Student Emotions with an Edu-Game: A Detailed Analysis

Mirela Gutica
Department of Curriculum and Pedagogy,
The University of British Columbia,
2125 Main Mall, Vancouver, BC, V6T 1Z4, Canada

Cristina Conati
Department of Computer Science,
The University of British Columbia,
2366 Main Mall, Vancouver, BC, V6T 1Z4, Canada

***Abstract*—We present the results of a study that explored the emotions experienced by students during interaction with an educational game for math (Heroes of Math Island). Starting from emotion frameworks in affective computing and education, we considered a larger set of emotions than in related research. For emotion labeling, we employed a standard method that relies on trained judges to report emotions over 20-second intervals. However, we asked judges to report all observed emotions in each interval, as opposed to only choosing one, as is standard practice. This variation allows us to discuss the appropriateness of this interval for emotion labeling. We present a detailed analysis of inter-coder reliability, both aggregated and over individual students, that considers not only the matching by judges over emotion type, but also the number of emotions detected.**

Keywords—affective states; learning; educational games; emotion labeling; inter-judge agreement

I. INTRODUCTION

In recent years, Intelligent Tutoring Systems (ITS) and game-based learning environments have attracted interest as technologies that harness motivation and support learning. Research has focused not only on the cognitive aspects of interaction, but also on affect recognition and response. There is increasing evidence that, to design an intelligent responsive tutor, the learner's emotions should be properly identified (e.g., [1, 2]). This paper focuses on the identification of emotions triggered during students' interaction with an educational game for mathematics, Heroes of Math Island (the term edu-game will be used in this paper), with the long-term goal of making the game capable of detecting and responding to these emotions. There has been extensive work on identifying and detecting emotions elicited by educational software [1, 2, 3, 4, 5, 6, 7, 8]. This paper adds to this work and contributes to the affective computing field by: (1) considering a larger set of emotions derived from a literature review and pilot studies, which includes emotions like *confidence,* that are considered important for learning [9]; (2) analyzing inter-judge agreement at the individual student level, to gain a better understanding of if and how individual differences in emotion expression can affect emotion recognition; and (3) looking in detail at how many emotions actually happened in a standard 20-second interval used in literature [2, 3, 5, 7, 8] and how easy it is for judges to identify them. In the remaining sections of the paper, we first discuss related work. Next, we present Heroes of Math Island, the game we used as a test-bed in this research. We then illustrate the study we ran and the protocol we used for emotion identification. Finally, we report on the results of emotion analysis.

II. RELATED WORK

Classifying, defining, and identifying emotion is controversial and challenging (e.g., [10, 11]). Cognitive-appraisal theories of emotion agree that emotions are triggered by events that elicit thinking and, in the end, arouse emotion [10, 11]. From the educational literature, we know that some emotions can be beneficial for learning because they drive attention, which in turn drives memorization, but we do not know exactly how to regulate emotions in learning activities [9, 12]. Several affect frameworks have been proposed for emotion detection, many of them based on Ekman's taxonomy of emotion [13]. In the context of detecting emotions during interaction with educational software, Conati et al. [1, 4] have proposed a framework that models the set of emotions taken from the Ortony, Clore and Collins (OCC) theory [14] (*admiration, joy, regret, reproach, pride* and *shame)*. Other work [3, 7, 8] has relied on a framework proposed by Graesser et al. [2, 5], which includes *boredom, confusion, delight, flow* (also named *engaged concentration*), *frustration,* and *surprise* as emotions considered relevant for learning. Using this framework, Graesser et al. [5] found that emotion classification performed by trained judges was more reliable than classification performed by peer judges. The same framework has been used by various researchers to measure the likelihood of transition from one affective state to another. D'Mello, Tyler and Graesser [2] focused on students interacting with a dialogue-based ITS. McQuiggan, Robison and Lester [6] investigated the likelihood of affective transitions in a narrative-centered edu-game (Crystal Island), showing that *engaged concentration* was the predominant state for learning. Baker et al. [3] studied the incidence, persistence, and impact of these emotions with three different learning environments. They found that *boredom* was the most persistent state in each environment; however, *engaged concentration* was the most frequent state, followed by *confusion*. Rodrigo et al. [7] found that two learning environments, Ecolab and M-Ecolab, were not able to disrupt the persistence of *boredom* and *frustration*. Researchers have adopted a variety of approaches for emotion identification, including video annotations [2, 5], quantitative field observation [3, 7, 8], self-reports [4, 6], and sensors [1]. In terms of how long the interval of observation should be, many studies used a 20-second interval [2, 3, 5, 7, 8].

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.94

III. HEROES OF MATH ISLAND

Fig. 1. Sample activity from Heroes of Math Island

The Heroes of Math Island game is designed for Grades 5 to 7 students. The game has a narrative and activities that occur on an island employing a castle as a central site, where students get "*quests*" (a hero's journey towards a goal) from a king or queen. Game-based learning in a rich interactive game environment employing activities on an *island* was also explored by Lester et al. for the Crystal Island game [6]. The idea of an island (also found in several commercial games, e.g., World of Warcraft) is related to adventure on an enclosed imaginary territory. The quest currently implemented in Heroes of Math Island targets three math learning outcomes: *divisibility, prime numbers* and *number decomposition*. The metaphor used in the quest is that of a mine (see Fig. 1): prime numbers are hard rocks that cannot be broken; composite numbers can be broken with picks. Once in the mine, students solve problems generated by the system. The level of difficulty increases based on student performance. The game provides progressive hints to help students overcome errors, similar to the Prime Climb edu-game for number factorization [1, 4]. At the first error of a kind, a hint appears acknowledging the error: e.g., "You picked 10, and that is incorrect." If a similar error is repeated, the next hint suggests what the student should be looking for: e.g., "There are still some rocks with prime numbers that you can pick." At the third error, an example of how to solve the problem is given.

Similar to [8], Heroes of Math Island includes an affective agent (personified by a monkey character) that uses emotional expressions to respond to situations in the game. These expressions (shown in Fig. 2) include a *neutral* state, two positive states (*happy* and *confident*) and two negative states (*sad* and *frustrated*). These expressions were designed based on standard depictions of emotions used in schools for helping students deal with their feelings and for conflict resolution, as well as in hospitals and in therapy [15]. The emotional state displayed by the monkey is based on a success score calculated from both an absolute score (number of mistakes minus number of correct responses) and the trend of the most recent actions. The monkey begins in the *neutral* state. If the student makes a mistake, the monkey displays *sadness,* and if the trend continues, *frustration*. From this state, if the student starts to improve, the monkey goes first into the *neutral* state, then into *happiness,* and ultimately into the *confident* state. We wanted the monkey to be encouraging; hence, it is slightly smiling in the *neutral* state.

Fig. 2. Monkey's emotions from left: frustrated, sad, neutral, happy and confident

IV. EMPIRICAL STUDY OF HEROES OF MATH ISLAND

We conducted an empirical evaluation of Heroes of Math Island designed both for understanding if and how the game stimulates learning and for analyzing the students' emotional reactions during game play. This paper focuses on this second objective; therefore, we provide a brief summary of the study design. The participants in this study were students aged 11 or 12. Each study session started with a short tutorial on the relevant math topics, to ensure that students started at comparable levels of knowledge. Next, students took a math pre-test, followed by game play, then a post-test, and finally a post-questionnaire for us to obtain their feedback on various aspects of the game. Pre- and post-tests were analogous and contained 23 questions. The duration of a study session ranged between 1½ and 2½ hours, of which game play was from 15 to 48 min (M = 32.3 min; SD = 10.3 min). The game interaction was videotaped: one video camera recorded the face of the student and one the computer screen. Afterwards, the two videos were merged and synchronized.

V. EMOTION LABELING PROCESS

In this study, the process of labeling emotions from the data collected took several iterations and consisted of the following phases:

- *Phase 1*: Definition of an initial set of emotions that can be relevant to learning with an edu-game, based on existing literature in education, emotional psychology and ITS.
- *Phase 2*: Pilot study to ascertain the adequacy of the emotions set from Phase 1. This pilot was also used as a first training session for the judges used in the analysis.
- *Phase 3*: Pilot study to finalize the emotion labeling process based on a variation of a standard protocol that involves reporting emotions during 20-second time intervals.

Phase 1 To define our initial set of emotions, we looked at the following emotion models: (1) the affect framework proposed by Graesser et al. [2, 5], which considers the following emotions: *boredom, confusion, delight, engaged concentration* (also known as *flow*)*, frustration* and *surprise*; (2) models found in the education literature: Astleitner's [12] model of emotions in the context of instruction (*anger, envy, fear, pleasure* and *sympathy*); (3) Ingleton's emotion model in learning mathematics [9] (*confidence, distance, fear, pride, shame* and *solidarity*); and (4) the OCC cognitive theory of emotions [14]. From each model, we selected the emotions that we thought would be relevant for our task and that could form a reasonable set to pilot test. From Graesser's framework [2, 5]we selected *boredom, confusion, delight, engaged concentration* and *frustration.* Note that *delight* is equivalent to what Astleitner [12] and OCC [14] refer to as *pleasure*, so we will refer to this emotion with both names from now on.

Surprise was not included in the first set of emotions because we wanted to focus on emotions only, and *surprise* is not considered an emotion by some authors [10, 16]. Ortony and Turner [16] considered that being "affectively valenced is a necessary condition for a state to be an emotion" and argued that *surprise* can be better qualified as a cognitive than an affective state (p. 317). From Ingleton's model [9], we selected *confidence*, *pride* and *shame*. Ingleton argues that *confidence* creates a "disposition to learn," as opposed to negative emotions such as anxiety, grief and dejection, which can prevent learning and lead to inactivity and isolation (p. 88). *Pride* and *shame* play a role in identity and self-esteem which influence the formation of *confidence* [9]. *Pride* and *shame* are also part of the OCC model [14] and were explored in studies with the PrimeClimb edu-game [1, 4]. Furthermore, they may be especially relevant in our game because the monkey is showing emotions that are explicit reactions to the student's game performance. From Ingleton's model [9] we excluded *anger*, *distance, fear* and *solidarity.* A*nger* (also found in the OCC and Astleitner's models) was excluded because we felt that the already included negative emotion of *frustration* is more appropriate for our study. *Fear* (also present in the OCC and Astleitner's models) was excluded because it is described as "fear of failure" in the context of learning [12] (p. 212), and our study participants were in a relaxed environment without any school pressure. *Distance* and *solidarity* from Ingleton's model [9] were excluded because we judged them to be more suited to classroom instruction. From the Astleitner's model [12] we only included *pleasure*, because the remaining emotions of *envy* and *sympathy* were also considered more suited for classroom instruction. The OCC model [14] contains 22 emotions in three categories: (1) emotions resulting from consequences of events, (2) actions of agents, and (3) aspects of objects. We did not consider reactions to aspects of objects because we felt that there were no objects in our math game that can trigger substantial emotions. Regarding the emotions resulting from consequence of events for self, we assumed that, for this initial round, the already included positive emotion of *delight* and negative emotion of *frustration* could cover the OCC emotions of *joy*, *pleasure*, *satisfaction*, *displeasure*, *disappointment* and *distress*. With respect to emotions for agents other than self (e.g., *admiration*, *gratitude, remorse, and reproach*), we decided to exclude them at this stage because the only agent in the game is the monkey, and it does not perform any game actions that can directly help or hinder the student's game performance. The monkey's affective displays may trigger emotions such as *reproach*, but this did not seem to be the case, based on informal observations of initial game sessions. The model included *neutral* to give judges a way to report no emotion, as opposed to a situation of invalid interval or missing data. This process resulted in eight emotions (plus *neutral*) to be pilot-tested in the second phase: *boredom, confidence, confusion, delight/pleasure, engaged concentration, frustration, pride* and *shame.* An initial instrument for emotion reporting was designed with vertical rubrics for each emotion in the set above, plus an entry for observed emotions not listed in the set. As part of our protocol for the emotion *judgment*, we adapted from Baker et al. [3] an observer's guide that had a short description for each emotion. The labeling system and the instrument's format were evaluated in several sessions, during which three observers (one of the co-authors of this paper and student research assistants) met, watched videos of pilot game segments on a big screen, and refined the material as needed.

Phase 2 Emotion labeling in similar studies (e.g., [2, 3, 5, 7, 8]) involved trained judges assessing emotions occurring during 20-second intervals. In this phase we wanted a simpler process because our goal was to see which emotions occurred during game play, but not a precise account of when and how often. Observers reported emotions during three-minute intervals from additional videos of pilot game segments. They reported instances of all emotions from the given list, although observers more frequently reported *engaged concentration*, *confusion* and *confidence*. However, also reported were emotions not included in the original set: *curiosity, surprise*, and *tentative* (also described by observers as *hesitancy*). Therefore, we extended the set of emotions by adding these three. We agreed that *surprise* should be added because, even if it is not consistently considered an emotion, it was used in several important previous emotion studies [2, 3, 5, 7, 8, 13] and was reported by observers. To summarize, the outcome of this phase was a revised set of 11 emotions (plus *neutral*): *boredom, confidence, confusion, curiosity, delight/pleasure, engaged concentration, frustration, hesitancy, pride, shame* and *surprise*. This pilot phase also confirmed that, as expected, too many emotions happen in a three-minute interval. Thus, for the subsequent phases, we adopted the 20-second interval used in previous studies of emotions during interaction with educational software. It should be noted, however, that in previous work judges were asked to report only one emotion per interval, even when more than one was observed. In contrast, we decided to allow judges to report all observed emotions, to have a better sense of whether 20-second intervals represent an adequate granularity for this process.

Phase 3 To increase the observers' confidence with the emotion labeling process, we conducted more of the group observations and brain-storming sessions conducted in Phase 1. Observers together examined three 8-minute slices of pilot videos from different students, reporting all affective states that were noted during each 20-second interval, and revised the observer`s guide. After observing the video once, they observed it again from the beginning, with the intention of correction. Based on feedback from the observers, we revised the emotions set as follows. We added rubrics for: (1) a negative emotion of *disappointment* or *displeasure* (also found in the OCC model), and (2) *excitement* (also found in emotion studies with the Cristal Island game [6]), to allow observers to report separately situations similar to *delight/pleasure* but of a higher intensity. We also merged *confusion* and *hesitancy* because it was difficult to differentiate between the two. The final set included several emotions that are joined and described with two names

because they are hard to discriminate (e.g., *confusion/hesitancy*) and to avoid too-fine of a granularity. Our final choice of affective states includes 12 emotions (plus *neutral*): *boredom, confidence, confusion/hesitancy, curiosity, delight/pleasure, disappointment/displeasure, engaged concentration, excitement, frustration, pride, shame* and *surprise.*

VI. RESULTS

The affective states of 15 study participants were classified by two judges from the original team of observers: one of the paper's co-authors and a graduate student assistant. It should be noted that, for these 15 students, there was a significant improvement from pre-test (M = 77.7%; SD = 9.3%) to post-test (M = 83.5%; SD = 8.7%), t (14) = 2.2; two-tailed $p < 0.007$, indicating that they did learn from the interaction with Heroes of Math Island.

A total of 1082 20-second data intervals were analyzed for the 15 students. We found that all emotions in the emotion set were present in the game interaction (see Fig. 3).

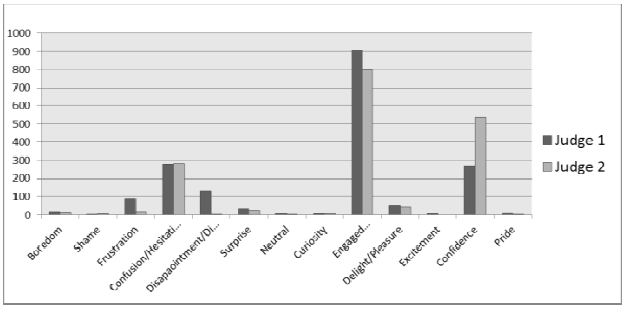

Fig. 3. Judges report on emotions: raw agreement

The least frequent emotions (below 1.2%) were: *curiosity*, *excitement*, *neutral*, *pride* and *shame. Boredom*, negatively associated to learning [2], was also rarely reported (1.7% by Judge 1 and 1.4% by Judge 2), followed by *surprise* (3.1% by Judge 1 and 2.3% by Judge 2), *delight/pleasur*e (5% by Judge 1 and 3.8% by Judge 2), and *confusion/hesitation* (26% by both judges). *Engaged concentration* was the state where students spent the majority of the time (83.9% reported by Judge 1 and 74.4% by Judge 2).

While the frequency of reports by the two judges for the emotions listed so far are quite similar, there are noticeable differences with respect to: (1) *confidence* (for which Judge 1 reported 24.9% and Judge 2 reported 49.5%; (2) *displeasure/disappointment* (for which Judge 1 reported 12% and Judge 2 reported only 0.3%); and (3) *frustration (*for which Judge 1 reported 8.1% and Judge 2 reported 1.6%). Both judges carefully followed the protocol and were very conscientious; however, Judge 1 is an experienced educator whereas Judge 2 is a graduate student without experience in teaching. We believe this may be the reason for Judge 2's higher tendency to interpret students' behaviors in terms of positive affect (more discussion on this point will follow in a subsequent section). In the next subsections, we provide a more detailed analysis of inter-coder agreement, at different levels of granularity.

A. Agreement over one emotion per interval

In this analysis, we discuss the level of agreement between the two judges, where only one emotion is selected per interval, even when several emotions were reported, to mimic the assumption made in previous emotion studies relying on the 20-second interval approach. We report Cohen`s Kappa scores for each student and for the aggregated data over all 15 students. When several emotions were reported per data point, only one emotion was taken into consideration to build the confusion matrix for agreement/disagreement, as follows. If there was agreement on one emotion only, that emotion was selected. If there was agreement over more than one emotion, one of the more prominent agreed-upon emotions was selected if possible (similar to [5]); otherwise, one was randomly chosen, unless one of them was *engaged concentration*. In that case, *engaged concentration* was excluded from the selection, because *engaged concentration* was observed more than other states. If there was no agreement over the emotions in the interval, the pair most likely to be mixed up was selected (e.g., if Judge 1 indicated *confusion* and *curiosity* and Judge 2 indicated *frustration* and *boredom*, we have chosen *confusion* and *frustration* because they are more likely to be mixed up)[1].The aggregated Cohen's Kappa was 0.676. Although there is no unified criterion to interpret Kappa values in the literature, values in the 0.6-0.7 range are generally considered good [17] or even substantial [18]. The best values of Kappa achieved in previous studies were: 0.63 [3], 0.68 [8], 0.71 [5], and 0.73 [7].

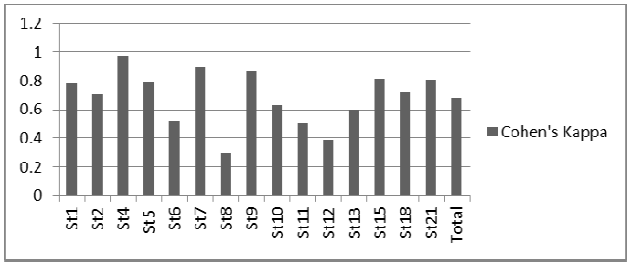

Fig. 4. Cohen's Kappa for individual students

In addition to the aggregated Kappa, we also computed Kappa scores for each individual student (Fig. 4), to ascertain the impact of individual differences in the observed subjects on the reliability of emotion labeling. As the figure shows, the inter-judge reliability was quite varied for some students, although it was substantial on average (M = 0.688, SD = 0.19). Five students had Kappa values at or above 0.8 (considered excellent), whereas four had values generally considered low (Student 8 with 0.297 and Student 12 with 0.388), or moderate (Student 11 with 0.504 and Student 6 with 0.526).

Table I shows the confusion matrix on the aggregated data, to give a sense of which emotions where harder to discriminate. Reports from Judge 1 are in the rows, and those for Judge 2 in the columns. Note that *excitement* is not in the matrix, because it always appears in intervals with other

[1] In most previous studies, judges reported the first emotion observed in an interval. We could not use observation time as a selection criterion, because we had no information on the sequencing of emotions *within* each interval.

emotions, with no agreement, and was never picked by our selection criteria.

TABLE I.

	Bo	Sh	Fr	CH	Dis	Su	Ne	Cu	Eng	PD	Co	Pr	Total
Bo	1	0	0	2	0	0	0	0	2	0	2	0	7
Sh	0	0	0	0	0	0	0	0	0	0	0	0	0
Fr	0	0	9	6	0	0	0	0	10	0	0	0	25
CH	1	1	0	116	0	1	1	0	34	0	19	0	171
Dis	3	0	0	4	1	0	0	0	7	0	3	0	18
Su	0	0	0	0	0	8	0	1	0	0	1	0	9
Ne	0	1	0	1	0	0	1	0	1	0	2	0	6
Cu	0	0	0	0	0	0	0	2	1	0	1	0	4
Eng	0	0	0	7	0	1	1	0	530	2	100	0	641
PD	0	0	0	0	0	0	0	0	1	23	1	0	24
Co	0	0	0	0	0	0	0	0	0	0	175	0	175
Pr	0	0	0	0	0	0	0	0	0	0	1	1	3
Total	5	2	9	136	1	10	2	2	586	25	304	1	1082

As we discussed earlier, Judge 2 (the graduate student) generated many more reports of *confidence* than Judge 1 (the experienced educator). The highest source of disagreement (100 instances) is between Judge 2 reporting *confidence* and Judge 1 reporting *engaged concentration*, a rather intuitive outcome since both states involve a positive attitude towards and an active involvement with the task at hand (*engaged concentration* is defined as immersion and focus, and *confidence* as knowing what to do and solving problems fast). More surprising are the situations in which Judge 2 reported a positive valenced emotion and Judge 1 reported a negative valenced one: there were 19 instances in which Judge 2 reported *confidence* and Judge 1 reported *confusion/hesitation*, 34 instances in which Judge 2 reported *engaged concentration* and Judge 1 reported *confusion/hesitation,* and 10 instances in which Judge 2 reported *engaged concentration* and Judge 1 reported *frustration.* We interpret these differences as due to Judge 1's experience as an educator, possibly resulting in a higher ability to detect affective signs of students having difficulties during learning. However, this result should be further explored in future studies.

B. Agreement over multiple emotions per interval

In this section, we look at all the emotions reported in 20-second intervals. Fig. 5 summarizes the frequency with which different numbers of emotions were reported per interval, showing that intervals with two emotions are almost as frequent as intervals with one emotion. There is a non-negligible number of intervals with three emotions, and even a small percentage of intervals with four emotions. Table II shows a confusion matrix indicating how many times the judges reported 1, 2, 3 or 4 emotions for the same interval. The Cohen's Kappa is negative (K = -0.011), indicating less than chance agreement between judges on this point.

We also looked at how often the judges agreed on the emotion type for 2, 3 and 4 emotions, regardless of how many they disagreed upon. There were 131 (12.1%) intervals in which judges agreed on the type of 2 emotions, 12 intervals (1.1%) with agreement over 3, and no interval with agreement over 4.

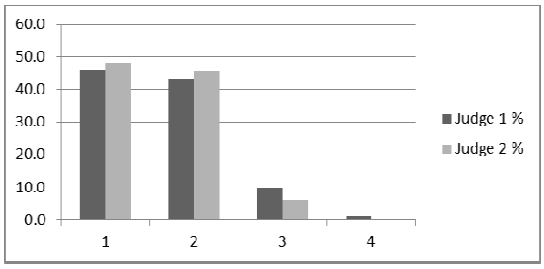

Fig. 5. Emotions reported per interval

TABLE II.

	1	2	3	4	Total
1	241	233	21	1	496
2	238	197	29	2	466
3	38	58	9	0	105
4	3	3	6	0	15
Total	520	494	65	3	1082

Looking at matches over individual students (see Fig. 6), there is a very high variance on raw agreement (M = 12.8%, SD = 8.5 for 2 matches; M = 1.2%, SD = 1.9 for 3), once again indicating that subject differences can affect accuracy of emotion labeling.

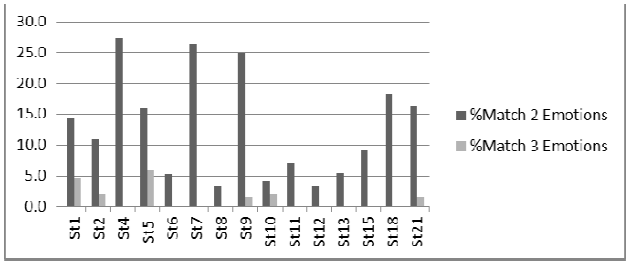

Fig. 6. Agreement for 2 or 3 emotions per interval for individual students

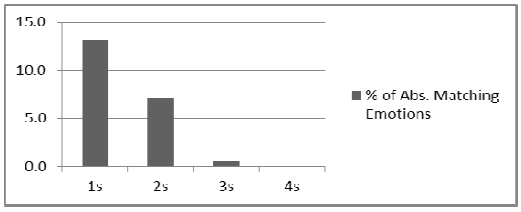

Fig. 7. Procentage of total instances when judges agreed on emotion type and number for 1, 2, 3, or 4 emotions

Finally, we looked at the instances when judges matched in both number and type of emotion (Fig. 7). In 225 cases (20.8%), there was a perfect match, with a higher occurrence for one emotion (142 intervals, or 13.1%), followed by two emotions (77 intervals, or 7.1%) and three emotions (6 intervals or 0.6 %). There was again a high variance between students (Fig. 8). Student 2 showed the most absolute matches (38%), followed by Student 1 (33.3%) and Student 9 (31.3%); the least are for Student 8 (2.2%), Student 12 (6.7%) and Student 11 (10.5%).

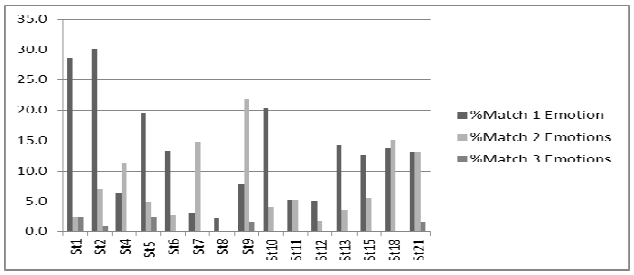

Fig. 8. Absolute agreement for 1, 2, or 3 emotions for individual students

VII. Discussion and Conclusions

This paper presented work that contributes to research on identifying the emotions arising during interaction with edu-games (here Heroes of Math Island), with the long-term goal of tracking and responding to these emotions in real time. Our contributions include: (1) considering a larger set of emotions compared to previous studies, including emotions like *confidence, curiosity, pride* and *shame*; (2) modifying a standard method for emotion labeling by asking judges to report all observed emotions in 20-second intervals, as opposed to only one as done in previous work; and (3) presenting a detailed analysis of inter-coder reliability both aggregated and over individual students (not done before). The analysis considers not only the matching by judges over emotion type, but also the number of emotions detected. Our results show that students' performance on ad-hoc math tests improved due to game playing, and these results are reflected in the observed affective states: a lack of reported *boredom* and high *engaged concentration*. Low levels of *boredom* (in the 3% to 7% range) have been reported by [3], [6] and [8]. Our percentages for *boredom* are even lower (less than 1.5% averaged between the two judges). Several studies showed *engaged concentration* to be the most reported state (68% in [3], 67.4% in [7], 43.45% in [8] and 42% in [6]). We found the same result, but in our study *engaged concentration* appeared even more frequently, with 79% of raw instances between the two judges. Our second most frequently reported emotion was *confidence*, one of the new emotions that we considered compared to previous studies. The third most frequently reported state was *confusion/hesitation*, similar to [3], which reports *confusion* to be the second most frequently observed emotion. The frequency of confusion in our study is actually higher than in [3] (about 26% instead of 13%), but this finding is still consistent with the observed positive learning outcomes, since confusion is considered an emotional state that can trigger learning [2, 5]. We found a high variance of reporting agreement over different students. Although it is common knowledge that different people have different propensity for showing their emotions, our results indicate that these differences can be quite substantial for some individuals, and may make it difficult to obtain reliable ground truth emotion information and subsequent accurate affective models. Another interesting finding is that the 20-second intervals used for emotion reporting often included more than one emotion. Studies so far have adopted the simplified approach of only considering one emotion per interval, but this may not necessarily be the most relevant for learning. Ignoring the other emotions too often may result in an inaccurate account of which emotions an edu-game should be able to detect and respond to. Our results, however, showed low inter-coder reliability on the number of observed emotions per interval, and few instances in which coders agreed on emotion type when more than one was present. This suggests that it will be challenging to develop methods for emotion analysis at a finer level of granularity, but it is an endeavor worthy of exploration.

References

[1] C. Conati, "Probabilistic Assessment of User's Emotions in Educational Games"*J. of Applied Artificial Intelligence, special issue on Merging Cognition and Affect in HCI,* vol. 16, no. 7-8, 2002.

[2] S. K. D'Mello, R. Taylor and A. C. Graesser, "Monitoring affective trajectories during complex learning," *29th Annual Cognitive Science Soc.,* pp. 203-208, 2007.

[3] R. Baker, S. D'Mello, M. Rodrigo and A. Graesser, "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," *Int. J. of Human-Computer Studies,* vol. 68, no. 4, 2010.

[4] C. Conati and H. Maclaren, "Modeling user affect from causes and effects," in *Proc. of the 17th Int. Conf. on User Modeling, Adaptation, and Personalization*, Berlin, Heidelberg, 2009.

[5] A. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello and B. Gholson, "Detection of emotions during learning with AutoTutor," in *Proc. of the 28th Annual Meetings of the Cognitive Science Soc.*, NJ:Erlbaum, 2006.

[6] S. McQuiggan, J. Robison and J. Lester, "Affective transitions in narrative-centered learning environments," in *In Proc. of the 9th Int. Conf. on Intelligent Tutoring Systems*, 2008.

[7] M. Rodrigo, G. Rebolledo-Mendez, R. Baker, B. d. Boulay, J. O. Sugay, S. Lim, M. B. E. Lahoz and R. Luckin, "The Effects of Motivational Modeling on Affect in an Intelligent Tutoring System," *International Conf. on Computers in Education,* 2008.

[8] M. Rodrigo, R. de Baker, J. Agapito, J. Nabos, M. C. Repalam, S. S. Reyes and M. O. C. Z. San Pedro, "The Effects of an Interactive Software Agent on Student Affective Dynamics while Using an Intelligent Tutoring System," *Affective Computing,* vol. 3, no. 2, pp. 224-236, 2012.

[9] C. Ingleton, "Emotion in learning - a neglected dynamic," *Research and Development in Higher Education,* vol. 22, pp. 86-99, 2000.

[10] R. S. Lazarus, "On the primacy of cognition," *American Psychologist,* vol. 39, pp. 124-129, 1984.

[11] J. E. LeDoux, "Emotion: Clues from the Brain," *Annual Reviews Psychology,* vol. 46, pp. 209-235, 1995.

[12] H. Astleitner, "Designing emotionally sound instruction: The FEASP-approach," *Instructional Science,* vol. 28, p. 169–198, 2000.

[13] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power , Eds., New York, Wiley, 1999, pp. 45-60.

[14] A. Ortony, G. L. Clore and A. Collins, The cognitive structure of emotions, Cambridge: Cambridge Univ. Press, 1988.

[15] "Creative Therapy Associates," [Online]. Available: http://www.ctherapy.com/Product_Home_Pages/feelings.asp.

[16] A. Ortony and T. J. Turner, "What's basic about basic emotions?," *Psychological Review,* vol. 97, pp. 315-331, 1990.

[17] J. Fleiss, Statistical methods for rates and proportions, 2nd ed., New York: John Wiley, 1981.

[18] J. Landis and G. Koch , "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, no. 1, pp. 159-174, 1977.