La Mort du Chercheur: How well do students' subjective understandings of affective representations used in self-report align with one another's, and researchers'?

Wixon[1], Danielle Allessio[2], Jaclyn Ocumpaugh[3], Beverly Woolf[2], Winslow Burleson[4] and Ivon Arroyo[1]

[1]Worcester Polytechnic Institute, Worcester Massachusetts
{mwixon, iarroyo}@wpi.edu
[2]University of Massachusetts, Amherst Massachusetts
{allessio@educ, bev@cs}.umass.edu
[3]Teachers College, Columbia University, New York, New York
jocumpaugh@wpi.edu
[4]New York University, New York, New York
Wb50@nyu.edu

Abstract. We address empirical methods to assess the reliability and design of affective self-reports. Previous research has shown that students may have subjectively different understandings of the affective state they are reporting [18], particularly among younger students[10]. For example, what one student describes as "extremely frustrating" another might see as only "mildly frustrating." Further, what students describe as "frustration" may differ between individuals in terms of valence, and activation. In an effort to address these issues, we use an established visual representation of educationally relevant emotional differences [3, 8, 25]. Students were asked to rate various affective terms and facial expressions on a coordinate axis in terms of valence and activation. In so doing, we hope to begin to measure the variability of affective representations as a measurement tool. Quantifying the extent to which representations of affect may vary provides a measure of measurement error to improve reliability.

Keywords: Affective States; Intelligent Tutoring Systems; Reasons for Affect

1 Introduction

The evaluation of students' affective states remains an incredibly difficult challenge. While recognized as a key indicator of student engagement [14, 17, 26], there remains no clear gold-standard for identifying an affective state, leading to researchers such as Graesser & D'Mello [13] to call for greater attention to the theoretical stances that certain research methods entail. A full theoretical review is beyond the scope of this paper; instead, the current work presents a pilot study designed to empirically evaluate the reliability of two different types of affective self-reports in an educational

adfa, p. 1, 2011.

context. Reliability is measured both in terms of inter-rater reliability (the degree of agreement between students), and "inter-method" reliability (i.e. given words or facial expressions as representations of affective states, which representation produces more consistent results).

A considerable body of research has been devoted to affect computing, and in particular to affect detection in educational software [9]. Progress has been made with methods that include self-report [8, 10], physiological sensors [1, 24], video-based retrospective reports [5, 15], text-based [11, 19], and field observation [16, 23] data. However, much of this research evaluates success based on the ability of a model to predict when a training label is present or absent, without giving deeper consideration to questions about the appropriateness of the training label itself.

Even within limited to the body of research that relies on self-report research, there are serious concerns about how methodological decisions might impact student responses. In addition to issues about the frequency and timing of surveys, one primary area of concern is that students may have subjectively different understandings of the state they are reporting [19], an effect that is likely to be even greater among younger students [10]. For example, Graesser and D'Mello [13] have suggested that a students' tolerance of cognitive disequilibrium (e.g., confusion or frustration) is probably conditioned by their knowledge and prior success with the topic they are interacting with. Further, what students describe as "frustration" in itself may differ between individuals in terms of dimensional component measures of affect: valence, activation, and dominance. The former two dimensions are typically used to differentiate affective states [4], and the latter used in some cases [7].

In this study, we explore these interpretative issues using three different types of representations that have been employed in previous self-report studies: words, facial expressions, and dimensional measures. In particular, we are interested in verifying that students' understanding of the meaning of these representations aligns with interpretations of these labels that are present in the literature (as constructed by experts). To this end, we use dimensional measures (valence & activation) to compare how students respond to both linguistic representations and pictorial representations, further testing hypotheses that the latter might be more appropriate for surveying students [19, 21, 22]. Our goal is to determine the extent to which this student population shows variance in the interpretation of these two different types of representations, since substantial variation in student perception should be accounted for in subsequent research. Last, while we might achieve researcher agreement in terms of methods and terminology for self-reported affects, that will do little good if there is a large degree of variance in terms of our subject pool's agreement on the meaning of these constructs.

1.1 Methods

Students surveyed included eighty one 7th graders from two Californian middle schools in a major city (among the 30 most populous cities in California), where a majority of census respondents identified as Hispanic or Latino and median household income was within one standard deviation of California's overall median household income. They were surveyed at the end of the academic year.

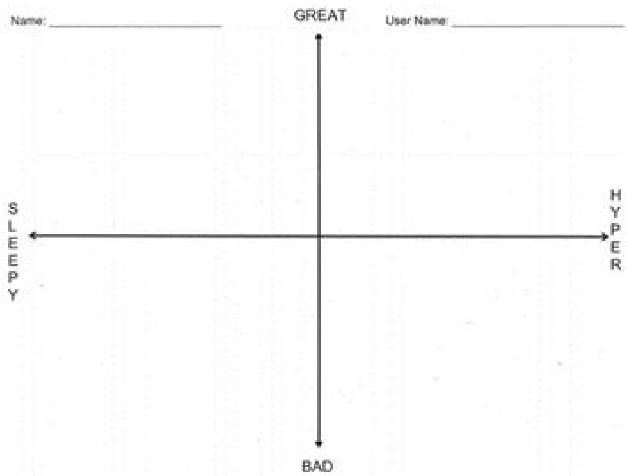

Fig. 1. Blank Valence & Activation Sheet given to Students

Students were asked to place both textual and facial representations of affect on an XY axis of Activation=X Valence=Y. Textual representations of affect were selected based on the affective states that have been used in the past [2, 12], that corresponded to quite different levels of activation x valence according to us researchers, so that words would theoretically cover all quadrants. These terms and their researcher-hypothesized valence x arousal placements included: Angry (low valence x high activation), Anxious (low valence x high activation), Bored (low valence x low activation), Confident (high valence x low activation), Confused (med-low valence x med-high activation), Enjoying (high valence x medium activation), Excited (high valence x high activation), Frustrated (low valence x high activation), Interested (high valence x medium activation) and Relieved (high valence x med-low activation). In general, it was clear to the researchers which word corresponded to which face, with a few exceptions, such as the level of activation that should be associated to enjoying and interest. An established set of emoticons were chosen from previous affective research [8] that corresponded to extreme emoticon states of activation x valence x dominance. While the emoticons possessed these three attributes, our participants were asked only to orient them based on activation and valence.

Each student was presented with a sheet of paper depicting a coordinate axis with activation from "sleepy" to "hyper" on the x-axis and "bad" to "great" on the y-axis. These terms were used to express what valence and activation mean experientially, using language that children are familiar with and could relate to. Activation is then expressed more as a physical experience of arousal, while Valence is expressed not as much as a physical experience but as a judgment of the positive or negative nature of the experience. Later, during coding, these axes were mapped discretized into a seven point scale of -3 to 0 to +3 at either extreme of each axis, defining a grid of 7 x 7.

Students were also given stickers for the 10 separate affective terms: Angry, Anxious, Bored, Confident, Confused, Enjoying, Excited, Frustrated, Interested, & Relieved, see Figure 2; as well as 8 stickers to depict each extreme emoticon expression from the ends of each of the 3 axis coordinate systems including: pleasure, activation, and dominance [8]. Students placed each of these stickers on their coordinate axes according to where they felt each term or emoticon should be placed with respect to valence and activation.

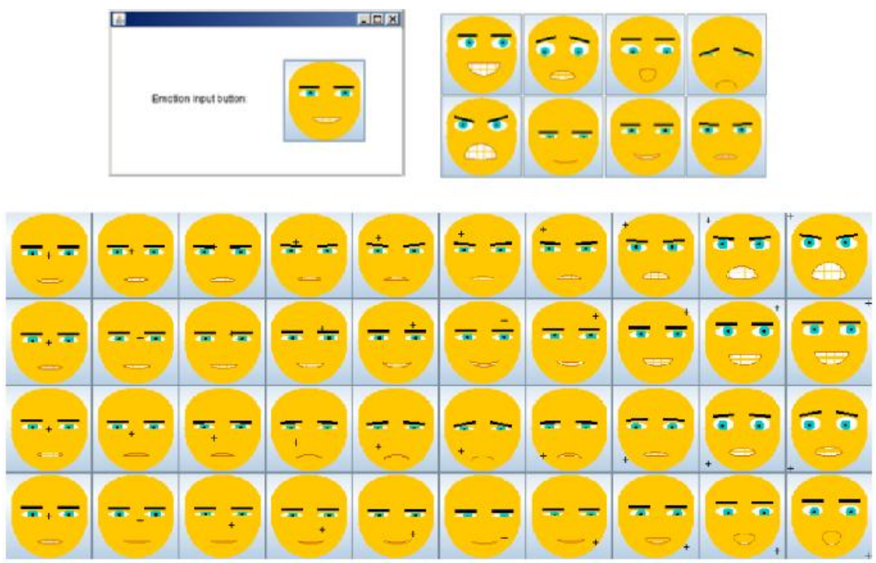

Fig. 2. Directly from Broekens, & Brinkman, 2013 [8]. Top left displays the affect button inter-face. Students use the cursor to change the expression in the inter-face. Depending on their actions, one of 40 affective expressions may be displayed; these expressions, shown across the bottom of this figure, are designed to vary based on pleasure (valence), activation, and dominance (PAD for brevity). From left to right first row: elated (PAD=1,1,1), afraid (-1,1,-1), surprised (1,1,-1), sad (-1,-1,-1). From left to right second row: angry (-1,1,1), relaxed (1,-1,-1), content(1,-1,1), frustrated (-1,-1,1). Top right displays PAD extremes, which serve as the basis for this research.

2 Results

Mean positioning results are displayed visually in figure 3, corresponding to the position that each word or emoticon sticker was placed averaged across all respondents. Missing data occurred in which students may not have placed every sticker. On average any given term or emoticon was missing 16.6 reports, with a maximum of 23 students of 81 missing reports for boredom, frustration, and relief. The average student was only missing 3.7 out of 18 terms and emoticons from their sheet, and there were 5 students who turned in completely blank sheets.

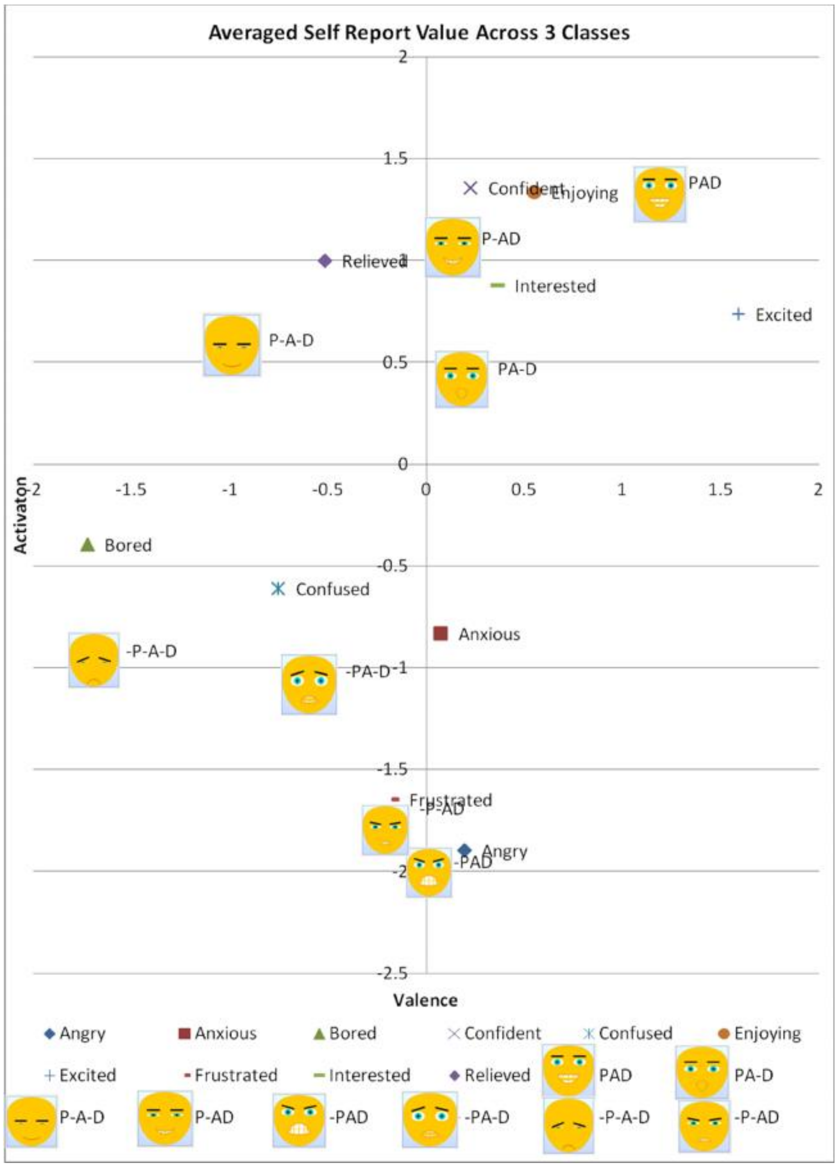

Fig. 3. Averaged Placement of Text and Emoticon Stickers

Interestingly, the placement of -PAD and -P-AD (negative sign indicating most extreme negative activation, pleasure, dominance, lack of a negative indicating most extreme positive, see figure 2 caption) match up with their respective terms "Angry" and "Frustrated" very closely. However, while both seem to be at the extreme end of negative valence, on average both seem to be viewed as fairly neutral in terms of activation by students. Although all emoticons and terms fall under the expected half of the coordinate axes in terms of valence (i.e. those we would expect to be pleasurable are categorized as above the origin, those displeasurable below it), activation does not follow this trend. For example anxiety is rated as neutral activation. One possible explanation, consistent with the results, is that students may be grouping activation and dominance together as a single measure. Emoticons with both negative activation and dominance were rated most negatively in terms of activation, those with either negative activation or dominance tended to fall in the middle, and the rating with all positive PAD was the emoticon with the highest rated activation.

Text or Emoticon	Activation Mean (StdDev)	Valence Mean (StdDev)
Angry	0.19 (1.09)	-1.9 (0.99)
Anxious	0.07 (1.78)	-0.87 (1.19)
Bored	-1.72 (1.28)	-0.4 (1.02)
Confident	0.23 (1.22)	1.35 (0.99)
Confused	-0.75 (1.36)	-0.61 (1.12)
Enjoying	0.55 (1.18)	1.34 (1.14)
Excited	1.59 (1.04)	0.74 (1.26)
Frustrated	-0.17 (1.33)	-1.65 (1.05)
Interested	0.36 (1.34)	0.88 (0.98)
Relieved	-0.52 (1.43)	1 (1.12)
Face_PAD	1.25 (1.3)	1.38 (1.13)
Face_PA-D	0.28 (1.86)	0.47 (0.93)
Face_P-A-D	-0.89 (1.57)	0.61 (0.91)
Face_P-AD	0.2 (1.26)	1.11 (1.08)
Face_-PAD	0.05 (0.95)	-1.95 (0.93)
Face_-PA-D	-0.5 (1.39)	-1.01 (1.01)
Face_-P-A-D	-1.61 (1.41)	-0.91 (1.11)
Face_-P-AD	-0.12 (1.15)	-1.69 (0.89)
Average	-0.08 (1.33)	-0.12 (1.05)

Table 1. Means and Standard Deviations of Students' placement of stickers.

One key goal of this work was to determine the degree of variance between students in terms of where they placed each term or emoticon. Given any affective term, there was little difference between the standard deviation for terms (mean S.D for terms = 1.20) and faces (mean S.D. for faces = 1.18). However, there was a larger difference between the standard deviation in activation (mean S.D for activation of terms or faces = 1.33) and valence (mean S.D for valence of terms or faces = 1.05), suggesting that students may have a greater degree of agreement in regarding rating the valence of affective representations than the activation it produces in them, which is consistent with the finding that affective representations fall on the division between positive and negative valence as we would categorize them, but not necessarily in terms of activation.

3 Discussion

The results presented in this article highlight a few different conclusions: a) students did not necessarily match emoticons or affective terms to the quadrants where researchers would have placed them, mostly in relation to activation; b) there is a large variation across these middle-school students in terms of where they placed a specific emotion within the axes of valence x arousal.

Characterizing researcher common expectations for arousal or activation is difficult, as many researchers only tentatively suggest how emotional states may be characterized in terms of activation. Pekrun found data to support boredom being somewhat deactivating, [18]. Russell [25] explores the components of affect and offers a few hypotheses which are summarized in figure 1 of Baker et al 2010 [3] wherein boredom is characterized as deactivating, while frustration, surprise, and delight are characterized as activating. Broekens' [8] emoticons follow the scheme outlined in the figure 2 caption: elation, fear, surprise, and anger are seen as activating, while sadness, relaxation, contentment, and frustration are seen as deactivating.

Students seem to agree that delight or elation is highly activating along with excitement, and boredom is deactivating along with sadness and relaxation. However, we found that students viewed an emoticon of fear as deactivating, and other affective states placed relatively close to neutral in terms of activation.

There are a few points of methodological concern. Firstly, the order that the students' place their stickers may be important: beyond a simple priming effect of considering one term/emoticon before another, by placing one item first students are changing the affordance of the coordinate axis itself by adding a milestone in the form of a term or emoticon. In future research, we could consider including fewer stimuli for placement or giving students a clean chart for each stimuli.

A second point of concern is one of validity. The terms, emoticons, and even the coordinate axis itself are abstract descriptors of affective states, which in this experiment are divorced from the actual experiences students may be having.

By placing our study outside the experimental environment we are likely reducing the validity of this work in exchange for simplicity of study design (i.e. not requiring students to respond with faces and words on the axis at various points in their experience).

The work of Bieg et al. [6] tells a much larger story than recommending against self-reports out of context. Out of context self-reports were found to bias in a consistent direction as compared to in context self-reports. However we maintain this method is “less valid” rather than “invalid”. Further, if we take into consideration the savings in class time an out of context self-report may actually be a better study design choice in some cases. It is our position that establishing more quantitative comparisons of reliability will yield better relative comparisons of validity and allow for improved study design.

This argument can be extended to affective research in general in the distinction between emotional experience and appraisal. We conceptualize the experience itself as the construct, and the cognitive appraisal process as a means of communicating that experience. The appraisal may be performed to send communication (e.g. having an experience and generating a representation of that experience for others), or receive communication (e.g. identify a representation as signifying an emotional state).

From this standpoint we suggest that the fewer steps of appraisal exist, the greater the face validity of an appraisal is in terms of reflecting an emotional experience. This is consistent with the findings of [6] wherein aggregate appraisal may differ from immediate contextual appraisal and we tend to view immediate appraisal as having greater face validity. This hypothesis also lends credence to the belief that external appraisal of an unconsciously generated representation (which may still be unconsciously meant to communicate an experience), in the form of facial expressions may be more valid than self-report measures wherein experiences are appraised by both subject and researcher. However, while passing through multiple appraisals may risk loss of information, the quality and richness of the appraisal may also play a role.

While validity remains very difficult to establish with regard to affect by testing “inter-method” or “representational” reliability perhaps we can building convergent and discriminant validity: multiple representations indicating the same construct across multiple participants. We maintain that reliability and validity are continuous rather than discrete traits of models. Therefore, we wish to reach consensus on methods of determining reliability and validity and then begin applying them to methods of inferring the experience of emotion. This work is a means of determining reliability between appraisals of representations of emotion rather than reliability of appraisals of emotions themselves. This is to say that matching particular facial expression to their personal lexicon of categorical affective terms, a high degree of agreement may validate the relationship between depictions of affect textually and facially, but not between either of those representations and the experience of an emotion.

A potential way towards greater validity and reliability could be to cognitively induce an emotional experience by asking students to respond to how they would feel given a particular situation (e.g. “Report on how you’d feel if you failed a math test.”). Of course there may be a distinction between induced affect and “organic” affect, further there will be a broad degree of subjectivity based on how individual students might feel about any given situation. Therefore the variance in responses could be attributed at least to two types of factors: those pertaining to both how students’ believe they would feel in a given context, and those pertaining to students’ ability to report that subjective experience through self-report measures. While there isn’t a clear way to disambiguate between which type of factor is responsible for the variance here, such an approach might be able to establish a conservative maximum of error in self-report measurements, because two students might have very different feelings about failing a math exam. In essence, we have measured variance in reliability here, not validity.

Finally, while reliability of self-report measures should inform their design, there may be cases of diminishing returns where a slight improvement in reliability has heavy costs for implementation workload, response time, or other practical concerns. We need not pick the measure with the highest available reliability; however it would be good to have some empirical handle on the relative reliabilities of different types of self-report measures. Perhaps the greatest thing to come out of this work would be future collaborations which might better address these concerns.

4 References

1. AlZoubi, O., D'Mello, S. K., & Calvo, R. A. (2012). Detecting naturalistic expressions of nonbasic affect using physiological signals. Affective Computing, IEEE Transactions on, 3(3), 298-310.
2. Arroyo, I., Woolf, B.P., Royer, J.M. and Tai, M. (2009b) 'Affective gendered learning companion', Proceedings of the International Conference on Artificial Intelligence and Education, IOS Press, pp.41–48.
3. Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C. (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.
4. Barrett, L. F. (2004). Feelings or Words? Understanding the Content in Self-Report Ratings of Experienced Emotion. Journal of Personality and Social Psychology, 87(2), 266–281.
5. Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., & Zhao, W. (2015). Automatic Detection of Learning-Centered Affective States in the Wild. In Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015). ACM, New York, NY, USA.
6. Bieg, M., Goetz, T., & Lipnevich, A.A. (2014). What Students Think They Feel Differs from What They Really Feel – Academic Self-Concept Moderates the Discrepancy between Students' Trait and State Emotional Self-Reports. PLoS ONE 9(3): e92563.
7. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. Journal of Behav Ther Exp Psychiatry, 25, 49-59.
8. Broekens, J., & Brinkman, W.-P. (2013). AffectButton: a method for reliable and valid affective support. International Journal of Human-Computer Studies, 71(6), 641-667.
9. Calvo, R. A., D'Mello, S., Gratch, J., & Kappas, A. (Eds.) (2015). The Oxford Handbook of Affective Computing. Oxford University Press: New York, NY.
10. Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction, 19(3), 267-303.
11. D'Mello, S. , Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. International Journal of Artificial Intelligence in Education, 16(1), 3-28.
12. D'Mello, S., & Graesser, A. C. (2012). Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, 5(4): 304–317.

13. Graesser, A., & D'Mello, S. (2011). Theoretical perspectives on affect and deep learning. In New perspectives on affect and learning technologies (pp. 11-21). Springer New York.
14. Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement. Introduction to the special issue. Contemporary Educational Psychology, 36, 1–3.
15. McDaniel, B. T., D'Mello, S., King, B. G., Chipman, P., Tapp, K., & Graesser, A. C. (2007). Facial features for affective state detection in learning environments. In Proceedings of the 29th Annual Cognitive Science Society (pp. 467-472).
16. Ocumpaugh, J., Baker, R.S., Rodrigo, M.M.T. (2015) Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.. Technical Report. New York, NY: Teachers College, Columbia University. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
17. Pardos, Z. A., Baker, R. S., San Pedro, M. O., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. Proc. 3rd Int.Conf. Learning Analytics & Knowledge, 117-124.
18. Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. Journal of Educational Psychology, 102(3), 531-549.
19. Porayska-Pomsta, K., Mavrikis, M., D'Mello, S., Conati, C., Baker, R.S.J.d. (2013) Knowledge Elicitation Methods for Affect Modeling in Education. International Journal of Artificial Intelligence in Education, 22 (3), 107-140.
20. Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. User Modeling and User-Adapted Interaction, 18(1-2), 125-173.
21. Read, J., McFarlane, S., and Cassey, C. (2002). Endurability, engagement and expectations: Measuring children's fun. In Proceedings of International Conference for Interaction Design and Children.
22. Read J. C. and MacFarlane, S.(2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In Proceedings of the 2006 conference on Interaction design and children (IDC '06). ACM, New York, NY, USA, 81-88.
23. Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., et al. (2007). Affect and Usage Choices in Simulation ProblemSolving Environments. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work (Vol. Frontiers in Artificial Intelligence and Applications 158). Amsterdam: IOS Press.
24. Rowe, J. P., Mott, B. W., & Lester, J. C. (2014) It's All About the Process: Building Sensor-Driven Emotion Detectors with GIFT. In Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2) (p. 135).
25. Russell J.A, Barrett L,F. (1999) Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. J. Pers. Soc. Psychol. 76(5):805–19.
26. San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. Proceedings of the 6th International Conference on Educational Data Mining, 177-184.