Prediction of Student Achievement Goals and Emotion Valence during Interaction with Pedagogical Agents

Socially Interactive Agents Track

Sébastien Lallé
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
lalles@cs.ubc.ca

Cristina Conati
Department of Computer Science
University of British Columbia
Vancouver, BC, Canada
conati@cs.ubc.ca

Roger Azevedo
Department of Psychology
North Carolina State University
Raleigh, NC, USA
razeved@ncsu.edu

ABSTRACT

There is evidence that Pedagogical Agents (PA) can influence students' emotions while learning with Intelligent Tutoring Systems, and that this influence is modulated by the students' achievement goals for learning. This suggests that students may benefit from personalized PAs that could rectify episodes of negative affect depending on their achievement goals. To ascertain the possibility of devising such personalized PAs, this paper investigates the real-time prediction of both students' achievement goals and affective valence while interacting with MetaTutor, an agent-based intelligent tutoring system. We train classifiers using eye-tracking data to make such prediction, and show that these classifiers can outperform a majority-class baseline at predicting both achievement goals and emotion valence.

ACM Reference format:

S. Lallé, C. Conati, and R. Azevedo. 2018. Prediction of Student Achievement Goals and Emotion Valence during Interaction with Pedagogical Agents. In *Proc. of the 17th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, July 10-15, 2018, Stockholm, Sweden, ACM, New York, NY, 10 pages.

1 INTRODUCTION

Pedagogical agents (PAs) are intelligent virtual agents that support learning with Intelligent Tutoring Systems (ITS) by providing students with adaptive scaffolding (e.g., hints, prompts, feedback) [3,17,50]. There is extensive evidence that PAs can improve learning and engagement with ITS [3,30,50]. However, there is also work showing that PAs can generate negative affect in students (e.g., boredom, frustration), and that such negative affect can be modulated by students' individual differences (e.g., gender, personality, achievement goals) [2,29,38]. In particular, work [38] has shown that students' achievement goals (motivational goals in learning situations [20]) can influence their affective reaction to

PAs' prompts and feedback during interaction with MetaTutor, an ITS designed to scaffold student cognitive and metacognitive processes [3]. Specifically, students with a *mastery-approach goal*, who focus on developing competencies, tended to experience more negative affect when receiving PAs' scaffolding than students with a *performance-oriented goal*, who focus on outperforming their peers. Such findings suggest that these mastery-oriented students may benefit from PAs that can recognise these episodes of negative affect, and take actions to rectify them.

Here, we take a first step toward providing such personalization by studying the real-time prediction of student *achievement goals* (mastery-oriented vs performance-oriented) and *emotion valence* (positive vs negative affect) during interaction with MetaTutor. To do so, we train classifiers using eye-tracking data collected during student interaction with MetaTutor, namely eye gaze movements and distance of the head to the screen. Our results show that these classifiers can significantly outperform a majority class baseline for both prediction tasks, with accuracies suitable to further investigate the value of personalization driven by these classifiers.

There has been a substantial interest and positive results in using eye-tracking data to predict user's *long-term cognitive abilities* relevant to perceptual tasks (e.g., perceptual speed, spatial memory) [12,52], as well as *short-term states* such as confusion [38], boredom [34], and learning [35]. Here we contribute to this research by showing that eye-tracking data can help predict a user's long-term trait beyond cognitive abilities, namely achievement goals, which can be seen as an attitude or preference. Achievement goals are considered a facet of motivation given that they provide a purpose for the learning task at hand, and guide students' learning behaviors by setting the standards with which they evaluate success [20]. There is extensive work on the impact of achievement goals not only on student learning with PAs [19], but also students' performance and motivation in general [20,32].

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10-15, 2018, Stockholm, Sweden.

Thus, showing the feasibility of predicting this trait is significant beyond the specific interaction with MetaTutor considered in this paper.

Although the real-time prediction of emotion valence (i.e., a short-term state indicating the presence of positive vs. negative affect) has been extensively studied using a variety of data sources (e.g., facial expressions, speech, blood pressure [14,44,45]), here we show that it can be predicted solely from eye tracking data. We also show that this prediction can be made in conjunction with the prediction of achievement goals (i.e., a long-term trait). Thus, these results are promising for the design of personalized PAs that can react to episodes of negative affect based on the student's goals, by solely leveraging eye-tracking information.

It should be noted that, in recent years the focus of research in affective student modeling has been on detecting specific student emotions (frustration, boredom, confusion, joy, hope and several others), e.g., [5,9,18,53,55]. Accuracy results for some of these emotions are very promising, other emotions (e.g., confusion) can be more difficult to predict in academic settings [10,46]. In general, most of this work predicts one emotion at a time, despite substantial evidence that emotions can co-occur [13,15,24,28]. Furthermore, there are still limited results on what can be done with these specific predictions in terms of personalization (see literature review). In this paper, we provide further evidence that students do feel multiple emotions during interaction with MetaTutor. Predicting these multiple emotions is a challenging multiclass classification problem, and even if we could solve it, we would still need to understand how to enable the MetaTutor's PA to respond to these multiple emotions. This is indeed an intriguing research endeavor, however in this paper we want to explore the potential of a simpler approach based on the detection of emotion valence only. To this end, we discuss guidelines about how our classification results on emotion valence and achievement goals can inform the design of personalized interventions for the MetaTutor's PAs. In the rest of this paper, we first discuss related work, followed by a description of MetaTutor and of the study that generated the data used in this paper. Next, we describe the datasets and machine learning set up, followed by results, discussion and conclusions.

2 RELATED WORK

Students' affective reactions to PAs can greatly vary among students. In particular there is evidence that PAs can generate negative affect in some students [2,29,38,41], which can in turn hinder learning [4,29,30]. To address this issue, researchers started investigating affect-aware PAs that can detect and rectify episodes of negative affect for learning, as stated in the introduction.

The first step to design such personalization is to detect episodes of negative affect during interaction with the PA. Extensive research has been dedicated to such real-time detection of emotions (see [11,56] for an overview). A simple approach consists in predicting the valence (i.e., positive or negative) of the occurring emotions, which has been done by leveraging a variety of data sources such as facial expressions [45,56,57], spoken/dialog cues [14,22,40], interaction data [49], physiological sensors (e.g., blood pressure, heart rate, skin conductance, pupil dilation [23,25,31,44]). Other work focused on detecting specific emotions by leveraging similar data sources as listed above for valence detection. Most of this work focused on predicting basic emotions (fear, anger, delight), e.g., [46,56]. However these basic emotions do not capture the full scope of affective reactions in learning or academic settings [47], therefore extensive research has also focused on detecting more subtle affective states that typically occur during learning, such as boredom, confusion, hope (e.g., [5,9,18,55]). Although high prediction accuracies were overall obtained in such work, some affective states may be more difficult to detect during learning, for example confusion or frustration in [10,46]. Recently, Jaques et al. [34] examined the value of eye tracking for affect detection, and found that eye gaze movements and distance of the eyes to the screen can be predictor of boredom in MetaTutor [34].

Although previous studies (e.g., [13,15,24,28]) showed that different emotions frequently co-occur while learning with ITS, almost all work on affect detection focused on building models that can predict only one emotion at a time. A notable exception is [13] where the authors studied a Bayesian Network that can detect pairs of co-occurring emotions (e.g., joy/pride) in an educational game.

A few works started leveraging affect detectors to react to episodes of negative affect when they occur [16,55]. In the Wayang Tutor, a PA offering students to move on to another exercise when they are bored was found to improve student's engagement [55]. In AutoTutor, empathetic feedback designed to respond to boredom, frustration or confusion improved learning for students with low prior knowledge [16].

Research has shown that learning outcome and affect can be influenced by specific student traits such as gender, personality or achievement goals during interaction with PAs [2,29,38]. In particular, achievement goals have been found to impact both learning and emotions in MetaTutor, with results showing that mastery-oriented students learned less from the PAs and experience more negative affect than performance-oriented students [19,38,39]. More generally, these achievement goals have been shown to influence various aspects of learning, including learning strategies [1,51], academic performance [26,27], self-regulation [19,43], and belief that success follows from one's effort [1].

Despite the strong interest for achievement goals in the fields of psychology and ITS, there has been no work on the real-time detection of achievement goals. However, eye tracking has been extensively used in HCI to predict other long-term traits, namely cognitive abilities (e.g., perceptual speed, working memory) [12,52]. Eye tracking is also a predictor of several short-term states, such as confusion [37], mind wandering [8], learning [35], and cognitive load [7].

3 MetaTutor

MetaTutor [3] is a ITS containing multiple pages of text and diagrams about the circulatory system, as well as mechanisms to help students apply meta-cognitive learning strategies known as Self-Regulated Learning (SRL), with the assistance of multiple

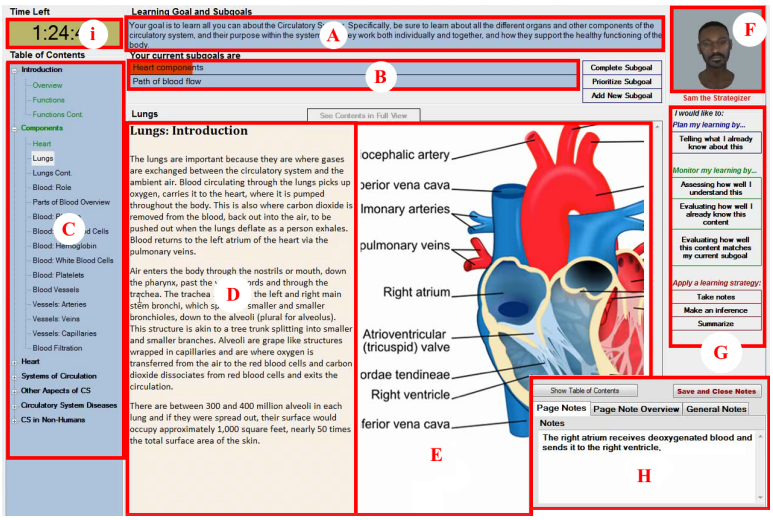

Figure 1. Screenshot of MetaTutor showing the main components of the interface: A) Overall Learning Goal, B) Current Subgoals, C) Table of Content, D) Learning Text, E) Diagram, F) Pedagogical Agent, G) SRL Palette, H) Note taking popup, I) Timer.

speaking pedagogical agents (PAs). When working with MetaTutor, students are given the overall goal of learning as much as they can about the human circulatory system, and they can set subsequent learning subgoals as they proceed through the available material. The MetaTutor interface and a description of its components are provided in Figure 1. A key element of the interface is the SRL palette (Fig 1G), designed to scaffold students self-regulatory processes by providing buttons they can select to initiate specific SRL activities (e.g., making a summary, taking a quiz, setting subgoals). Further SRL scaffolding is provided on each of these activities by three PAs, which in turn appear at the top right corner of the interface (Fig. 1F).

These PAs provide both feedback on the outcome of students' SRL activities (e.g., on quiz performance or on the quality of a summary), as well as prompts to guide these activities when needed. *Pam the Planner* scaffolds planning by assisting the student in creating subgoals (e.g., learning about the path of blood flow or heart disease) adequate to their progress through the material. *Mary the Monitor* scaffolds students' metacognitive monitoring processes, such as the self-assessment of their progress towards the established subgoals, and relevance of content to the subgoals. *Sam the Strategizer* supports students in applying cognitive learning strategies such as taking notes on the content or summarizing it in their own words. All PAs provide audible assistance through the use of a text-to-speech engine (Nuance) and are visually rendered using Haptek virtual characters. More details about the design of the PAs can be found in [3].

4 USER STUDY

The data used in this paper derives from a study designed to gain a general understanding of how students learn with MetaTutor [3]. Here we provide a brief summary of the study specific to the purposes of the paper. Thirty college students participated in the study, which started with a session during which students took a pretest on the circulatory system and questionnaires on demographics and learning-related traits, including the Achievement Goal Questionnaire that generated some of the data we use in this paper. In a second session, participants first underwent a calibration phase with a non-intrusive, monitor-mounted eye tracker (SMI RED 250). Next, they had 90 minutes with MetaTutor to learn as much as possible about the circulatory system, while their gaze was tracked with the SMI RED 250. During the interaction, the PAs spoke to the students for 24mins on average (SD = 10.5) [39], to scaffold the use of SRL strategies as described in the previous section. Thus, although MetaTutor has several components, the PAs have a considerable presence. At various points during the session participants were asked to report their current emotions, via the self-reports described next.

4.1 Measures of Achievement Goals and Emotions

Here we describe the study material that provided the data we use in this paper for predicting achievement goals and emotion.

Achievement Goals. The Achievement Goal Questionnaire Revised (AGQ-R) [20] is a 12-item self-report questionnaire that assesses four components of motivation in learning situations: (a) mastery-approach (e.g., goal to develop competence and skills), (b) mastery-avoidance (e.g., goal to avoid a failure to learn a skill), (c) performance-approach (e.g., goal to outperform others), and (d) performance-avoidance (e.g., goal to avoid being outperformed by others). All work done on MetaTutor has focused on mastery-approach and performance-approach goals only, given that avoidance goals are typically considered less useful to scaffold effective learning [19,38]. In the AGQ-R, students indicate their agreement with a series of items using a 7-point Likert scale. A sample item for the mastery-approach subscale is: "My aim is to completely master the material presented during this learning

session." A sample item for the performance-approach subscale is: "My goal is to perform better than the other student participants."

Emotion Reports. During the interaction with MetaTutor, at regular intervals of about 9 minutes students had to complete an on-line Emotions and Value (EV) Questionnaire [3]. This questionnaire consists of 15 items, each asking if the student currently feels a specific emotion: *enjoyment, pride, hope, curiosity, eureka, anxiety, boredom, frustration, contempt, confusion, sadness, shame, hopelessness, surprise, neutral*. One example item is: ''Right now I feel bored''. These items were rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), and students had to rate each emotion before they could proceed. The instructions and wording of the items are based on a subscale of the Academic Emotions Questionnaire (AEQ, [47]), that assesses students' emotions commonly experienced in learning or academic settings.

5 DATASET AND DATA LABELING

5.1 Data for Predicting Achievement Goals

Given the students' responses to the AGQ-R questionnaire from the study, following [19,38,39] we assign students a *dominant goal* with respect to *mastery-approach* vs *performance-approach*, based on which of the two received the highest score in the AGQ-R. In the case of a tie, it is assumed that the student had no dominant goal and the student is excluded from the analysis**.** There are 5 such students in our dataset. Based on this criterion, 16 of the remaining students had a mastery-oriented dominant goal (mastery-oriented students from now on), and 9 had a performance-oriented dominant goal (performance-oriented students from now on). These are the data that will be used to train our classifiers for achieving goal.

5.1 Data for Predicting Valence

206 EV emotion self-reports (simply *reports* from now on) were collected in the study (M = 6.8, SD = 1.4). Following [34], an emotion is considered present or *reported* at the time of a report if it is rated 4 or 5. 182 reports include at least one *reported* emotion, with 2.3 different emotions reported on average (SD = 1.44).

The fact that often students reported more than one emotion is consistent with previous work, as mentioned in the related work (e.g., [13,15,24,28]). Recall, however, that our goal is to predict emotion valence, thus we need to extract labels for valence from the reports. There is generally a well-established mapping between individual emotions and their valence [33], however the presence of co-occurring emotions complicates our task when there are *mixed reports* that include both emotions with a negative and with a positive valence. This is because there is no emotion theory that gives a formal definition of how the valence of co-occurring emotions integrates into an overall affective valence.

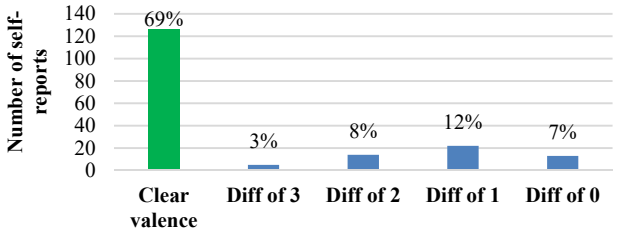

Figure 2. Histogram of self-reports valence.

Figure 2 shows how often these mixed reports appear in our dataset. The first (green) bar in Figure 2 denotes reports with a *clear valence*, i.e., where students reported solely positive or negative emotions. The rest of the bars (in blue) represent different types of *mixed* reports, i.e., reports with a specific absolute difference between the number of positive and negative emotions reported. For example, a report with 4 positive and 1 negative emotion is categorized as "Difference 3". As Figure 2 shows, about 30% of the reports (54) have mixed valence. Thirteen of these include the same number of positive and negative emotions (Diff of 0 in Fig 2), and thus may be considered as "undetermined" in terms of valence. The remaining 41 reports have a difference of at least one, meaning that there were either more positive or more negative emotions reported. Table 1 provides additional descriptive statistics for the 5 report categories in Figure 2. The first row shows the total number of students who generated reports in each category. The second row shows statistics on how many emotions per report there were in each category. The last row shows the number of reports with a majority of positive emotions (+), and a majority of negative emotions (-) in each category.

Given the challenge of clearly defining the overall valence of these mixed reports, we will initially focus on predicting valence as expressed in clear reports where students reported only positive or only negative emotions. These are the 72 positive reports and 54 negative reports listed in the last cell of the "clear valence" column in Table 1. This choice is suitable for our goal of investigating the feasibility of predicting emotion valence using eye-tracking data, because it removes possible confounds due to the unclear valence of the mixed reports. In section 7, however, we will also discuss an analysis that includes the classification of these mixed self-reports.

Table 1. Descriptive statistics of the EV reports.

EV report:	Clear valence	Mixed valence			
		Diff of 3	Diff of 2	Diff of 1	Diff of 0
#students	26	5	9	13	8
Statistic on number of emotions per report	*M=1.93* *SD=.86* *Max=4* *Min=1*	*M=5* *SD=0* *Max=5* *Min=5*	*M=4* *SD=0* *Max=4* *Min=4*	*M=3.36* *SD=.76* *Max=5* *Min=3*	*M=1.15* *SD=.18* *Max=2* *Min=1*
Valence	+ 72 - 54	+ 5 - 0	+ 13 - 1	+ 8 - 14	N/A

6 MACHINE LEARNING SET UP

We describe here the machine learning analysis we conducted to predict the binary labels generated for achievement goals (mastery vs. performance) and for emotion valence (positive vs. negative).

6.1 Data Windows

As described in section 4.2, students were prompted by MetaTutor to report their emotions at regular intervals (about 9 minutes). To predict the valence of a given report *r*, we use gaze data collected between the submission of the previous report and the appearance of *r*. In order to ascertain how much data leading up to a report is needed to predict it, we built our machine learning models over two different windows of data: (ii) a *short window* capturing data 15 seconds immediately before a report, (ii) and a *full window* capturing all data between two reports. We selected the 15-seconds window because it has been extensively used for affect detection [9,13,34,42]. As for the full window, it was found to be among the best at predicting boredom with MetaTutor in [34].

Unlike emotions, achievement goals do not change overtime (or they might change over long periods of time beyond the duration of a session with MetaTutor) [1,22]. What can change overtime during interaction with MetaTutor is the accuracy of a classifier for AG, as more eye-tracking data becomes available for classification. To ascertain how early we can predict a student's achievement goal during interaction with MetaTutor, we use an approach similar to [12,52] for predicting different long-term user traits: given the sequence of 6 reports seen by all the participants in the study, we generated a data window for each report, including the eye-tracking data available from the beginning of the interaction to that report (i.e., 9 minutes on average, 18 minutes, etc). We predict achievement goal for each of these 6 data windows.

6.2 Eye Tracking Features

The SMI eye tracker provides information on user gaze patterns (*Gaze* from now on) as well as on the distance of the user's head from the screen (*Head Distance* from now on). From both gaze and head distance information, we derived a set of features listed in Table 2 that we leveraged to predict student's emotion valence and achievement goals during interaction with MetaTutor. We used EMDAT (*github.com/ATUAV/EMDAT*), an eye tracking data analysis toolkit, to generate the features in Table 2, described next.

Table 2. Set of features considered for classification.

***a)* Gaze Features (154)**
Overall Gaze Features (19):
Fixation rate, Mean & Std. deviation of fixation durations
Mean, Std. deviation of saccade length & Longest saccade
Mean & Std. deviation of saccade duration
Mean, Std. deviation, Min & Max of saccade speed
Mean, Rate & Std. deviation of relative saccade angles
Mean, Rate & Std. deviation of absolute saccade angles
Ratio between total fixation duration and total saccade duration
AOI Gaze Features for each AOI (135):
Fixation rate in AOI
Longest fixation in AOI, Time to first & last fixation in AOI
Proportion of time, Proportion of fixations in AOI
Prop. of transitions from this AOI to every AOI
***b)* Head Distance Features (6)**
Mean, Std. deviation, Max., Min. of head distance
Head distance at the *first* and *last fixation* in the data window

Gaze Features: Users' gaze patterns are captured in terms of *fixations* (gaze maintained at one point on the screen), and *saccades* (quick eye movement between two fixations). EMDAT generated the gaze features listed in Table 2 (part a) by calculating various summary statistics (e.g., *sum*, *mean*) over a user's fixations and saccades. These statistics were computed for gaze patterns over the whole interface, generating the gaze features labelled as *Overall Gaze Features* in Table 2a, as well as over specific areas of interest (AOI) in the MQ interface, generating the *AOI Gaze Features* in Table 2a. There are nine AOIs defined over nine regions of MetaTutor, shown Figure 1.

Head Distance: Head distance is obtained by averaging the distances from both eyes to the screen. Using EMDAT, we computed a set of summary statistics on users' head distance, suitable for describing fluctuations of this measure over the course of the interaction with MetaTutor. These include *min*, *max*, *mean*, and *std. dev.* of users' head distance (see Table 2, part b). We also included head distance at the *first* and *last* fixation in the data window as a way to capture variations of the measures between the closest and farthest datapoints to the emotion report in that window.

6.3 Classifiers

We tested 4 standard machine learning algorithms available in the Caret package in R [36] for our classifiers: Boosted Logistic regression (BL), Random Forest (RF), Neural Network (NN) and Support-Vector Machine (SVM). We chose these classifiers because they have been extensively used for affect detection, without concluding evidence as for which one is the best (see overview in [12]). As a baseline, we use a majority class classifier. For each of these 5 algorithms, we built:

- Six classifiers for achievement goal, one for each data window at reports 1 through 6,
- Two classifiers for valence, one for the 15-second window and one for the full window described in section 6.1.

All classifiers were trained and evaluated with a process of *10-runs-10-folds stratified cross-validation* over students, meaning that all data for a given student are either in the training or in the test set. Stratification ensures that the class distribution in the folds is similar to that in the whole dataset. Due to the high number of features in our dataset, we discarded highly correlated features ($r > .8$) within the train folds only, to reduce data dimensionality. As a result 82 features were discarded on average across folds. The performance of each classifier was averaged across the 10 folds, and then again over the 10 runs. We report performance in terms of: *overall accuracy* (number of correct predictions divided by the total number of predictions) as well as

class accuracy (for each class, the number of correctly identified datapoints divided by the total number of datapoints in this class).

7 RESULTS

7.1 Prediction of Achievement Goals

Figure 3 shows overall accuracy for the binary prediction of students' achievements goal, namely whether students are *mastery-oriented* or *performance-oriented*. Figure 4 and Figure 5 show class accuracy respectively for mastery-oriented students and performance-oriented students. Prediction accuracy is shown for the various machine learning algorithms we tried (defined in Section 6.3), along with the majority class baseline, for classification done from the first to the sixth emotion self-report. Notice that the baseline always predicts students as "mastery-oriented", the majority class, and thus is unable to identify performance-oriented students.

To identify which combinations of classifier and self-report yield the best predictive performance, we ran a MANOVA[2] with:

- *overall accuracy* and both *class accuracies* at each run of cross-validation as the dependent variables;
- *self-report* (6 levels) and *classifier* (5 levels) as the factors.

The MANOVA reveals a significant[3] main effect of self-report ($F_{5,810} = 67.2$, $p < .000$, $\eta^2 = 0.77$) and of classifiers ($F_{12,810} = 109.1$, $p < .000$, $\eta^2 = 0.87$).

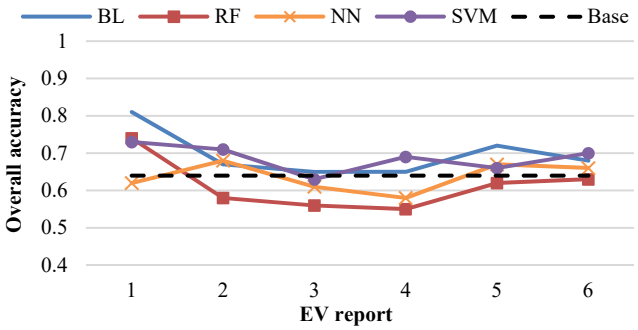

Figure 3. Overall prediction accuracy for the prediction of students' achievement goals.

Main effect of self-report. Post-hoc univariate ANOVAs indicate that there is a main effect of self-report on all three dependent variables, namely overall accuracy ($F_{5,270} = 278.36$, $p < .000$, $\eta^2 = 0.65$), class accuracy for mastery-oriented students ($F_{5,270} = 89.64$, $p < .000$, $\eta^2 = 0.56$), and class accuracy for performance-oriented students ($F_{5,270} = 1672.8$, $p < .000$, $\eta^2 = 0.72$). For each dependent variable, we ran pairwise comparisons between the reports to identify which report yields the best accuracy, using the Wilcoxon signed-rank test. To account for family-wise error, we used the Holm method to adjust *p*-values based on the number of comparisons made per dependent variable [33].

Results of the pairwise comparisons reveal that prediction at the first report significantly outperform all other reports in terms of *overall accuracy* and class *accuracy for performance-oriented students*, with medium to large effect sizes (from η^2=.16 to η^2=.57). This first report also ties for best with the fifth report in terms of *class accuracy for mastery-oriented students*. The fact that the best overall accuracy and class accuracy can be obtained at the very first report (i.e., the first 9 minutes of interaction) is important as it enables the possibility of early personalization to the student's achievement goal. Although it may seem surprising that having more interaction data available for training our classifiers at later reports does not yield better prediction of achievement goals, this result is consistent with previous work on predicting other long-term user traits from gaze data [12,52]. A possible explanation is that these user traits make the most difference at the onset of the interaction, when they heavily influence how users assess the task at hand and start reasoning about it.

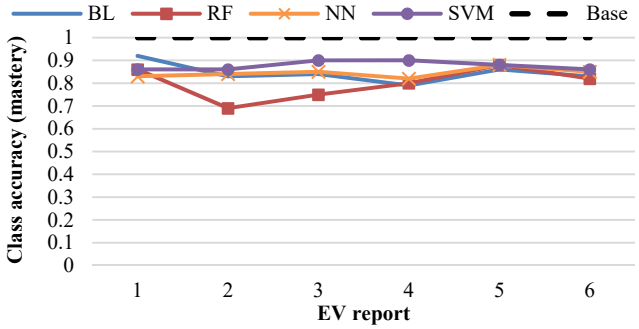

Figure 4. Class accuracy for mastery-oriented students.

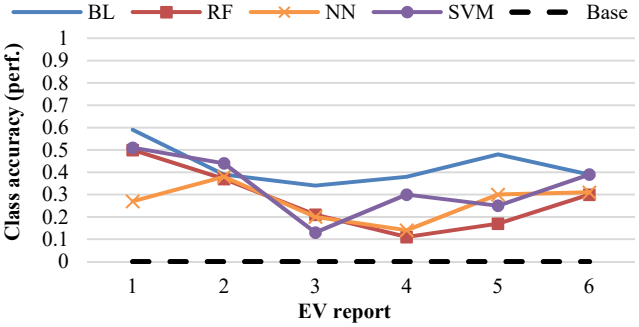

Figure 5. Class accuracy for performance-oriented students.

Main effect of classifiers. For the best report identified above (report 1), we ran Holm-adjusted pairwise comparisons between the classifiers for each independent variables, using the Wilcoxon signed-rank test. Results reveal that:

- BL reaches significantly higher overall accuracy than the baseline ($Z = 2.8$, $p = .005$, $\eta^2 = .38$), and so do SVM ($Z = 2.8$, $p < .005$, $\eta^2 = .38$) and RF ($Z = 2.8$, $p < .005$, $\eta^2 = .38$). BL also

[2] The data normality assumption was not met, thus we aligned the data using the *Aligned Rank Transform* method implemented in the *ARTool* package in R, widely used to run ANOVAs on non-parametric data [21,48,54].

[3] In this paper statistical significance is reported at the .05 levels, and effect sizes as large for $\eta^2 > .26$, medium for $\eta^2 > .13$, and small otherwise.

significantly outperforms both RF (Z = 2.7, p = .007, η^2 = .36) and SVM (Z = 2.8, p = .005, η^2 = .38).

- BL outperforms all other classifiers in terms of both class accuracies (obviously excluding the 100% accurate baseline for mastery-oriented students).

These results show that the BL classifier trained at the first report is the best at predicting achievement goals, reaching an overall accuracy of .81 over a .64 baseline (Figure 3). As shown Figure 4 and Figure 5, BL can detect mastery-oriented students with a very high accuracy (.92) at this first report, while still being able to recognize 59% of the performance-oriented students. Previous work has shown that mastery-oriented students tend to experience more negative affect with MetaTutor's PAs than performance-oriented students [38]. Thus the fact that BL can detect mastery-oriented students with a very high accuracy is important to provide these students with dedicated support when they experience negative affect.

7.2 Prediction of Emotion Valence

Here we evaluate the ability of BL, RF, SVM, NN and the majority class baseline (which always predicts *positive)* to predict student emotion valence (positive vs negative) with the two data windows described in section 6.1 (Full and 15-seconds). As done for achievement goals, we ran a MANOVA with: *overall accuracy* and *class accuracies* at each run of cross-validation as the dependent variables; *window* (2 levels) and *classifier* (5 levels) as the factors.

Results reveal a significant main effect of *classifier* ($F_{12,420}$ = 45.95, p < .000, η^2 = 0.54) only. Post-hoc univariate ANOVAs for each dependent variable show a significant main effect of *classifier* for all dependent variables, namely overall accuracy ($F_{4,140}$ = 321.4, p < .000, η^2 = 0.58), class accuracy for negative reports ($F_{4,140}$ = 221.7, p = .002, η^2 = 0.39), and class accuracy for positive reports ($F_{1,140}$ = 644.1, p = .002, η^2 = 0.39). To identify the best classifier, we started by running holm-adjusted pairwise comparisons between the classifiers with *overall accuracy* as the dependent variable, using the Wilcoxon signed-rank test. Figure 6 shows overall accuracy for each classifier and each of the two windows.

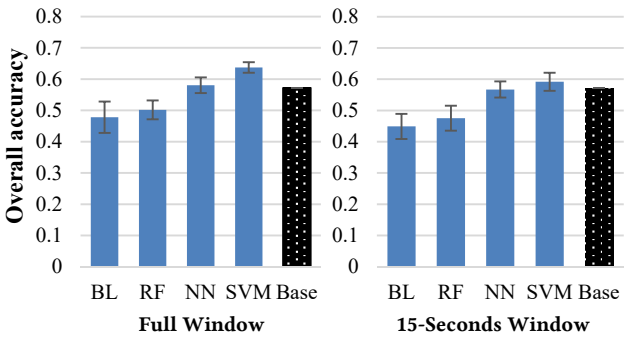

Figure 6. Accuracy and confidence intervals for the binary prediction of emotion valence in reports with a clear valence.

Results reveal that only SVM with the full window significantly outperforms the baseline (Z = 2.29, p = .02, η^2 = .51). Specifically, this SVM classifier reaches 0.64 accuracy, over a .57 baseline (see Figure 7, left). This finding indicates that the most promising approach to predict emotion valence from eye tracking in this context is to use a full window and SVM. Thus, we will focus the discussion of results for class accuracy on Full window only.

Figure 7 shows class accuracy with Full window for the negative and the positive reports. Although the baseline can correctly detect all positive reports (the majority class), it cannot identify negative reports. Pairwise comparisons (with the suitable adjustments described in the previous section) reveal that SVM and NN tie as the best at predicting positive reports. SVM is the second best classifier at predicting negative reports, being outperformed only by BL. Specifically, as shown Figure 7, SVM with data from the Full window can correctly identify 26% of the negative reports, and 88% of the positive reports. Thus, the higher classification error pertains to false positives, namely the classifier misclassifies 74% of the report with negative valence as positive (31% of all reports). These would translate into missed opportunities for MetaTutor to rectify episodes with negative affect for the students, but would not otherwise interfere with the interaction. The classification error for false negatives is rather low, namely only 12% of the positive self-reports are misclassified (6% of all reports). These are the misclassifications that could generate unwarranted interventions from the PAs, thus it is encouraging that their number remains low.

Pairwise comparisons also show that BL is the best classifier in terms of class accuracy for negative reports, being able to correctly identify 34% of them (see Figure 7, left). However, as shown Figure 7 (right), BL has the worst performance for positive reports, misclassifying as negative 44% of them (25% of all reports). Therefore, although BL would allow rectifying more episodes with negative affect than SVM, it would also lead to substantially more unwarranted interventions due to the high false negative rate. As future work, it may be worth examining ensemble techniques that could combine the strengths of the SVM and the BL classifiers.

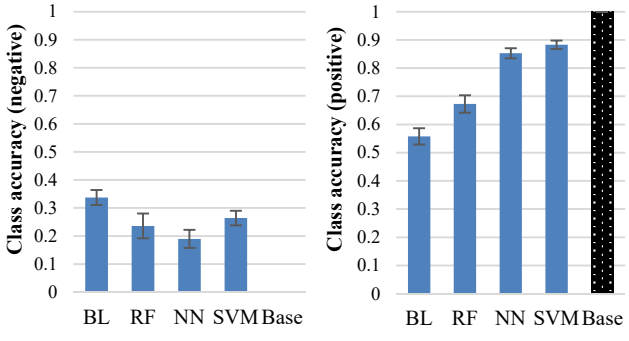

Figure 7. Class accuracy and confidence intervals for the negative (left) and positive (right) reports in the full window.

7.3 Prediction for the Mixed Reports

Results reported in the previous section were obtained for reports with a clear valence only, as this is the best way to show whether valence can be predicted, without incurring in possible confounds due to the difficulty of defining valence for mixed reports. Here we examine if the winning classifier from the previous section (*SVM+ Full window*) can also make predictions on the mixed reports with a difference of at least 1 (cf. Section 5), if we interpret their valence as *positive* when they include more positive than negative emotions, and *negative* otherwise. Thus, we repeat the analysis performed in the previous section, for the accuracy of SVM at full windows trained only on clear valence reports and tested (with 10-run 10-fold stratified cross-validation) on test sets including both clear and mixed reports. Table 3 reports the results (3rd row), as well the accuracies for predicting just the mixed reports (2nd row).

Both these sets of accuracies are very similar to the accuracies obtained on clear valence reports, indicating that perhaps student affective valence when they generated the mixed reports is indeed dictated by the valence of the dominant emotions reported. We also re-run the analysis by adding the mixed reports to both training and test sets, to ascertain whether training on more data would improve the accuracy of the SVM classifier with full window data, but there was no substantial change on the results (see last row in Table 3).

Table 3. Performance of the SVM+Full window classifiers for different approaches to include the mixed reports.

	Overall accuracy	Class accuracy	
		Negative	Positive
Trained and test on Clear (Section 7.2)	.637	.264	.882
Trained on Clear and tested on Mixed	.626	.248	.874
Trained on Clear and tested on Clear+Mixed	.638	.262	.885
Trained and tested on Clear+Mixed	.643	.275	.884

8 IMPLICATIONS FOR PERSONALIZATION

As explained in the introduction, our main goal is to enable personalization to rectify episodes of negative affect based on the student's achievement goal, leveraging the findings reported in [38]. That work found that PAs' scaffolding in MetaTutor tended to generate negative affect in mastery-oriented students. As a possible reason for this result, [38] suggests that the PAs' scaffolding may generate a sense of lower perceived autonomy for mastery-oriented students, based on an analysis of this achievement goal in [6]. Thus, possible forms of personalization for a student who is predicted to be mastery-oriented and experiencing negative affect during interaction with MetaTutor could include:

- reducing the number of prompts and feedback provided,
- offering the prompts and feedback in a way that gives the student more autonomy with them.

One possible concern with our results is that some of the reported class accuracies are not high. This concern might be addressed by devising more sophisticated classifiers trained on more and richer data that combines eye tracking with other data sources known to be predictive of valence (e.g., facial expressions, blood pressure [9,53,55]). Here, however, we discuss whether our gaze-based classifiers are still worth considering to drive the personalization described above, despite their inaccuracies. Specifically, we discuss the implications for misclassification. Note that we will not consider the effects of misclassification of affect for performance-oriented students, because we propose no personalization for these students.

(i) Misclassifying episodes of negative affect as positive for mastery-oriented students would result in missed opportunities to rectify negative affect. Although such missed opportunities would be frequent because our SVM classifier detects only 26% of the negative self-reports, they would not hinder students in any way since no intervention would be triggered.

(ii) Misclassifying episodes of positive affect as negative would result in reducing PAs' scaffolding or designing it for more autonomy if the student is predicted to be mastery-oriented. This might be problematic if the mastery prediction is incorrect and the student is in fact performance-oriented, because these students have been shown to benefit from the PAs' scaffolding [19] and this personalization would reduce it or make it more discretionary. For correctly classified mastery-oriented students, making the scaffolding less prominent when it is not generating negative affect might unnecessarily remove the chance for them to benefit from it. However, these episodes of unwarranted personalization are limited, because our SVM valence classifier has high accuracy on positive reports (only 12% of all positive reports are misclassified).

(iii) Misclassifying mastery-oriented students as performance-oriented would also result in missed opportunities to support these students in the presence of negative affect. Fortunately, such misclassifications occur rather unfrequently with our BL classifier, which correctly identified up to 92% of mastery-oriented students. Misclassifying performance-oriented students as mastery-oriented, on the other hand, is quite frequent (40% of them are misclassified), and in that case students would receive unwarranted and possibly hindering personalization when negative affect is detected, as discussed above. This said, if those performance-oriented students are truly experiencing negative affect, they might actually benefit from receiving less scaffolding from the PAs, although further analysis would be needed to assess this hypothesis.

Given the discussion above it appears that our gaze-based classifiers are still worth considering to drive personalization for mastery-oriented students who have negative affective reactions to the scaffolding provided by the MetaTutor PAs. Specifically, the benefits in personalization generated by correct predictions and the consequences of misclassification should be tested against MetaTutor with no such personalization or with the personalization provided to all students detected to be mastery-oriented, regardless of their affective states.

9 CONCLUSION

This paper contributes to research on intelligent Pedagogical Agents (PAs) by investigating the real-time prediction of students' achievement goals and emotion valence during interaction with several PAs featured in MetaTutor, an ITS that scaffolds self-regulated learning. The focus on these classification tasks is driven by findings indicating that the prompts and feedback generated by the MetaTutor' PAs can generate negative affective reactions in mastery-oriented students, which the PAs could address if detected in real time. We trained several classifiers on eye-tracking data capturing student gaze patterns as well as head distance to the screen. Our results show that a Boosted Logistic Regression (BL) classifier predicts achievement goals with an overall accuracy of .81, significantly over a .64 baseline. Furthermore, this accuracy is achieved after seeing about 10 minutes worth of data over a 90 minutes interaction, which is very promising for early personalization. This is to our knowledge the first result on predicting a student's long-term trait during interaction with an ITS. As for emotion valence, a SVM classifier significantly outperformed a .57 baseline by reaching an accuracy of .64. Based on these results, we provided suggestions on how to design personalized PAs in MetaTutor that can regulate episodes of negative affect based on the student's achievement goal.

As future work, we first plan to investigate other machine learning techniques to improve classification accuracy (e.g., ensemble modeling), as well as other data sources (e.g., face videos, log files). Second, we plan to evaluate the forms of personalization suggested in this paper, driven by our classifiers.

REFERENCES

[1] Carole Ames and Jennifer Archer. 1988. Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of educational psychology* 80, 3, 260–267.

[2] Ivon Arroyo, Beverly Park Woolf, James M. Royer, and Minghui Tai. 2009. Affective Gendered Learning Companions. In *Proceedings of the International Conference on Artificial Intelligent in Education*. Springer, 41–48.

[3] Roger Azevedo, Jason Harley, Gregory Trevors, Melissa Duffy, Reza Feyzi-Behnagh, François Bouchet, and Ronald Landis. 2013. Using trace data to examine the complex roles of cognitive, metacognitive, and emotional self-regulatory processes during learning with multi-agent systems. In *International handbook of metacognition and learning technologies*. Springer, 427–449.

[4] Ryan Sjd Baker, Sidney K. D'Mello, Ma Mercedes T. Rodrigo, and Arthur C. Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4, 223–241.

[5] Ryan Baker, Sujith Gowda, Michael Wixon, Jessica Kalka, Angela Wagner, Aatish Salvi, Vincent Aleven, Gail Kusbit, Jaclyn Ocumpaugh, and Lisa Rossi. 2012. Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. In *Proceeding of the International Conference on Educational Data Mining*, 126-133

[6] Moti Benita, Guy Roth, and Edward L. Deci. 2014. When are mastery goals more adaptive? It depends on experiences of autonomy support and autonomy. *Journal of Educational Psychology* 106, 1: 258.

[7] Pradipta Biswas, Varun Dutt, and Pat Langdon. 2016. Comparing Ocular Parameters for Cognitive Load Measurement in Eye-Gaze-Controlled Interfaces for Automotive and Desktop Computing Environments. *International Journal of Human-Computer Interaction* 32, 1, 23–38.

[8] Robert Bixler and Sidney D'Mello. 2015. Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness. In *Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization*. Soringer, 31–43.

[9] Nigel Bosch, Yuxuan Chen, and Sidney D'Mello. 2014. It's written on your face: detecting affective states from facial expressions while learning computer programming. In *Proceedings of the International Conference on International Conference on Intelligent Tutoring Systems*. Springer, 39–44.

[10] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2015. Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, 379–388.

[11] Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing* 1, 1, 18–37.

[12] Cristina Conati, Sébastien Lallé, Md. Abed Rahman, and Dereck Toker. 2017. Further Results on Predicting Cognitive Abilities for Adaptive Visualizations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. AAAI: 1568–1574.

[13] Cristina Conati and Heather Maclaren. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction* 19, 3, 267–303.

[14] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks* 18, 4, 407–422.

[15] John Dillon, Nigel Bosch, Malolan Chetlur, Nirandika Wanigasekara, G. Alex Ambrose, Bikram Sengupta, and Sidney K. D'Mello. 2016. Student Emotion, Co-occurrence, and Dropout in a MOOC Context. In *Proceedings of the International Conference on Educational Data Mining*, 353–357.

[16] Sidney D'mello and Art Graesser. 2012. AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems* 2, 4, 23.

[17] Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies* 70, 5, 377–398.

[18] Sidney D'Mello, Rosalind W. Picard, and Arthur Graesser. 2007. Toward an Affect-Sensitive AutoTutor. *IEEE Intelligent Systems* 22, 53–61.

[19] Melissa C. Duffy and Roger Azevedo. 2015. Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior* 52, 338–348.

[20] Andrew J. Elliot and Kou Murayama. 2008. On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology* 100, 3, 613.

[21] Jos Feys. 2016. Nonparametric tests for the interaction in two-way factorial designs using R. *The R Journal* 8, 1, 367–378.

[22] Kate Forbes-Riley and Diane Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*: HLT-NAACL 2004.

[23] Eric Granholm and Stuart R. Steinhauer. 2004. Pupillometric measures of cognitive and emotional processes. *International Journal of Psychophysiology* 52, 1, 1–6.

[24] Mirela Gutica and Cristina Conati. 2013. Student emotions with an edu-game: a detailed analysis. In *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, 534–539.

[25] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and research workshop on affective dialogue systems*, 36–48.

[26] Judith M. Harackiewicz, Kenneth E. Barron, Suzanne M. Carter, Alan T. Lehto, and Andrew J. Elliot. 1997. Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social psychology* 73, 6, 1284.

[27] Judith M. Harackiewicz, Kenneth E. Barron, Paul R. Pintrich, Andrew J. Elliot, and Todd M. Thrash. 2002. Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology* 94, 3, 638–654.

[28] Jason Harley, François Bouchet, and Roger Azevedo. 2012. Measuring learners' co-occurring emotional responses during their interaction with a pedagogical agent in MetaTutor. In *Proceedings of the International Conference on Intelligent tutoring systems*. Sringer, 40–45.

[29] Jason M. Harley, Cassia K. Carter, Niki Papaionnou, François Bouchet, Ronald S. Landis, Roger Azevedo, and Lana Karabachian. 2016. Examining the predictive relationship between personality and emotion traits and students' agent-directed emotions: towards emotionally-adaptive agent-based learning environments. *User Modeling and User-Adapted Interaction* 26, 2–3, 177–219.

[30] Steffi Heidig and Geraldine Clarebout. 2011. Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review* 6, 1, 27–54.

[31] Alicia Heraz and Claude Frasson. 2007. Predicting the three major dimensions of the learner's emotions from brainwaves. *International Journal of Computer Science* 2, 3, 187–193.

[32] Suzanne Hidi and Judith M. Harackiewicz. 2000. Motivating the academically unmotivated: A critical issue for the 21st century. Review of educational research 70, 2: 151–179.

[33] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* 6, 2, 65–70.

[34] Natasha Jaques, Cristina Conati, Jason M. Harley, and Roger Azevedo. 2014. Predicting Affect from Gaze Data during Interaction with an Intelligent Tutoring

System. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*. Springer, 29–38.

[35] Samad Kardan and Cristina Conati. 2012. Exploring gaze data for determining user learning with an interactive simulation. In *Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization*, 126–138.

[36] M. Kuhn. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28, 5, 1–26.

[37] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2016. Predicting confusion in information visualization from eye tracking and interaction data. In *Proceedings on the 25th International Joint Conference on Artificial Intelligence*. Springer, 2529–2535.

[38] Sébastien Lallé, Nicholas V. Mudrick, Michelle Taub, Joseph F. Grafsgaard, Cristina Conati, and Roger Azevedo. 2016. Impact of Individual Differences on Affective Reactions to Pedagogical Agents Scaffolding. In *Proceedings of the International Conference on Intelligent Virtual Agents*. Springer, 269–282.

[39] Sébastien Lallé, Michelle Taub, Nicholas V. Mudrick, Cristina Conati, and Roger Azevedo. 2017. The Impact of Student Individual Differences and Visual Attention to Pedagogical Agents during Learning with MetaTutor. In *Proceedings of the 18th International Conference on Artificial Intelligence in Education*. Springer, 149-161

[40] Chul Min Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* 13, 2, 293–303.

[41] Nicholas Mudrick, Roger Azevedo, Michelle Taub, and François Bouchet. 2015. Does the Frequency of Pedagogical Agent Intervention Relate to Learners' Self-Reported Boredom while using Multiagent Intelligent Tutoring Systems? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 1661–1666.

[42] Kasia Muldner, Robert Christopherson, Robert K. Atkinson, and Winslow Burleson. 2009. Investigating the Utility of Eye-Tracking Information on Affect and Reasoning for User Modeling. *Proceedings of the International Conference on User Modeling, Adaptation and Personalization*. Springer, 138–149.

[43] John C. Nesbit, Philip H. Winne, Dianne Jamieson-Noel, Jillianne Code, Mingming Zhou, Ken Mac Allister, Sharon Bratt, Wei Wang, and Allyson Hadwin. 2006. Using cognitive tools in gStudy to investigate how study activities covary with achievement goals. *Journal of Educational Computing Research* 35, 4, 339–358.

[44] Serina A Neumann and Shari R Waldstein. 2001. Similar patterns of cardiovascular response during emotional activation as a function of affective valence and arousal and gender. *Journal of Psychosomatic Research* 50, 5, 245–253.

[45] Mihalis A. Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Transactions on Affective Computing 2, 2: 92–105.

[57] Zhihong Zeng, Yun Fu, Glenn I. Roisman, Zhen Wen, Yuxiao Hu, and Thomas S. Huang. 2006. Spontaneous emotional facial expression detection. *Journal of multimedia* 1, 5: 1–8.

[46] Luc Paquette, Jonathan Rowe, Ryan Baker, Bradford Mott, James Lester, Jeanine DeFalco, Keith Brawner, Robert Sottilare, and Vasiliki Georgoulas. 2016. Sensor-Free or Sensor-Full: A Comparison of Data Modalities in Multi-Channel Affect Detection. *Proceedings of the International Conference on Educational Data Mining*, 93–100.

[47] Reinhard Pekrun, Thomas Goetz, Anne C. Frenzel, Petra Barchfeld, and Raymond P. Perry. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology* 36, 1, 36–48.

[48] Judy Robertson and Maurits Kaptein. 2016. *Modern Statistical Methods for HCI*. Springer.

[49] Sergio Salmeron-Majadas, Olga C. Santos, and Jesus G. Boticario. 2014. An evaluation of mouse and keyboard interaction indicators towards non-intrusive and low cost affective modeling in an educational context. *Procedia Computer Science* 35, 691–700.

[50] Noah L. Schroeder and Olusola O. Adesope. 2014. A systematic review of pedagogical agents' persona, motivation, and cognitive load implications for learners. *Journal of Research on Technology in Education* 46, 3, 229–251.

[51] Corwin Senko and Kenneth M. Miles. 2008. Pursuing their own learning agenda: How mastery-oriented students jeopardize their class performance. *Contemporary Educational Psychology* 33, 4, 561–583.

[52] Ben Steichen, Cristina Conati, and Giuseppe Carenini. 2014. Inferring Visualization Task Properties, User Performance, and User Cognitive Abilities from Eye Gaze Data. *ACM Transactions on Interactive Intelligent Systems* 4, 2, 11.

[53] Noam Tractinsky and Dror Zmiri. 2006. Exploring attributes of skins as potential antecedents of emotion in HCI. *Aesthetic computing*, 405–422.

[54] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 143–146.

[55] Beverly Woolf, Winslow Burleson, Ivon Arroyo, Toby Dragon, David Cooper, and Rosalind Picard. 2009. Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology* 4, 3–4, 129–164.

[56] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. 2009. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1: 39–58.