, the experimenter had to act as a companion in the climbing task, but he also directed the pedagogical agent in a Wizard of Oz fashion. This could be done because the student could not see the experimenter's screen (see Figure 5 for a similar study setup).

All together, 10 students participated in the study. 4 students used the version of the game with the permanent dialogue box only (PDB), 6 used the version including the pop-up-box (PDB+POPUP)[1]. After 10-15 minutes of game playing, the students completed a questionnaire targeting the questions on interface acceptance described earlier (see Table 2). We also recorded how often students volunteered information in the emotion dialogue boxes (see Table 1).

The "# per student" column in Table 1 reports for each interface how many self-reports each student volunteered in the permanent dialogue box. The numbers show that some students tend to volunteer few self-reports, suggesting that the permanent dialogue box by itself may not consistently generate sufficient data for model testing. On the other hand, the numbers for volunteered self-reports in the PDB-POPUP row also shows that some students continued to volunteer information in the permanent dialogue

[1] For technical reasons we could not divide students equally between the two groups.

box even if they knew that the system would ask for the information explicitly when it needed.

Table 1: Number of self-reports generated in the two interfaces

	Volunteered self-reports		Reports in pop-up box	
Group	**# per student**	**Mean**	**# per student**	**Mean**
PDB only	1,2,4,10	4.25	NA	NA
PDB-POPUP	0,3,3,4,7,12	4.48	16,20,26,18,16,36	21.6

Thus, this pilot study suggests that we do need the pop-up box to proactively elicit self-reports, although it is worth keeping the permanent box around for those students who tend to volunteer information. We then checked the post questionnaire for students in the PDB-POPUP group to see how they tolerated the pop-up dialogue box. Table 3 reports the results. Students' answers are on a Likert scale from 1 (strongly disagree) to 5 (strongly agree), where 3 represents indifference.

Table 2: Average scores for post questionnaire items in the pilot study; scores are on a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

The popup dialog box showed up too frequently.	4.2
The popup dialog box interfered with my game playing.	3.2
The permanent dialog box interfered with my game playing.	2.56
The questions in the dialog box were clear to me.	4.11
It bothered me having to tell the system how I feel.	2.56

The averages in Table 2 suggest that, although students on average found that the pop-up box showed up too frequently, it did not seem to interfere too much with game playing. This gave us confidence that, by adjusting the pop-up frequency, we could have an interface students can live with and that provides enough data for model evaluation.

Another encouraging result from the questionnaire is that students didn't seem to mind having to express their emotions to the system. This is important to reduce the possibility that the mere act of having to express their emotions is upsetting for students, regardless of the available self-reporting mechanism. The questionnaire also asked for suggestions on how to improve the dialogue boxes or the emotion input mechanism in general, but we did not get any relevant answers.

Given the results of the pilot study, we decided to run the empirical evaluation of our model by using the interface with both dialogue boxes. However, we adjusted the algorithm that manages the pop-up box so that the box would appear less frequently.

4. Using the Two-box Interface for Model Evaluation

Twenty 7th grade students from a local Vancouver school participated in the study to evaluate the Prime Climb affective model. The study set up was exactly the same as for the pilot study described earlier, except that all the students used the two-box interface, and we had two sessions running in parallel, with two students playing with two experimenters who also directed the agent via the Wizard of Oz interface (See Figure 5). All

Figure 5: Study set up

students' inputs in the dialogue boxes were collected for comparison with the corresponding model assessments. After game playing, students filled out the same questionnaire on interface acceptance that was used in the previous study. In this section, we discuss whether the two-dialogue boxes set up matched our expectations as a technique for model testing. In particular, we focus on how to address a set of questions that are fundamental to ascertain the reliability of any technique that relies on self-reports for direct evaluation of an affective user model.

(1) What was the user acceptance of the interface to elicit emotion self-reports?

(2) If there are subjects annoyed by this interface, do we have to discard their data?

(3) How reliable are the self-reports elicited through this interface?

What is the acceptance of the interface for emotion self-report?

Table 3 reports the average scores on each post-questionnaire item. Although some students are still bothered by the pop-up dialogue box, the average level of declared interference is fairly low. As in the first study, acceptance of the permanent dialogue box is very high, students did not seem to mind telling the system how they felt, and were pretty clear about what the dialogue boxes were asking.

Table 3: average scores for questionnaire items in the second study

	Question type	Mean	StDev
C	The popup dialog box showed up too frequently.	3.4	1.5
D	The popup dialog box interfered with my game playing.	2.8	1.4
E	The permanent dialog box interfered with my game playing.	1.9	1.4
F	The questions in the dialog box were clear to me.	4.5	0.9
G	It bothered me having to tell the system how I feel.	2.1	1.1

What to do with subjects who were annoyed by the interface?

The questionnaire results showed that 10 students gave a rating higher than 3 to the question asking whether the pop-up box showed up too frequently (question C in Table 3), or to the question asking whether the dialogue box interfered with game playing (question D). Because the ratings indicate that these students were somewhat annoyed by the popup box, what should we do with their data? If they were truly upset by the pop-up box, perhaps we should discard their self-reports when computing the model's accuracy. The model would never be able to detect their negative affect since it does not take into account the appearing of the dialogue box in its assessment (recall from Section 3.1 that the model was not built using data from emotion self-reports). More specifically, the model would tend to underestimate the players' emotion toward the game (*Distress*) and perhaps toward the agent (*Reproach*).

To test whether this was the case, we computed the model accuracy in detecting *distress* and *reproach* for the subset of students who gave a score higher than 3 to questions C and D in the post-questionnaire (see Table 4 and Table 5). We then compared this accuracy with the accuracy for the students who did not report annoyance with the dia-

logue boxes. Model accuracy is computed as the fraction of the students' reported emotions that the model predicted correctly. The accuracy for students who were not annoyed with the dialogue box is 100% for *Distress* and 75% for *Reproach*. Table 4 and Table 5 show lower accuracies on *Distress* and *Reproach* for those students who reported annoyance with the dialogue box. Unfortunately we don't have sufficient data to make reliable conclusions based on these numbers. The difference between the two accuracies is not statistically significant. Thus, we cannot reject the null hypothesis that annoyance with the dialogue box did not affect the player attitude toward the game, and we have no basis to eliminate the self-reports of the annoyed subjects from the analysis of model accuracy. On the other hand, we are aware that lack of statistical significance could also be due to the limited number of subjects and their uneven distribution between the two groups.

Table 4: Analysis of *Distress* accuracy for students who reported annoyance with the dialogue box

Questionnaire Answers	Students	Students who reported *Distress*	# of *Distress* Datapoints	Correct predictions
4 or 5 to question D	X5,X6,X10 E2,E5,E6	X5, E2	4	3 (75%)
4 or 5 to question C	X1,X4,X5,X6 E3, E4, E5	X5	3	2 (66.7%)

Table 5: Analysis of *Reproach* accuracy for students who reported annoyance with the dialogue box

Questionnaire Answers	Students	Students who reported *Reproach*	# of *reproach* Datapoints	Correct predictions
4 or 5 to question D	X5,X6,X10 E2,E5,E6	0	0	---
4 or 5 to question C	X1,X4,X5,X6 E3, E4, E5	X1,E4	2	2 (100%)

What we can say, however, is that negative emotion reports were only a small fraction (4% for *Distress* and 2% for *Reproach*) of all the emotion reports generated by the 10 students who had declared annoyance with the dialogue box (9 reports on average for each emotion pair). Thus, if we can trust the students' self-reports, these results could be interpreted as an indication that annoyance with the dialogue box, in fact, *does not* always translate into annoyance with the game or the agent. This would be in itself a quite encouraging finding for researchers interested in evaluating affective user models, because it shows that subjects can tolerate to some extent the interference caused by the artifacts designed to elicit their emotions. We obviously need more data before we can draw any reliable conclusion on this issue. And because any such conclusion would have to rely on an analysis of the student emotion self-reports similar to the one discussed above, we also need to understand what the reliability of these self-reports is. This takes us to the third and final question in this discussion.

How reliable are students self-reports elicited through the interface?

All in all, in this study we had far fewer self reports of negative than of positive emotions. Of the 130 self-reports on the joy/distress pair, only 9 were for distress. Of the 103 reports on admiration/reproach pair, only 6 were for reproach. On the one hand, this could be taken as further evidence that user self-reports are unreliable because users often tend to give answers out of politeness. On the other hand, the reader should recall that in our study the experimenters were not trying to induce negative emotions, they were simply trying to provide help every time they thought a student was not learning well from the game. Our initial expectation was that these tutorial interventions would often be annoying because they would interfere with the game-like nature of the interaction. However, it may be that we underestimated the students' desire to learn from the game or the novelty effect of interacting with an animated pedagogical agent. Furthermore, students may not have encountered many situations in which the game itself became annoying or frustrating.

To gain a better understanding of how reliable our subjects' self-reports are, we looked at the log files to identify those situations that did generate negative reports from some of the subjects, to see if and how often they appeared in conjunction with positive reports. A preliminary analysis shows that only 1 student (X1 in Table 5), gave both a negative and a positive report in response to the same, potentially negative situation. In a single sequence where the student is not making climbing progress, X1 gave an admiration report of 5, followed by 1 (indicating reproach), followed again by 5. This is the only instance we have been able to find in our log files of potentially inconsistent self-reports, providing evidence toward the hypothesis that in our study we did not get many negative self-reports not because of the subjects tendency to please the experimenter, but because they mostly did not experience negative emotions. This does not prove, of course, that our interface for emotion self-reports is generally reliable. More testing should be done with interactions that *do* induce more negative emotions. However, this analysis shows how log data can be used as an alternative to (or in conjunction of) video data to integrate and validate specific sets of emotion self-reports.

5 Conclusions and Future Work

In this paper, we have addressed the problem of how to evaluate affective user models. Because affective user modeling is a relatively new research field, there is very little knowledge on how to best evaluate these models, especially if they try to assess a variety of specific emotions in fairly unconstrained interactions that tend to generate different affective reactions in different users. We have reviewed the techniques that can be used to validate an affective user model or the sources of data that it uses, and we have discussed the application of one of these techniques, direct model evaluation, to the validation of a model of student affect during the interaction with an educational game. Direct model evaluation is advantageous both because it does not require a complete system that uses the affective model and because it gives more precise information on model performance and the factors that influence it. However, it poses the challenge of obtaining a measure of the user's actual emotions during the interaction. We have illustrated a mechanism that we have devised to obtain this information as unobtrusively as possible, and we have presented an analysis aimed at understanding whether we have succeeded. In particular, we have shown how we tried to answer three questions that are key to defining the effectiveness of any mechanism for emotion self-report to directly evaluate

affective models: (1) how intrusive the mechanism turns out to be; (2) what to do with the data from subjects that do find the mechanism intrusive; (3) how to assess the reliability of the obtained self-reports. Our answers are preliminary, and currently limited to the specific application and user population involved in the study presented here. However, as research in affective modeling progresses, we hope that more and more of these answers will be provided through empirical model evaluations. This would help create a set of standards and guidelines that can streamline the evaluation process and allow researchers to adopt a specific evaluation method with a clear understanding of its possible sources of inaccuracies and related compensation strategies.

References

1. Ball, G. & Breese, J., Modeling the emotional state of computer users. Workshop on 'Attitude, Personality and Emotions in User-Adapted Interaction, Proc. of UM '99, Banff, Canada (1999).
2. Bosma, W. and André, E. Recognizing Emotions to Disambiguate Dialogue Acts. Proc. of IUI '04, Madeira, Portugal (2004).
3. Conati C., Probabilistic Assessment of User's Emotions in Educational Games, Journal of Applied Artificial Intelligence, vol. 16 (7-8), p. 555-575 (2002).
4. Conati, C., Chabbal R., and Maclaren, H., A study on using biometric sensors for monitoring user emotions in educational games. Proc. of Workshop on Modeling User Affect and Actions: Why, When and How. Proc. of UM' 03, , Pittsburgh, PA (2003)(*).
5. Conati C. and Zhao X. Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game. Proc. of IUI '04, Madeira, Portugal. (2004).
6. Costa, P.T. and R.R. McCrae. Four ways five factors are basic. Personality and Individual Differences 1. **13**: p. 653-665. (1992).
7. Gratch, J. and Marsella, S., Evaluating a General Model of Emotional Appraisal and Coping. Proc. of AAAI Spring Symposium '04 "Architectures for Modeling Emotions", (2004)
8. Guinn, C., and Hubal, R., Extracting Emotional Information from the Text of Spoken Dialogue, in (*) (2003).
9. Healy, J. and Picard, R. SmartCar: Detecting Driver Stress. in 15th Int. Conf. on Pattern Recognition. Barcelona, Spain (2000).
10. Hudlicka, E. & McNeese, M., Assessment of User Affective and Belief States for Interface Adaptation: Application to an Air Force Pilot Task. User Modeling and User Adapted Interaction, 12(1), 1-47, (2002).
11. Mori, J., Prendinger H., Mayer, S., Dohi and Ishizuka, M., Using Biosignals to track the effects of a character-based interface. in (*) (2003).
12. Nasoz. F., Lisetti, C., Alvarez, K., Finelstein, N., 2003 Emotion recognition from Physiological Signals for User Modeling of Affect. in (*) (2003).
13. Ortony, A., G.L. Clore, and A. Collins, The cognitive structure of emotions. Cambridge University Press, (1988).
14. Picard R.W., Vyzas E., & Healey J., Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), 1175-1191, (2001).
15. Zhou, X. and C. Conati. 2003. Inferring User Goals from Personality and Behavior in a Causal Model of User Affect. Proc. of IUI '03, Miami, FL.