When the Question is Part of the Answer: Examining the Impact of Emotion Self-reports on Student Emotion

Michael Wixon and Ivon Arroyo

Learning Sciences and Technologies. Social Sciences and Policy Studies Department
Worcester Polytechnic Institute

Abstract. A variety of methodologies have been put forth to assess students' affective states as they use interactive learning environments (ILEs) and intelligent tutoring systems (ITS), such as classroom observations and subjective coding, self-coding by students after replays, as well as self-reports of student emotion as students are using the learning environment. Still, it is unclear what the disadvantages of each methodology are. In particular, does measuring affect by asking students to self-report alter student affect itself? The following work explores this question of how self-reports themselves can bias affective states, within one particular tutoring system, Wayang Outpost.

Keywords: affect, assessment, modeling of emotions.

1 Motivation

Several methods have been proposed to measure how students feel as they interact with digital learning environments. This data has been used to research how to facilitate learning by enhancing students' positive valence affective states, mitigate potentially harmful negative valence affective states such as boredom [1, 2], and explain the complex relationship between frustration/confusion and learning [3, 4]. In pursuit of these goals, emotion assessments are collected to generate affect models and classifiers that may be used to automatically predict students' emotions within the digital learning environment.

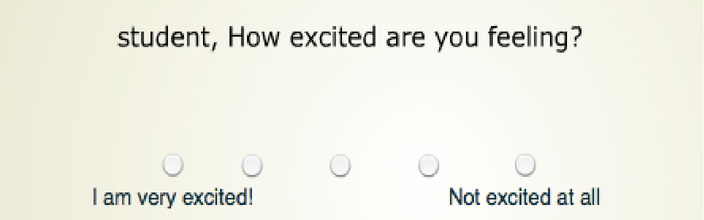
The first step in developing these automated detectors is to establish a "ground truth" label of affect that a detector can approximate. There are several different methodologies to obtain such "ground truth" label. A large body of work in affect detection utilizes videos of students' facial expressions with posterior coding by students themselves, for posterior detection through behaviors such as gaze tracking, and galvanic skin response sensors [5-6]. Another approach is BROMP, which employs specially trained observers to identify students' affective states through unobtrusive observations and inter-rater reliability of observers to establish construct validity [8,9]. Finally, students may self-report their emotions as they are learning, to obtain "ground truth" labels of student affective states [7].

We have employed this third method of self-report. Specifically, our approach has been minimally invasive, similar to the concurrent forced-response technique [10] which uses Likert scales in between problems [11]. Our endeavors to be minimally invasive by placing our self-reports between problems and furthermore by not requiring responses, are meant to avoid the pitfalls that come the interruptions to

V. Dimitrova et al. (Eds.): UMAP 2014, LNCS 8538, pp. 471–477, 2014.

students' work that self-report necessitates. Prior work has found that interrupting students during primary tasks can cause an increase in annoyance [12], and it is our hope that judicious use of self-reports will mitigate effects such as these. Nonetheless, there are still concerns that even unobtrusively collecting self-report data may influence a student's affective state [10].

Our most troubling evidence of self-report negatively impacting student affect, although anecdotal, comes from our own data collections. In a prior study, after students were asked to give a self-report of their affective state, they were asked to explain their self-report. Student responses included "[I am frustrated] Because you keep asking me if I am frustrated." Such responses were rare, but might indicate a larger unreported trend. So we resolved to quantitatively address concerns with self-reported affect influencing students' affective states.



student, How excited are you feeling?

○ ○ ○ ○ ○

I am very excited! Not excited at all

Fig. 1. Emotion Question that appears on average every5-7 minutes on average. "student" is replaced by the student's first name. "Why is that?" question (not shown) allows students to expand on their reasons for their rating.

2 Method

Participants. Participants consisted of two hundred and ninety five (295) students, 7th, 8th, 9th and 10th graders from three semi-rural area schools in Massachusetts from several studies involving the Wayang Outpost math tutoring system, in 2009. Students used the tutoring system for several days (3-5 days) during 1-2 weeks.

Wayang Outpost. Wayang Outpost is a mathematics ITS which covers K-12 material such as number sense, pre-algebra, algebra, geometry. Wayang Outpost[1] adapts content presented to students depending on mastery learning. It emphasizes scaffolding students through multimedia hints and pedagogical agents also known as "learning companions" who provide both motivational and cognitive support [13]. Affective measures for self-report were selected based on prior work used to model a range of various emotional states during learning [14]: Confidence/Anxiety (bipolar scale), Excitement (unipolar), Frustration (unipolar), and Interest/Boredom (bipolar scale), that overlap with metrics from the Theory of Achievement Emotions [19].

Overview of Analyses. We performed three levels of analysis, each with finer granularity than the previous one. First, we examined the correlation between the affective state reported by a student to the total number of self-reports a student was

[1] `http://wayangoutpost.com`

asked so far, while controlling for time in tutor session. We expected that this analysis might reveal a relationship between sheer quantity of self-reports and student affective state. Second, we examined the correlation between affective self-reports and the interval of time that had passed since the last self-report. This analysis was also correlational and similar to the first one, except that total self-reports was now replaced with "time since last self-report". Finally, we examined changes in students' affective states at an even finer grain size: from one problem to the next. We used predictive models of affect for this analysis, which allowed us to understand the affective state of individual students at any math practice problem. We considered the difference in emotional state between a pair of problems preceding a self-report as a control condition, and compared this general trend to the difference between the problem preceding a self-report and the problem following a self-report. The next section describes our results.

Predictive Models. The affect detectors were trained under five fold batch student-level cross validation using simple linear regression in Rapidminer 5.0 [15]. Post-hoc discretization (i.e. survey responses and predictions 1 to 2 are negative affect, 3 is neutral affect, and 4 to 5 are positive affect) was employed to obtain weighed Cohen's Kappa [16] values measuring agreement between actual self-report and prediction. Results of these detectors are in Table 1. We concede that the performance of these detectors is poorer than the typically accepted Cohen's Kappa of 0.4; however, generally accepted Kappas in sensor-free affect detection tend to be lower than Kappas detected for other constructs [9,17].

Table 1. Affect Detector Performance

	Confidence	Excitement	Frustration	Interest
Pearson's *R* for Continuous Prediction	0.404	0.224	0.372	0.232
Kappa for Discretized Classification	0.200	0.151	0.173	0.100

Additionally, due to its sensitivity to affect as a continuous rather than binary variable this detector suffers a handicap: it is more difficult to select affect correctly due to chance (e.g this detector which distinguishes between "bored", "neutral", and "interested" and may be outperformed by a detector which need only distinguish between the binary "bored" vs "not bored"). The affect model generates a prediction of each of each affect after each solved problem. Details on how similar models are created, which relies on a classifier based on linear regression, may be found in [18].

3 Results

First Analysis. This analysis measured the correlation between self-reported affect and the number of times a student had been asked to self-report on any affect so far, for each of our four emotions while controlling for "time in session" to account for any changes in affective state that might be due to fatigue. We obtained only one near significant correlation between frustration and requests for self-report. This correlation was negligible (r=0.043, p=0.108, df=1408).

Second Analysis This analysis examined how the spacing between requests for self-report influenced students' self-reported affect. This is possible because while reports were set to happen at intervals of 3-5 minutes, Wayang Outpost would "wait" until the student had finished the actual math problem. Thus, the correlation between the interval of "time since last self-report" and self-reported affect were considered, while controlling for overall "time spent in the tutoring session". We found a negligible relationship between "time since last self-report" and change in interest (*r*=-0.074, p=0.011, df=1185).

Third Analysis. In the last analysis, we considered the change in affect from one problem to the next. Here we looked at the problems which are adjacent to self-reports in order to get a better idea at how self-reports influence affect, within a small window of time. For this analysis, we were able to estimate affective states for problems where affect was not self-reported by using our predictive models of affect. The models predict a student's affect given information from pretest surveys and their log files. Since our prior analyses had shown little to no change in affect due to self-report, we used our models to detect change in affect between two problems that have no self-report between them (e.g. between Time1 and Time2 as displayed in figure 2) as compared to the change in affect between two problems that did have a self-report between them (e.g. between Time2 and Time3). These time and self-report immediately follow one another which minimizes the chance of other intervening events influencing our effects.

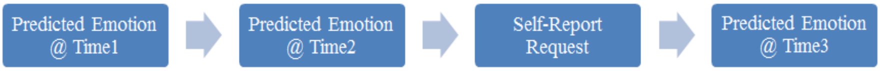

Predicted Emotion @ Time1
Predicted Emotion @ Time2
Self-Report Request
Predicted Emotion @ Time3

Fig. 2. Methodology for the Third Analysis

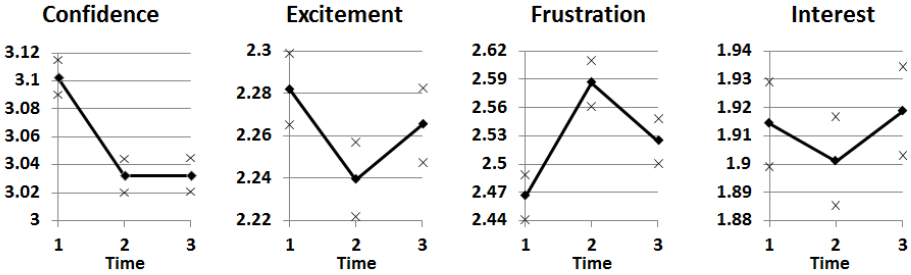


Fig. 3. Predicted Affect at Each Time (Paired Samples T-Test shows significant difference in affect p < 0.05 between Times 1 & 2 and 2 & 3 for all. Self-Report Occurs between Times 2 & 3, but not 1 & 2. 95% Confidence Interval denoted by Xs. N = 2878 for Confidence, Exceitment & Interest. N = 3001 for Frustration).

Figure 3 illustrates how students' affect changes between problems when there is no intervening self-report (Time 1 to Time 2) and how their affect changes when there is an intervening self-report (Time 2 to Time 3). The changes we have detected in affect here seem negligible given that students typically respond on a Likert scale from 1 to 5. Our analyses have been ordered in progressing sensitivity, examining information at smaller and smaller grain size. At the individual problem level we have detected a negligible, but consistent effect. With no intervening self-report students' affect states appear to become very slightly more negative (i.e. confidence, excitement, and interest decrease while frustration increases). However, with an intervening self-report the valences of students' affective states become more positive (excitement and interest increase while frustration decreases). While the change in affect appears to be negligible in magnitude it is statistically significant at a large sample size (Paired sample T-Tests indicate $p < 0.05$ for all cases when comparing each affect at Time 1 to Time 2, and Time 2 to Time 3).

4 Discussion

Most of the results of the first two analyses aren't significant. We are certainly not proving the null here that self-reports do not impact student affect, but we are addressing that concern that self-reports may influence students' affective state. In this way, it is our belief that these results are a valuable contribution to researchers modeling affect using self-report measures.

However, the data of the third analysis indicates that self-report appears to have an small but positive effect on the valence of the affects measured herein, from activity to activity. To the extent that the question is part of the answer, it seems to be a positive part, improving students mood as compared to not intervening. This may be due to the system exhibiting some degree of empathy with the student, which has been shown to improve students' overall mood [20]. Another possible reason for this positive effect is that students have the possibility of venting any negative affect, not only through the actual scale but also through the "Why is that?" text box that accompanies the question. This change from negative affective trends (before the self-report) to positive affective trends (after the self-report) are causing an oscillating effect, which is harder to see in the more global first two analyses we report on.

A major weakness of this work is that it treats affective reports as independent when many of them may come from the same student. While these results appear indicate a small and overall positive effect of self-report on students' affect they of course do not preclude large or negative effects of self-report on students' affect in different environments or with distinct samples of students. It is our hope that the analytical methods outlined in this work may serve as a means of easily checking on the effects of self-report confounds in other learning environments.

References

1. D'Mello, S., Person, N., Lehman, B.: Antecedent-Consequent realtionships and Cyclical Patterns between Affective States and Problem Solving Outcomes. In: Dimitrova, V., Mizoguchi, R., du Boulay, B., Grasser, A. (eds.) Artificial Intelligence in Education. Building Learning Systems that Care: from Knowledge Representation to Affective Modelling, pp. 57–64. IOS Press (2009)
2. Pekrun, R., Goetz, T., Daniels, L., Stupnisky, R.H., Raymond, P.: Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. Journal of Educational Psychology 102(3), 531–549 (2010)

3. Baker, R.S.J.D., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C.: 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 (2010)
4. Gee, J.P.: Situated Language and Learning: A Critique of Traditional Schooling. Routledge Taylor & Francis, London (2004)
5. D'Mello, S., Graesser, A.: Multimodal Semi-Automated Affect Detection from Conversational Cues, Gross Body Language, and Facial Features. User Modeling and User-Adapted Interaction 20(2), 147–187 (2010)
6. D'Mello, S.K., Picard, R.W., Graesser, A.C.: Towards an Affect-Sensitive AutoTutor. Special Issue on Intelligent Educational Systems IEEE Intelligent Systems 22(4), 53–61 (2007)
7. Arroyo, I., Woolf, B., Cooper, D., Burlesom, W., Muldner, K., Cristopherson, R.: Emotion sensors go to school. In: Dimitrova, V., Mizoguchi, R., Du Boulay, B., Graesser, A. (eds.) 14th International Conference on Artificial Intelligence In Education, IOS Press, Amsterdam (2009)
8. Ocumpaugh, J., Baker, R.S.J.D., Rodrigo, M.M.T.: Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. EdLab, Ateneo Laboratory for the Learning Sciences, New York, Manila (2012)
9. Baker, R.S.J.D., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L.: Sensor-free automated detection of affect in a Cognitive Tutor for Algebra. In: Proceedings of the 5th International Conference on Educational Data Mining, pp. 126–133 (2012)
10. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.S.J.D.: Knowledge Elicitation Methods for Affect Modeling in Education. International Journal of Artificial Intelligence in Education 22(3), 107–140 (2013)
11. Conati, C.: How to Evaluate a Model of User Affect? In: André, E., Dybkjær, L., Minker, W., Heisterkamp, P. (eds.) ADS 2004. LNCS (LNAI), vol. 3068, pp. 288–300. Springer, Heidelberg (2004)
12. Bailey, B.A., Konstan, J.A.: On the need for attention-aware systems: measuring effects of interruption on task performance, error rate, and affective state. Computers in Human Behavior 22, 685–708 (2006)
13. Woolf, B.P., Arroyo, I., Muldner, K., Burleson, W., Cooper, D., Dolan, R., Christopherson, R.M.: The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 327–337. Springer, Heidelberg (2010)
14. Kort, B., Reilly, R., Picard, R.W.: An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In: Proceedings of the IEEE International Conference on Advanced Learning Technologies, pp. 43–46 (2001)
15. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), pp. 935–940 (2006)
16. Cohen, J.: Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213–220 (1968)
17. Pardos, Z.A., Baker, R.S.J.D., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M.: Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In: Proceedings of the 3rd International Conference on Learning Analytics and Knowledge, pp. 117–124 (2013)

18. Arroyo, I., Cooper, D.G., Burleson, W., Woolf, B.P.: Bayesian Networks and Linear Regression Models of Students' Goals, Moods, and Emotions. In: Handbook of Educational Data Mining. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series (2010)
19. Arroyo, I., Shanabrook, D., Woolf, B.P., Burleson, W.: Analyzing Affective Constructs: Emotions, Motivation and Attitudes. In: International Conference on Intelligent Tutoring Systems (2012)
20. Nguyen, H., Masthoff, J.: Designing empathic computers: the effect of multimodal empathic feedback using animated agent. In: Proceedings of the 4th International Conference on Persuasive Technology (2009)