Affect detection in home-based educational software for young children

Roger Smeets
Squla
roger@squla.com

Francette Broekman
Squla
francette@squla.com

Eric Bouwers
Squla
eric@squla.com

ABSTRACT

Research on automated affect detection in educational software using play log data has shown promising results. Yet most studies use classroom-based software designed for adolescents or adults. In this paper, we aim to detect affection in an online educational platform primarily aimed at home use by young children. This presents two challenges: we have to rely on a self-report instrument of affect that users can utilize at home, and we have to make sure that this instrument is properly understood by children. To this end, we developed and validated an emoticon-based self-report instrument to derive ground-truth labels of four emotions: Joy, frustration, confusion, and boredom. Training a number of different classifiers for automated affect detection yields promising results, in particular for detecting joy and frustration.

1. INTRODUCTION

It is by now well established that student engagement is an important correlate of learning efficacy, academic performance, and even long-term professional achievements [1, 2]. Consequently, monitoring student engagement is of key interest, so that appropriate interventions can be applied when necessary. Educational software holds a strong promise in this regard, as it enables the collection of troves of data that can be mined for engagement and learning patterns.

Previous research has demonstrated that student engagement is a multifaceted concept, typically encompassing behavioral, cognitive, and emotional aspects [5–7, 14]. There is a large body of empirical literature that aims to identify these facets of engagement during educational software use. Emotional engagement – or affect detection – in particular has garnered a lot of recent interest. Previous studies in this area have collected data from sensors and play logs to construct and identify metrics that can detect different types of emotions (and changes therein) [1, 3–5, 12].

Our objective in the current paper is to explore the possibility for affect detection based on play logs in a particular online learning platform, called Squla. This platform embodies two traits that inhibit a straightforward application of previous research. First, Squla is primarily developed for use at home, whereas virtually all previous research considers educational software used in the classroom. Second, Squla is used by children aged 4-12, whereas previous studies mostly consider adolescents or adults.

In order to train models for automated affect detection, we first have to collect instances (i.e. ‘ground truth labels’) of students’ emotions during software use. Ideally, these data are collected during normal use of the product, i.e. when students play at home. This in turn requires a self-report instrument that can be understood and used by children. In particular, the instrument should take into account that young children cannot read, and that identifying their own emotions might be problematic.

To address these issues, we first design and validate a self-report instrument of affect that is based on emoticons. In particular, we aim to identify four emotions: Joy, boredom, confusion, and frustration [1, 3]. We then use this instrument to collect ground truth labels of affect during software use. Finally, we construct a number of features from the play logs and correlate these with the labeled emotions, building four separate affect detectors (one for each emotion).

2. THE ONLINE LEARNING PLATFORM

The software used in this study is an online educational gaming platform for K-6 children (ages 4-12) in the Netherlands, called Squla. Squla’s primary business model is aimed at home-use, i.e. paid subscriptions are sold to parents. The platform is available through web as well as via a native app (iOS and Android). It is set up primarily along 8 education classes that correspond with the grades in kindergarten (KDG), pre-school (pre-K), and elementary school. Children can play a variety of subjects, both curriculum based (e.g. math or spelling) as well as outside the curriculum (e.g. social skills or 21st century skill).

Within a subject, students can choose a topic, within which they can choose a mission, that itself is typically subdivided into multiple (hierarchical) levels.[1] One level typically consists of ten questions that have to be answered correctly be-

Roger Smeets, Francette Broekman and Eric Bouwers "Affect detection in home-based educational software for young children" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 760 - 765

[1] For example, grade 3 has the subject geography, containing the topic ‘weather’, which has two missions. The first is ‘seasons’, consisting of five levels, and the second is an instruction video about rain.

fore the level is completed and the next one is unlocked. Importantly, students have complete freedom regarding what to play. The only restriction is that within a mission, levels have to be played in a predetermined order.

There are various different question formats, such as multiple choice questions, puzzles, open-answer questions, bubble poppers (popping foam bubbles containing answers), and catapult questions (shoot at answers with a catapult). If a question is answered incorrectly, the right answer is highlighted, and the question is moved to the end of the queue, so that the user has to answer it again after cycling through the remaining questions. At any point during playtime, a user can choose to quit the current level. She can always go back to that level and pick up where she left off.

When answering a question correctly, the user is rewarded with a coin. There are also numerous ways to earn bonus coins, e.g. when participating in a thematic campaign. Coins can be used to buy new avatars, or non-digital goodies from the goodie shop (e.g. cinema vouchers, toy-store vouchers, or paper-cut animals). During play, users also collect experience points, which gradually increases their experience level. Each experience level unlocks new benefits, such as new avatars in the avatar shop.

When users log onto the platform, they land on the home screen depicted in Figure 1. Here, they can navigate between the different subjects, but also to their playing statistics, to the Squla shop – where they can buy avatars and goodies with the collected coins – and to the ‘social’ screen – where they can search for friends, send and accept friend requests, and send or read messages to and from friends and parents. The top of the screen shows their experience level, their experience points, and the number of coins collected. Finally, the home screen also offers play recommendations in the three tiles in the top bar.

3. PREVIOUS RESEARCH

3.1 Log-based affect detection

Previous research on automated affect detection has used both sensor data and play log-based data to find correlates of affect. The educational software used in this study does not utilize sensors, so that we have to rely on the latter approach. Below we briefly review a number of studies that have successfully done so before.

In a study involving an automated tutoring system, D’mello et al. [5] use a combination of conversation-based log metrics and self and third-person reports – by one peer and two expert judges – of affective states. Their results show that machine predictions of affect are on par with those of ‘novice’ judges (self and peer) but inferior to those of the trained judges.

Contati and MacLaren [4] and Sabourin et al. [12] adopt a (dynamic) Bayesian network approach to develop a causal model linking user goals and (inter)actions to affective states. The resulting models significantly outperform default class prediction, and further allow the researchers to infer user play-goals.

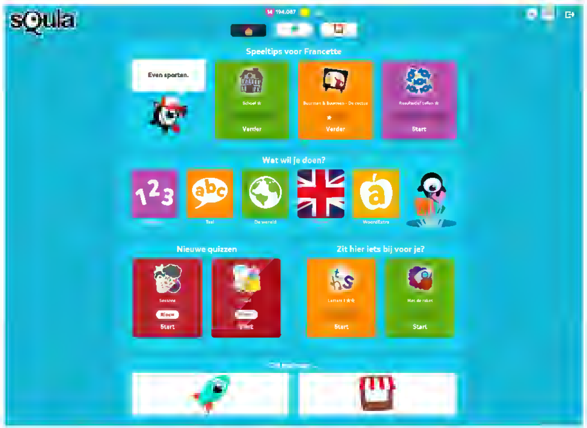

Figure 1: Squla home screen (web)

Lee et al. [8] show that persistent confusion negatively correlates with learning achievement, yet that resolved confusion has a positive effect. Botelho et al. [3] resort to Recurrent Neural Networks to improve the performance of these models. Unlike previous studies, they combine all emotions in one detector, rather than training one detector per emotion.

3.2 Affect detection for young children

All the studies discussed above share two traits: First, they all involve educational software that is used in a classroom setting, and second, the study populations are usually either adolescents or adults. Yet in this paper we consider educational software designed for home use by children. This means that collecting data in a controlled setting – such as a classroom or lab – would interfere too much with the intended user experience. As a result, affect labeling has to happen at home, using self-reports. Yet this creates another challenge, which is that (young) children cannot be expected to provide reliable survey feedback [9].

A number of studies have developed instruments for usability testing of software products with children [10, 11]. This research stream restricts the notion of child engagement or satisfaction to “fun”. One validated instrument that has been successfully applied in this context is the *smiley-o-meter*. This is essentially a 5-point Likert scale with emoticons capturing the intensity of fun (vs lack of fun).

However, Padilla-Zea et al. [13] find that this instrument is not properly understood by young children (ages 3-5), since they are not able to grade the strength or intensity of a feeling. Yet their tests demonstrate that young children *are* able to differentiate between different *types* of emotions, based on different emoticons.

Our approach toward affect labeling mirrors that of Padilla-Zea et al.[13]: Instead of aiming to capture the *intensity* of one emotion, we want students to identify the relevant *type* out of several emotions. To this end, we develop an emoticon-based instrument, while taking into account that the youngest respondents cannot read.

Following Baker et al. [1] and Botelho et al. [3], we aim to capture four affective states: Joy, confusion, frustration, and boredom.[2] However, before putting the emoticon-based instrument to use, we first have to ensure that we design a set of emoticons whose meaning is understood by the students. This design and validation process is described in the next section. Section 5 then discusses our initial efforts towards automated affect detection, using this instrument to collect ground-truth labels.

4. AFFECT INSTRUMENT VALIDATION

4.1 Method

We first designed six sets of three emoticons, or 18 in total. The emoticons in the first four sets were designed to capture joy, confusion, frustration, and boredom. We added two more sets to represent surprise and sadness. We then proceeded in three steps.

In step one (the offline adult test), we presented each of the six sets to 14 adults, asking them (1) to express the emotion they inferred from each of the 18 emoticons, and (2) to choose the emoticon (out of three) they found most fitting for each of the six emotions. Achieving consistency across feedback from adults served as a lower threshold in this case. Eventually, this step resulted in six emoticons preferred by adults, one for each affect type.

In step two (the offline child test), we presented these six emoticons to 23 children aged 4-12 in one-on-one sessions, asking them (1) to express the emotion they inferred from each of them, and (2) to choose the emoticon (out of six) they found most fitting for a particular scenario. This step was intended to flag potential interpretation differences between adults and children, allowing for a redesign if necessary. Eventually, this step resulted in a set of emoticons preferred by children.

In step three (the online child test), we used the emoticons from step two to develop a new online 'affect mission' with two levels on the Squla platform. In the first level, children were presented each of the emoticons individually, while being asked to answer the following question: "Look at the character in the picture. What do you see?" They were then offered four answer options, one of which contained the intended meaning of the emoticon.[3]

In the second level of the mission, users were asked questions of the type "Which of the characters is [. . .]?" where [. . .] was substituted for a particular emotion. They were then shown five emoticons, one of which contained the one we intended to actually capture the emotion. For children aged 4-7 we provided voice-recordings of both the question (in both levels), as well as the answer options (in level one).

Table 1: Respondents by grade level

Grade level	# Respondents	Respondent share
KDG	1169	5.4
Pre-K	1972	9.1
Grade 1	2965	13.7
Grade 2	2421	11.1
Grade 3	3222	14.8
Grade 4	3411	15.7
Grade 5	3570	16.4
Grade 6	2985	13.7

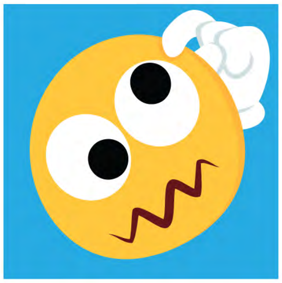

Figure 2: Initial confusion emoticon

The two mission levels were live on the platform for one week (from November 19th 2018 to November 26th 2018). During that time period, we collected responses from 21715 unique players. Table 1 shows the distribution of respondents by grade level. Although respondent shares are notably lower in the bottom two grades, they are representative of the user population across grades.

4.2 Results

The offline adult test resulted in one emoticon per affective state that was preferred by adults. In the offline child test, the emoticons for joy and boredom were correctly interpreted by children. The same holds for frustration, although most children identified this as anger instead. However, given their relatedness, as well as the fact that young children are generally not familiar yet with the notion of frustration, we deem these as interchangeable.

However, confusion was not well identified. Figure 2 depicts the original emoticon that was used. Wen asked what the character was feeling, typical responses were 'nauseous', 'funny', and 'dizzy'. Further probing revealed that both the wriggly mouth as well as the cross eyes confused children's interpretation regarding the intended meaning.

We used this feedback to redesign the confusion emoticon. In particular, two new designs were made. Additionally, we also designed an additional emoticon for frustration, in order to verify that the response 'angry' was not driven by the emoticon but rather by children's lack of awareness of the concept of frustration.

All in all, we ended up with a set of eight emoticons, pre-

[2] To be precise, these studies aim to capture "engaged concentration" rather than "joy". However, in designing the instrument we ran into difficulties capturing the former. Instead, we opted for "joy" as a substitute positive affective state that is more easily understood by children.

[3] An alternative setup would have been to ask children open-answer questions. However, since our youngest users are not able to write, this option is unfeasible. Furthermore, our experience with these types of questions is that many children abuse this freedom of answering by providing nonsensical or offensive answers, increasing the effort of identifying relevant answers. Nonetheless, we attempted to ensure that one or two of the alternative answer options served as detractors.

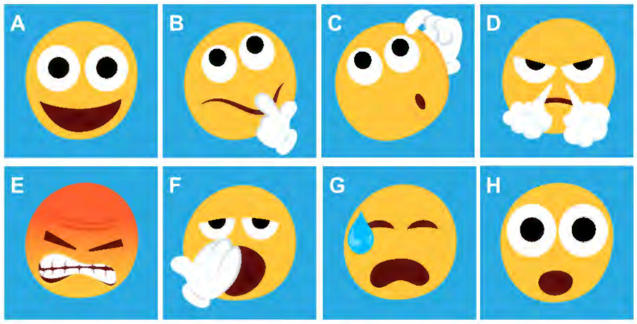

Figure 3: Final test emoticons

Table 2: Answer response shares per trigger

Answer	Level completed	Level interrupted	Navigating
happy	54.4	28.5*	33.2*
bored	15.4	14.8	11.9*
confused	9.6	9.7	6.7*
frustrated	9.5	18*	8.8
not-answered	11.1	29.1*	39.5*

* Response share different from quiz-completed ($p<0.05$)

sented in Figure 3: Joy (A), thoughtfulness (B), confusion (C), anger (D), frustration (E), boredom (F), sadness (G), and surprise (H). Recall that we are primarily interested in capturing joy, confusion, frustration, and boredom. We had two emoticons for confusion (B and C) and frustration (D and E). We used the results of the online child tests to choose the preferred/optimal one. The emoticons for surprise and sadness were used as additional detractors in the online child test.

Using the data from the online child test, we compare the share of respondents that chose a particular answer option. Chi-squared tests show that the distribution of response shares on all questions in all grade levels are significantly different from uniform, indicating that *at least* one answer received a disproportionate share of responses. Computing Chi-squared statistics on all answer pairs per question reveals there is only one instance where *two* answer options are equally preferred: This concerns the level one question regarding the confused emoticon (C) in kindergarten.

The four answer options presented to users in this case are 'struggling', 'confused', 'surprised', and 'sad', with 'struggling' added as a detractor. Indeed, in Kindergarten the detractor received a sufficient number of responses so as to not be statistically differentiable from the 'confused' answer option. Yet for present purposes, the distinction between struggling and confused is negligible. Both these emotions correspond to what we aim to capture, which is that the user is having difficulties understanding or answering questions.

In all of the other cases, the preferred answer option corresponds to the intended answer option. The degree of variation across answer options is notably higher in the lower grades. Figure 4 illustrates this. It shows the response shares across the five emoticon options for the question 'Which character is bored?' Although the bored emoticon (F) receives the highest response shares in all grades, the variation is notably greater in the bottom two grades, with the angry emoticon (D) as a close second.

Finally, the online child test also enables us to identify the preferred emoticon for confusion and frustration. In all grades, emoticon C is preferred for confusion, and emoticon E is preferred for frustration. An added benefit is that the primary detractor for Kindergarten and Pre-K, i.e. emoticon D, is not part of the final set.

5. AFFECT DETECTION

5.1 Method

Using emoticons A, C, E, and F in Figure 3, we designed a modal to prompt users for feedback regarding their affective state during playtime. Since we are targeting young children whose parents pay for the product, we did not want to excessively disrupt their play experience. We therefore present the modal following three specific triggers (with a maximum of three modals per week):

1. When a user finishes a mission-level in one attempt
2. When a user prematurely interrupts a mission-level for the first time
3. When a user has been navigating the platform without playing a mission-level for three consecutive minutes

In all three cases, users are shown a modal that prompts them for their affective state. Their response is logged, together with the relevant trigger, a timestamp, an id of the particular gamesession[4] that they are playing (in case of the first two triggers), or the url that they were visiting when they were shown the modal (in case of the third trigger). Users have the option to close the modal without providing an answer, which is logged as well.

In Section 5.2, we first compare the affect response distribution following the first trigger with that following the other two. The first trigger serves as the benchmark case: We expect that users are predominantly happy when they finish a mission-level, and that the two other triggers typically correlate more strongly with a lack of enjoyment.

In Section 5.3, we then present the results for four affect detectors (one for each affect state) for users playing the subject *math*. We construct 51 mission-level features (such as the share of correct questions, the total number of questions answered, or the minimum, median, maximum and standard deviation of answer time per question) to serve as inputs, with the collected affect responses serving as outputs. We trained the following models (hyper parameters within parentheses):

- Logistic regression with elastic-net regularization (mixing parameter between L1 and L2 regularization, and regularization weight parameter) [**glmnet**]

[4] A gamesession is a particular play instance of a mission-level by a student.

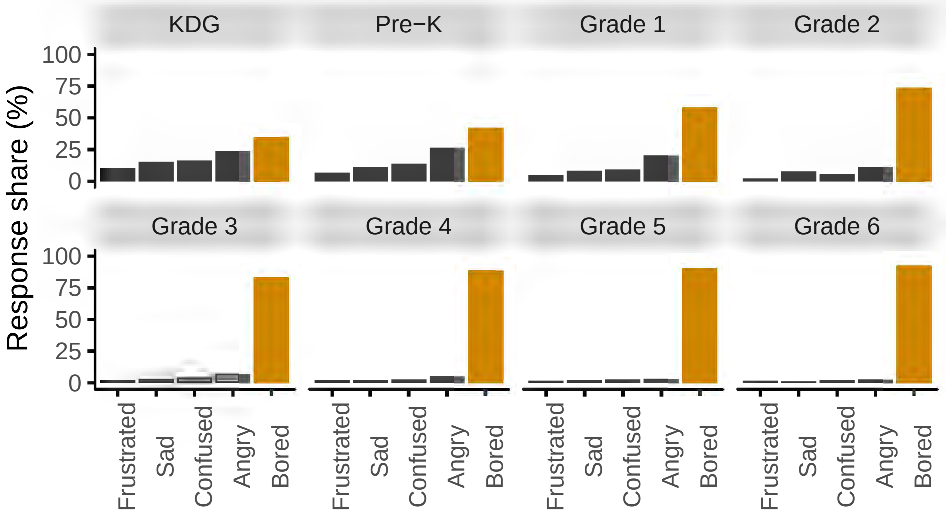

Figure 4: Which character is bored? (intended answer in orange)

- C5.0 classification tree with boosting (tree vs rule-based model, number of boosting iterations, and without or without feature selection) [**C5.0**]
- Naive Bayes classifier (Laplace correction and bandwidth adjustment) [**nb**]
- Random forest (the number of selected features to try during bagging) [**rf**]
- Gradient boosting machine (number of boosting iterations and maximum tree depth) [**gbm**]

Following Baker et al. we use 5-fold student level cross-validation to guide hyper-parameter tuning of the model, and conduct down-sampling to achieve response balance during cross-validation [1]. In terms of hyperparameters, we perform grid searches to determine the optimal parameters during cross-validation, and use the optimal model to compute performance on the (held out) test set. The objective metric for model training is Kappa.

5.2 Trigger comparisons

Table 2 presents the share of responses per answer-type and per trigger. In all three cases, the dominant answer option is 'happy'. Yet the share is notably higher when users complete a level than when they interrupt it or when they are navigating. Conversely, the share of frustrated users is notably higher when a level is interrupted. There is little difference in boredom and confusion between completing and interrupting a level, however. Finally, players are much more likely to ignore the modal when they are navigating or after they interrupt a level, compared to when they complete a level.

It is noteworthy that a non-negligible share of users chooses to ignore the modal, in particular when interrupting a level or after navigating for a while. Given the difference with the 'level completed' share on this response, this *could* be indicative of a lack of enjoyment as well, yet without more detailed user feedback this is impossible to verify.

These patterns are by and large replicated across grades and subjects (for the level completed and interrupted triggers). The most notable differences are that players in kindergarten and pre-school (1) more often choose to ignore the modal (regardless of the trigger), and (2) are less frustrated when interrupting a mission-level.

Table 3: Model training results

Affect state	Model	Kappa	AUC
happy	gbm	0.212	0.643
bored	C5.0	0.051	0.541
confused	rf	0.077	0.617
frustrated	glmnet	0.173	0.669

5.3 Affect detectors

Table 3 presents the model results in terms of Kappa and AUC.[5] The detector for boredom is weakest, followed by that for frustration, confusion, and boredom, respectively. This partly reinforces the results presented in Table 2, where we observed that happiness and frustration exhibit most variation between the level completed and interrupted triggers.

Table 4 shows the top-3 features per detector. The share of correctly answered questions appears in three of the four detectors. This is also true for the response time between the moment the modal is shown, and when it is answered. Finally, note that *happy* and *frustrated* share two of the top-three features. This could indicate that these two emotions are two sides of the same coin.

6. DISCUSSION & CONCLUSION

In this paper we have developed and validated an instrument for affect detection that can be used during home-based educational software use by (young) children. Consistent with previous research, we find that children aged 4-12 are able to

[5] The model performance statistics reported in Table 3 are based on a test set with the original, non-resampled response distributions. The test data are again completely disjoint from the training data at the student

Table 4: Top-3 model features

happy	share of correctly answered questions response time to modal trigger that generates the modal
bored	player grade level share of correctly answered questions a repetitive streak of calculator questions
confused	response time to modal the type of questions that was answered last the average difficulty of questions in the level
frustrated	the difference between user performance and average question difficulty share of correctly answered questions response time to modal

differentiate between four different affective states (joy, boredom, confusion, and frustration) based on a set of emoticons.

We find that users in our study that complete a mission-level are predominantly happy, whereas those that interrupt it are predominantly unhappy. Furthermore, this unhappiness mostly reveals itself as frustration. For users that have been navigating for an extended period of time, the conclusion are less clear, in particular given the high share of users which ignore the modal.

In comparison to existing studies on affect detection in educational software, we find that the response distribution across affect states – reported in Table 2 – is much less skewed. For example, Baker et al. (2012) find that close to 85% of their responses capture engaged concentration, and just 0.9% confusion [1]. The differences with the results reported in this study could be driven by a variety of factors, most notably the fact that out target population includes children instead of adults/adolescents, and the fact that students are playing at home rather than in the classroom. More generally, recall that students are completely free to pick and choose what they play on the platform. This is very different from many of the previous studies, where software is used in the classroom and as part of the curriculum, including instruction. At the very least, it illustrates that the findings in this literature cannot simply be extended to other contexts.

Compared to Baker et. al. [1], the performance of our automated affect detectors is quite a bit weaker. In part, these differences might be driven by the different number of model features (51 vs 258) and in target populations, i.e. children vs adolescents. Yet another explanation might be that our emoticons leave more room for interpretation differences than text surveys. As a result, out affect identification might be less precise than in earlier studies.

Going forward, the ultimate goal is to be able to detect affect during playtime, in order to provide meaningful and effective interventions. The processes and results reported in this paper are a first step towards that goal. The best performing models are still weaker than those found in previous research. However, the models for detecting joy and frustration perform notably better than chance. This suggests that it is worthwhile to put more effort into feature construction and algorithm selection for these detectors.

7. REFERENCES

[1] Baker, R.S. et al. 2012. Towards sensor-free affect detection in cognitive tutor algebra. *Educational data mining 2012, june 19-21, proceedings* (2012).

[2] Baker, R.S.J.d. et al. 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. *Int. J. Hum.-Comput. Stud.* 68, 4 (Apr. 2010), 223–241.

[3] Botelho, A.F. et al. 2017. Improving sensor-free affect detection using deep learning. *Artificial intelligence in education AIED 2017, june 28 - july 1* (2017), 40–51.

[4] Conati, C. and Maclaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction.* (2009), 267–303.

[5] D'Mello, S.K. et al. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction.* 18, 1-2 (Feb. 2008), 45–80.

[6] Fredricks, J.A. et al. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research.* 74, 1, 59–109.

[7] Hershkovitz, A. and Nachmias, R. 2008. Developing a log-based motivation measuring tool. *Educational data mining 2008, june 20-21, proceedings* (2008), 226–233.

[8] Lee, D.M.C. et al. 2011. Exploring the relationship between novice programmer confusion and achievement. *4th international conference on affective computing and intelligent interaction* (Memphis, TN, 2011), 175–184.

[9] Markopoulos, P. et al. 2008. *Evaluating children's interactive products: Principles and practices for interaction designers.* Morgan Kaufmann Publishers Inc.

[10] Read, J. and Macfarlane, S. 2002. Endurability, engagement and expectations: Measuring children's fun. *Interaction design and children, shaker publishing* (2002), 1–23.

[11] Read, J.C. 2008. Validating the fun toolkit: An instrument for measuring children's opinions of technology. *Cogn. Technol. Work.* 10, 2 (Mar. 2008), 119–128.

[12] Sabourin, J. et al. 2011. Modeling learner affect with theoretically grounded dynamic bayesian networks. *Proceedings of the 4th international conference on affective computing and intelligent interaction* (2011), 286–295.

[13] Zea, N.P. et al. 2013. A method to evaluate emotions in educational video games for children. *J. UCS.* 19, 8 (2013), 1066–1085.

[14] Zhu, E. 2006. Interaction and cognitive engagement: An analysis of four asynchronous online discussions. *Instructional Science.* 34, 6 (Nov. 2006), 451.