# A comparison of students' emotional self-reports with automated facial emotion recognition in a reading situation

Franziska S. Hirt†
Institute for Research in Open-, Distance, and eLearning (IFeL)
Swiss Distance University of Applied Sciences (FFHS)
Brig, Switzerland
franziska.hirt@ffhs.ch

Ivan Moser
Institute for Research in Open-, Distance, and eLearning (IFeL)
Swiss Distance University of Applied Sciences (FFHS)
Brig, Switzerland
ivan.moser@ffhs.ch

Egon Werlen
Institute for Research in Open-, Distance, and eLearning (IFeL)
Swiss Distance University of Applied Sciences (FFHS)
Brig, Switzerland
egon.werlen@ffhs.ch

Christof Imhof
Institute for Research in Open-, Distance, and eLearning (IFeL)
Swiss Distance University of Applied Sciences (FFHS)
Brig, Switzerland
christof.imhof@ffhs.ch

Per Bergamin
Institute for Research in Open-, Distance, and eLearning (IFeL)
Swiss Distance University of Applied Sciences (FFHS)
Brig, Switzerland
per.bergamin@ffhs.ch

## ABSTRACT

This study investigated the measurement of students' emotional states during a common learning activity, digital reading of factual texts. The objective was to compare emotional self-reports with automated facial emotion recognition. The latter promises non-intrusive measurements of emotions, which could inform adaptive learning systems. We used an established facial emotion recognition software trained on experts' ratings of facial expressions (FaceReader). For basic emotions, previous studies have reported high agreement of the software with human raters. However, little evidence exists a) on its performance for the epistemic emotions of interest and boredom, b) on its agreement with self-reports, and c) in naturalistic reading situations. We compared the facial expression-based recognition of interest, boredom, and valence of affect to students' self-reports of those emotional states. Analyses of webcam recordings of 103 students revealed no relationship between facial emotion recognition and self-reports. Due to the low agreement of the facial emotion recognition software with self-reports, it remains unclear what the facial expression-based recognition of interest, boredom, and valence actually implies. We advise to wait for more comprehensive evidence a) on the agreement of facial emotion recognition software with self-reports or b) on its predictive validity for learning before applying it in educational practice (e.g., in adaptive learning systems).



*TEEM'18, October 24-26, 2018, Salamanca, Spain*

ACM ISBN 978-1-4503-6518-5/18/10…$15.00
http://dx.doi.org/10.1145/3284179.3284230

## CCS CONCEPTS

• **Applied computing~Psychology**
• **Applied computing~Computer-assisted instruction**
• **Applied computing~Interactive learning environments**

## KEYWORDS

epistemic emotions, facial emotion recognition, FaceReader, emotional self-reports, adaptive learning

**ACM Reference format:**
Franziska Hirt, Ivan Moser, Egon Werlen, Christof Imhof and Per Bergamin. 2018. Continuing Engineering Education and Sustainability: IACEE Contribution with SERINA and Porto Declaration. In *Proceedings of the 6th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2018) (Salamanca, Spain, October 24-26, 2018), F. J. García-Peñalvo Ed. ACM, New York, NY, USA, 8 pages.* https://doi.org/10.1145/3284179.3284230

## 1 INTRODUCTION

### 1.1 Adaptive Learning and Innovations in Learning Analytics

In the context of the establishment of digital and lifelong learning, the methods for learning and teaching are changing. One promising idea is to enable personalized learning experiences in the absence of personal tutors or teachers. This approach is summarized under the term *adaptive learning systems*. Adaptive learning systems modify the learning experience, for example, according to specific characteristics of

the learner [33]. Technology-driven adaptive learning systems have been propagated for decades now [4]. One idea is that adaptive learning might improve educational standards particularly in developing countries with a high percentage of children (combined with few teachers) and in less accessible rural areas. Adaptations on the basis of students' knowledge (measured for example through online tasks) have been implemented in a respectable number of learning systems (for some examples see [2, 32]). However, the promise of adaptive learning systems to deliver a truly personalized experience adapting to different aspects of learning such as emotions has not been met so far [10].

One challenge which must be met to fulfill this promise is to measure students' non-cognitive states – such as emotions – frequently and without hindering the learning process. The most prominent measure of emotions in learning studies, self-report, is not a satisfying sensor for continuously adapting learning systems. First, it would interrupt the learning process. Second, it can be manipulated (e.g. by answering incorrectly on purpose in order to control the adaptive mechanism).

Future adaptive learning systems might benefit from upcoming technological and data science innovations (e.g., wearables, eye-tracking, automated facial emotion recognition, deep leaning algorithms), which promise to estimate students' states such as emotions or focus of attention. In the context of adaptive learning systems or learning analytics, modern sensors are often promised, some are tested in experimental settings, but only rarely are they implemented in broad field settings. Accordingly, it is difficult for practitioners who build adaptive learning systems to understand the current as well as future potential and limitations of those sensors.

Promising innovation is available in the form of facial emotion recognition algorithms (e.g., OpenFace [3], Affectiva [1], FaceReader [22]). FaceReader, for example, estimates the basic emotions (happy, sad, angry, surprised, scared, disgusted) as well as arousal and valence of affect from video captures or pictures of faces [19, 22]. Version 7.1 of the software further offers estimations of the affective attitudes 'interest', 'boredom', and 'confusion' [22]. FaceReader's calculations are based on the Facial Action Coding System (FACS; [9]), which was developed through human observeration. The FACS allows the observer, or in this case, FaceReader, to analyze facial expressions on the basis of Action Units, which represent groups of facial muscles and in certain spatial configurations correspond to specific emotions [9].

## 1.2 Agreement of FaceReader's Estimates with Other Instruments

Facereader's recently available affective attitudes (interest, boredom and confusion) are labeled as "experimental". We are not aware of any information on how accurate those new estimates actually are. Even for FaceReader's more established estimations (e.g., the basic emotions), we are not aware of broad validations in different contexts.

Most of the validation studies available were based on pictures (in which actors were instructed to express specific emotions). Lewinski and colleagues [16] found that FaceReader recognized 89% of the target basic emotions in two high quality picture data bases. They concluded that "FaceReader is as good at recognizing emotions as humans" (p. 4) which in those observations correctly categorized 85% of the pictures. In different contexts, however, FaceReader seems to perform more poorly. Brodny and colleagues [7] found that the accuracy differs between photo and video stimuli. In their study, FaceReader's estimations of basic emotions in good quality video clips (where participants were asked to express specific emotions) matched human ratings in only 56% of the cases. In his master thesis, Suhr [29] analyzed negative emotions in a video data set with the FaceReader and facial electromyography (fEMG). Comparing both measures revealed that they were inconsistent.

Validation studies of facial emotion recognition software are often based on pictures or videos, for which actors were asked to express specific emotions. Videos from naturalistic reading or learning settings probably differ from this rather artificial and highly emotion inducing settings. To our knowledge, there is also little evidence on FaceReader's validity in the context of factual text reading.

We opted to compare this promising, continuous measurement method with the most prominent measure in learning studies, self-report. Thus, we assessed the relationship of FaceReader's estimates with individual self-reports in a reading situation. The question was whether FaceReader's estimations are indeed a valuable substitute for self-reports of emotional states in adaptive learning scenarios. Answering this question is a first step to understand the value of the emotion recognition software for future adaptive learning systems or learning analytics in general.

In this study we did not focus on the basic emotions. Instead, we focused on emotions particularly relevant during complex learning (e.g., reading of factual texts). We refer to those as epistemic emotions [23]. In particular, this study investigated two selected epistemic emotions, interest and boredom, which are supposed to have an impact on learning behavior and performance [14, 31]. As epistemic emotions were available by FaceReader on an experimental basis, we opted to additionally take a more established emotional state into account, namely valence of affect. Valence refers to whether one is in a positive or negative emotional state and does not constitute a discrete emotion (such as basic or epistemic emotions are). In fact, valence and arousal build a dimensional model of emotional states [25]. According to this simple dimensional system, the discrete emotions can be classified as being more/less arousing and more/less positive or negative.

### 1.3 Measuring Different Components of Emotions

A general difficulty in the measurement of emotional states is that there is no uniform understanding of their nature. Different views and approaches are still discussed controversially. Ekman's universal basic emotions propagate a genetic basis for facial expressions of some emotions [8]. This view supports the measurement of emotions via facial expressions. Another point of view is that facial expressions emerge less from the emotions per se, but rather from their underlying appraisals (e.g., considering the social context) [21]. Respectively, the context might affect not only the expression of emotions but also their detection (e.g., through facial action coding). Considering this view suggests some challenges for facial emotion recognition and its interpretation. We expect that the measurement of emotions through facial expression recognition is more difficult in non-social settings where participants might be less expressive compared to situations with active social interactions.

Considering the complexity of the construct, emotions (and in our view also valence) can be regarded as multi-component responses. In Scherer's *Component Process Modell of Emotions*, emotions are composed of five different components [26]:

a) a cognitive component (appraisal),
b) bodily symptoms (e.g., heart rate, electrodermal activity),
c) a motivational component (action tendencies),
d) motor expressions (e.g., face, gesture, inflection), and
e) subjective feelings.

We aimed at comparing FaceReader's emotional estimates with self-reports. Both measurement methods are supposed to measure specific emotions – or as for valence, specific dimensions – , however, they might not measure exactly the same as they refer to different emotional components [11]. Self-reports are typically designed to measure the subjective feeling (e), whereas facial emotion recognition is based on motor expressions (d). Nevertheless, we expected an overlap between both types of measurement as both represent components of a given emotion, respectively, emotional dimension.

In the following sections, we describe the sample, design and analysis of the conducted study. We further present results on the agreement between the automatic facial expression recognition software and self-reports. Finally, we discuss our results and suggest directions for future research.

## 2 METHOD

This study was conducted at a secondary school in the German-speaking part of Switzerland in autumn of 2017. In total, 103 students participated, 87 of which were female (median year of birth = 2000, *SD* = 1.66). The study took place in a temporary laboratory within the school area. The duration of the study was around 30-40 minutes. Students participated during regular school hours (missing one class) and received no financial reimbursement.

### 2.1 Experimental Design and Material

Each participant read two texts (randomly selected from a pool of three text pairs). The texts were each between 200 and 230 words long. One text pair was of low, one of medium, and one of high readability (according to the FLESCH-index [12]). The topics of the texts were all based on topics taught in the psychology classes at the students' school. The texts were not expected to be particularly emotion inducing and were instead representative of common texts that may be found in the syllabus. The order of text one and text two (as well as the between-subject-manipulation of text readability) was assigned randomly.

For each text, the participants first had to read its title, which was separately presented on the computer screen. Participants were then asked about their attitudes and emotions towards the topic (e.g., interest and boredom) and about how they felt at the moment (valence of affect). Afterwards, the participants were presented the corresponding text which they could read without time restrictions. Subsequently, they again rated their attitudes and emotions (e.g., interest and boredom) towards the topic of the text and their current valence of affect.

As we were interested in fluent state measures, we opted against long and time-consuming scales. Interest and boredom were each measured with one item from Pekrun and colleagues [23][1]. Participants were asked about the intensity of emotions they felt towards the topic of the text (e.g., how interested and bored). They rated the intensity on a scale from 1 *(not at all)* to 5 *(very)*.

Valence of affect was measured with one item: a modified version of the SAM (Self-Assessment-Manikin; originally developed by Lang [15]. This modified version was first used by Suk [30]. Participants had to choose between 9 icons to describe their current emotional state (cf. figure 1).

We used the Logitech webcam Pro 9000 (attached at the top of the screen) at 15 to 30 frames per second (variations due to technical issues) to capture the participants' facial expressions. We also recorded participants' eye gaze and heart rate plus electrodermal activity with an eye-tracker and a wearable device respectively. For this paper, however, we solely focused on the facial expression data recorded by the webcam. The experimental script (stimuli presentation and data collection) was operated by the software OpenSesame [20]. Participant-to-screen distance (as well as to camera or remote eye-tracker) was around 60 cm.

For the facial emotion recognition in the videos we used Noldus' FaceReader version 7.1 with its default settings (e.g., using the general face model). Similar to former studies [16] and as recommended by the manual [19] we used no continuous calibration in FaceReader's estimations. The basis for the calculation of valence is the estimates of the basic emotions. FaceReader calculates valence as the intensity of the positive emotion 'happy' minus the intensity of the negative

[1] The item for interest was originally part of a 3-item measure of curiosity, which we extracted here to measure the hardly distinguishable construct, interest.

expression with the highest intensity (either 'sad', 'angry', 'scared' or 'disgusted'). In contrast to the basic emotions, FaceReader's estimates for boredom and interest are not calculated on a frame basis [22] but over a window of 1-5 seconds. A further particularity about the estimation of interest, boredom, and confusion is that they are not only based on action units, but also facial cues such as nodding or head shaking are taken into account. Each emotional state is expressed as a value between 0 and 1, indicating the intensity of the emotion; only valence ranges from -1 to 1.

## 2.2 Analysis

FaceReader provides emotional estimates for each available frame (15-30 per second). Since we were only interested in comparing the self-reports with the overall intensity and not the time course of the emotions, we aggregated FaceReader's estimates over each trial. We did so by using three different approaches which seemed most reasonable to us:

.

**Figure 1: SAM scale as used by Suk** [30]

(1) We calculated mean values over the whole trial. However, aggregating the estimates by calculating their mean might flatten effects and lead to underestimation of differences between trials.

(2) We calculated the mean over 10% of the highest estimates within each trial (for boredom and interest). Such an approach has been successfully used in previous studies [18]. Aggregating the estimates by only using peak values might, however, be strongly influenced by minor movements (e.g., sneezing or coughing [29]) .

(3) As the distributions of the estimates were highly skewed (particularly interest and boredom), we additionally broke them down into dichotomous variables.

This led to the computation of the following predictor variables. The FaceReader estimates for *interest* and *boredom* were aggregated over the duration of the text in three ways:

(1) As mean over the trial (cf. table1).

(2) As mean over the 10% of the highest values (peaks) per trial (cf. table1).

(3) As dichotomous value with the median of the 10% of the highest values being the threshold for dividing the averaged 10% peak values per trial. Respectively, half of the trial was labeled with "high intensity" and half with "low intensity" of the respective emotion or valence.

The FaceReader estimates for *valence* were aggregated over the duration of the trial in two ways:

(1) As mean over the trial (cf. table1).

(3) As dichotomous value with 0 being the threshold dividing the mean valence of trial in positive and negative. Resulting overall in 182 trials with negative mean valence and 22 trials with positive mean valence.

All those aggregated estimates were introduced as level 1 predictors in separate linear mixed-effects regression models predicting the three emotional self-reports. Additionally, we controlled for the specific text (6 levels) as another level 1 predictor. Random intercepts for the participants were included to account for between-subject variability in the self-reports. Calculations were done using the lme4 package [5] within the statistical software R [24]. We compared the model with intercept and text as predictor with the model additionally including the predictor of interest, the FaceReader estimate. For statistical comparison the likelihood ratio test, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) were calculated.

# 3 RESULTS

On average, participants read the short texts for 83.6 seconds (*SD* = 26.0). This reading duration defined the number of estimates available from the FaceReader. In 3% of the video frames no estimation was possible for the FaceReader as it could not find or model a face.

Table 1 presents median and standard deviation for interest, boredom, and valence measured via self-reports (before and after reading the text); furthermore, the mean per trial and the mean of the 10% of the highest values per trial estimated with FaceReader are presented**.** The distributions of the self-reports were all skewed (valence and interest left skewed, boredom right skewed). All FaceReader estimates were skewed to the right (many near zero values for interest and boredom, for valence rather negative than positive values).

**Table 1: Descriptives of FaceReader's estimates (aggregated as mean and mean of peak values) and students' self-reports**

| | Self-report before | Self-report after | FaceReader (0-1) mean | FaceReader mean over the 10% highest values |
|---|---|---|---|---|
| Interest | 3.91 (0.87; *1-5*) | 3.80 (0.93; *1-5*) | 0.01 (0.04; *1-5*) | 0.03 (0.08; *0-1*) |
| Boredom | 1.57 (0.89; *1-5*) | 1.44 (0.79; *1-5*) | 0.06 (0.15; *1-5*) | 0.32 (0.28; *0-1*) |

| | | | | |
|---|---|---|---|---|
| Valence | 6.82 (1.19; *1-9*) | 6.94 (1.13; *1-9*) | -0.12 (0.15; *1-9*) | 0.03 (0.16; *-1 to 1*) |

*Note:* The table presents the mean (SD, *Scale range – higher values indicating higher intensity*).

As model comparisons revealed no significant effect for any of the calculated FaceReader estimates, we only report likelihood ratio test and fit indices of each model (cf. table 2).

Figure 2 presents the distribution of the self-reports grouped by the dichotomous FaceReader estimates. In line with the regression models, visual inspection of those plots also indicated no difference in the median of participants' self-reports when FaceReader estimated a high versus low intensity of those emotions and valence. Accordingly, we found no indication of a relationship between FaceReader's aggregated estimates and the self-reports.

**Table 2: Overview of the model fit indices and their comparisons**

| | Model with aggregation method for FaceReader's estimates | Comparing M0 with M1 | Model fit indices of M0 and M1 |
|---|---|---|---|
| Interest | mean as predictor | $\chi^2(1) = 2.354$, $p = 0.125$, $N = 103$ | $AIC_{M0} = 533.92$; $BIC_{M0} = 560.50$ |
| | | | $AIC_{M1} = 533.56$; $BIC_{M1} = 563.47$ |
| | mean over the 10% of the highest values as predictor | $\chi^2(1) = 0.2852$, $p = 0.2852$, $N = 103$ | $AIC_{M0} = 533.92$; $BIC_{M0} = 560.50$ |
| | | | $AIC_{M1} = 534.77$; $BIC_{M1} = 564.68$ |
| | dichotomous predictor (split by median) | $\chi^2(1) = 0.0683$, $p = 0.7938$, $N = 103$ | $AIC_{M0} = 533.92$; $BIC_{M0} = 560.50$ |
| | | | $AIC_{M1} = 535.85$; $BIC_{M1} = 565.76$ |
| Boredom | mean as predictor | $\chi^2(1) = 0.4991$, $p = 0.4799$, $N = 102$ | $AIC_{M0} = 471.34$; $BIC_{M0} = 497.89$ |
| | | | $AIC_{M1} = 472.84$; $BIC_{M1} = 502.71$ |
| | mean over the 10% of the highest values as predictor | $\chi^2(1) = 0.9771$, $p = 0.3229$, $N = 102$ | $AIC_{M0} = 471.34$; $BIC_{M0} = 497.89$ |
| | | | $AIC_{M1} = 472.36$; $BIC_{M1} = 502.23$ |
| | dichotomous predictor (split by median) | $\chi^2(1) = 1.0149$; $p = 0.3137$, $N = 102$ | $AIC_{M0} = 471.34$; $BIC_{M0} = 497.89$ |
| | | | $AIC_{M1} = 472.33$; $BIC_{M1} = 502.19$ |
| Valence | mean as predictor | $\chi^2(1) = 0.2366$, $p = 0.6267$, $N = 97$ | $AIC_{M0} = 533.39$; $BIC_{M0} = 559.49$ |
| | | | $AIC_{M1} = 535.15$; $BIC_{M1} = 564.52$ |
| | dichotomous (positive/negative) as predictor | $\chi^2(1) = 0.0681$, $p = 0.7942$, $N = 97$ | $AIC_{M0} = 533.39$; $BIC_{M0} = 559.49$ |
| | | | $AIC_{M1} = 535.32$; $BIC_{M1} = 564.68$ |

*Note: M0 = restricted model (without FaceReader's estimate as predictor); M1 = full model including FaceReader's estimate as predictor.*

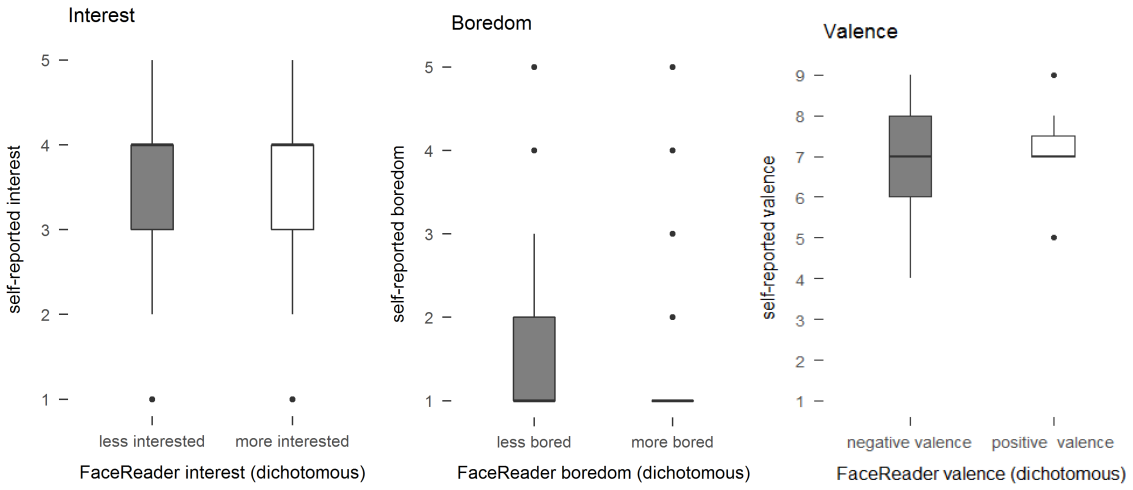


**Figure 2: Boxplots presenting the values of the self-reports grouped by the dichotomous FaceReader estimates**

*Note.* Scales for self-reported interest and boredom range from 1 *(not at all)* to 5 *(very)*, and for valence from 1-9 with 5 being neutral, 1-4 being negative and 6-9 being positive; the dichotomous FaceReader estimates for interest and boredom were built by median split of the mean of the 10% highest values per trial, valence was built by splitting the mean values by the threshold 0. The bold horizontal line in the boxplots represents the median, with boxes indicating the range between the first and third quartile. Whiskers extend from the hinges to data points no further than, respectively at most, 1.5 times the inter-quartile range from the hinge. "Outlying" points, are individually plotted data falling beyond the end of the whiskers.

## 4 DISCUSSION

The objective of this study was to compare students' self-reports with automated facial recognition of interest, boredom, and valence in a reading situation. We investigated whether aggregated scores of a facial emotion recognition software predicted emotional states reported by 103 students after having read a short factual text. Irrespective of different types of aggregation of the facial emotion estimates, neither the recently available ("experimental") epistemic emotions (i.e., boredom and interest), nor the established estimates of valence predicted the respective self-report measure. We therefore argue that FaceReader's automated facial emotion recognition does not correspond to single-item self-reports of interest, boredom, and valence.

### 4.1 Construct and Predictive Validity of Components of Emotional States

The actual meaning of this result remains unclear as self-reports of emotional states cannot necessarily be considered the "true" benchmark. Inspecting the means of interest, boredom, and valence indicated clear differences in intensity compared to FaceReader's estimates. Self-reports all differed from normal distributions in the direction of social desirability (higher interest, lower boredom, more positive valence compared to FaceReader's estimates). Such skewed distributions might have been affected by some form of instrument or response bias. Particularly the distribution for boredom, where over half of the participants indicated not to be bored (marking 1 on a scale from 1-5), is noteworthy and a reason to consider the results with caution. The skewedness in direction of social desirability raises the question if self-reports are a valid or useful instrument for measuring emotions at all [34]. This controversial question and the lack of an evident "gold standard" in measuring emotions clearly represents a huge challenge for the understanding of emotions.

Previous studies comparing facial expressions with self-reports yielded mixed results. Soleymani [27] for example found no agreement of facial expressions (recorded while watching different images and micro-videos) with participants' self-reports of interest and curiosity. Soleymani and Mortillaro [28] found rank correlations between $\rho = 0.23$ to 0.33 ($SD = 0.06$ to 0.11) between facial expressions (recorded while watching different images and micro-videos) and participants' self-reports of curiosity. In the context of bereavement interviews Bonanno and Keltner [6] found correlations between facial expressions and self-reports of anger, but non-significant results for sadness and joy. Furthermore, Harley and colleagues [13] compared FaceReader with self-reports of emotions, finding an agreement of 75.6%, however by using a different method for comparison[2]. Although it remains unclear what different measurement methods (i.e., components) of "emotions" actually measure and how they relate to each other (lack of established *construct validity*), it might be even more important for learning analytics to understand the *predictive validity* of the emotional components for learning behavior and performance. It is crucial to understand which combination of measurement types of emotions predict learning performance or behavior most accurately – irrespective of what those measurements actually mean. Which combination of, for example, physiological data, facial expression recognition, or self-report predict learning performance best? The answer to this question might vary for different emotional states. Interaction effects between different measurement types are conceivable as well. Further research should investigate these questions. It is important for practitioners to understand which combination of sensors (e.g., webcam, eye-tracker, wearables, and logfiles) and algorithms might deliver estimates the most predictive of learning performance or behavior. Such evidence could help to understand which affective sensors learning instructions should be adapted in dependence of in order to improve learning.

### 4.2. Specific Issues in the Computation of Facial Emotion Recognition Estimates

We argue that FaceReader's estimates do not deliver valid indicators for students' self-report of interest, boredom and valence towards factual texts in non-social settings (without applying previous calibration or providing a baseline). Testing facial emotion recognition in a reading situation is rather conservative as we expect students to be much less expressive while reading alone compared to social learning contexts. Results might be much different in social and more emotion inducing settings. Furthermore, we might find higher agreement between self-reports and facial emotion recognition for basic emotions; then again, they formed the basis for the estimation of valence which did not reveal different results than interest or boredom. Further research is needed to clarify if facial emotion recognition yields better accuracy and agreements with self-reports when more video data is provided per participant (e.g., a baseline).

A possible explanation for our results is that the algorithm behind the FaceReader might have been trained under circumstances very distinct from our setting. First, FaceReader was trained on image data sets for which human observers' ratings ("annotations") – and not actors' self-reports – were applied as ground truth [22]. Accordingly, comparing FaceReader's estimates with assessments by others (e.g., teachers) and not one-self might lead to higher agreement rates. However, other authors did use emotional self-reports as

[2] They used FaceReader's most dominant emotional state during the 10 seconds before the self-report measure and counted it as agreement, when participants rated a similar emotion with at least 3 (on a scale of 1-5).

ground truth with little [27] to moderately [28] encouraging results. Second, an algorithm specifically trained on videos of people reading factual texts might yield different results; however, we are not aware of such an algorithm. Accordingly, this study prompts the need to further investigate the usefulness of algorithms for facial emotion recognition in different settings. For practice it is relevant to understand how they perform in specific contexts (social/isolated, emotional content of the stimuli) and different formats of recording (picture/video, quality). A further idea is to train algorithms for the facial recognition of epistemic emotions particularly with the focus on learning settings.

Another need for research is to provide information on how to analyze facial expression data and compare them to other measurement types. A limitation of this study is the use of different time points for the measurement by FaceReader (simultaneously during reading) and self-reports (afterwards). Measuring emotions during reading via self-reports in a simultaneous manner seems very hard to realize in a reasonable way. This complicates the comparison of FaceReader's estimates with self-reports. There is an almost infinite number of possibilities to compare those two measurement types (e.g., mean values, peak values, and progression tendency over time). We aggregated the FaceReader's estimates in three different ways, but different approaches with potentially different results are conceivable.

### 4.3 Conclusions for Learning Analytics and Adaptive Learning

This study did not yield high agreement rates of automated facial emotion recognition with other forms of measurement such as previous studies did [16, 17]. We assume that changes in the experimental setting (reading situation) and comparing the estimates with a different measurement type of emotions (subjective self-reports) have led to drastic reduction in agreement rates, which we consider as remarkable and of interest. Given the poor agreement between an established emotion recognition software and students' self-report of interest, boredom, and valence in this study, we advise practitioners to wait for further research on those estimates before applying them in learning analytics or adaptive learning systems.

For adaptive learning systems, the accuracy and predictive power of the sensors used as basis for adaptation is crucial for the performance of the whole adaptive system. Thus, the importance to develop valid and accurate sensors on which to base adaptations cannot be stressed enough. We consider the lack of information on non-intrusive, accurate sensors predictive for learning performance as a decisive obstacle for the use of such sensors in adaptive learning systems. Practitioners need tools at hand, which measure variables relevant for learning without impeding the learning process. Otherwise, how can adaptive learning systems, which effectively adapt to a variety of students' needs, be implemented in practice?

The variety of innovative measures of emotions is increasing and so are low cost versions of such sensors. We consider it worthwhile to evaluate the usefulness of such sensors not only for the gaming, marketing, or sports industries, but also for learning settings. Considering the lack of conclusive research on the nature of emotions and the relationship between their components, we propose to systematically compare different measurement types with regard to their predictive validity for different learning parameters in applied settings.

## REFERENCES


[1] Affectiva Homepage: *https://www.affectiva.com/*. Accessed: 2018-06-01.

[2] Aleven, V. et al. 2016. Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of Research on Learning and Instruction*. Routledge.

[3] Amos, B. et al. 2016. Openface: A general-purpose face recognition library with mobile applications. DOI:https://doi.org/10.5281/zenodo.32148.

[4] Atkinson, R.C. 1976. *Adaptive Instructional Systems: Some attempts to optimize the learning process*. Stanford University, Institute for Mathematical Studies in the Social Sciences.

[5] Bates, D. et al. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. DOI:https://doi.org/10.18637/jss.v067.i01.

[6] Bonanno, G.A. and Keltner, D. 2004. The coherence of emotion systems: Comparing "on-line" measures of appraisal and facial expressions, and self-report. *Cognition and Emotion*, 18(3), 431–444. DOI:https://doi.org/10.1080/02699930341000149.

[7] Brodny, G. et al. 2016. Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions. *Proceedings - 2016 9th International Conference on Human System Interactions, HSI 2016*, 397–404.

[8] Ekman, P. and Cordaro, D. 2011. What is meant by calling emotions basic. *Emotion Review*, 3(4), 364–370.

[9] Ekman, P. and Friesen, W. V. 1976. Measuring facial movement. *Environmental Psychology and Nonverbal Behavior*, 1, 56–75.

[10] Essa, A. 2016. A possible future for next generation adaptive learning systems. *Smart Learning Environments*, 3. DOI:https://doi.org/10.1186/s40561-016-0038-y.

[11] Feldman Barrett, L. et al. 2005. Interoceptive sensitivity and self-reports of emotional experience. *Journal of Personality and Social Psychology*, 87(5), 684–697. DOI:https://doi.org/10.1016/j.molcel.2009.10.020.The.

[12] Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

[13] Harley, J.M. et al. 2013. Aligning and comparing data on emotions experienced during learning with metatutor. In H.C. Lane et al. (eds), *Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science, vol 7926*.Springer. 61–70.

[14] Krapp, A. et al. 1991. Interest, learning, and development. *The role of interest in learning and development*. Erlbaum. 3–25.

[15] Lang, P.J. 1980. Behavioral treatment and bio-behavioral assessment: computer applications. In J.B. Sidowski et al. (eds), *Technology in mental health care delivery systems*. Ablex. 119–137.

[16] Lewinski, P. et al. 2014. Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics*, 8(1), 58–59. DOI:https://doi.org/10.1037/npe0000028.

[17] Lewinski, P. 2015. Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets. *Frontiers in Psychology*, 6(1386). DOI:https://doi.org/10.3389/fpsyg.2015.01386.

[18] Lewinski, P. 2015. Don't look blank, happy, or sad: Patterns of facial expressions of speakers in banks' YouTube Videos predict video's popularity over time. *Journal of Neuroscience, Psychology, and Economics*, 8(4), 1–9. DOI:https://doi.org/10.13140/RG.2.1.4653.6409.

[19] Loijens, L. and Krips, O. 2018. *FaceReader Methodology Note. A white paper by Noldus Information Technology*.

[20] Mathôt, S. et al. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, 44(2), 314–324. DOI:https://doi.org/10.3758/s13428-011-0168-7.

[21] Moors, A. 2013. Appraisal theories of emotion: State of the art and future development, *Emotion Review*. 5(2), 119–124.

[22] Noldus Homepage: *https://www.noldus.com/facereader/whats-new-facereader-71*. Accessed: 2018-05-18.

[23] Pekrun, R. et al. 2016. Measuring emotions during epistemic activities: The Epistemically-Related Emotion Scales. *Cognition & Emotion*, 31( 6), 1–9. DOI:https://doi.org/10.1080/02699931.2016.1204989.

[24] R Core Team 2018. R: A language and environment for statistical

computing. 2018.

[25] Russell, J.A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.

[26] Scherer, K.R. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4), 695–729. DOI:https://doi.org/10.1177/0539018405058216.

[27] Soleymani, M. 2016. Detecting cognitive appraisals from facial expressions for interest recognition. *preprint arXiv*. DOI:https://doi.org/arXiv:1609.09761v2.

[28] Soleymani, M. and Mortillaro, M. 2018. Behavioral and Physiological Responses to Visual Interest and Appraisals: Multimodal Analysis and Automatic Recognition. *Frontiers in ICT*, 5, 17. DOI:https://doi.org/10.3389/fict.2018.00017.

[29] Suhr, Y.T. 2017. *FaceReader, a promising instrument for measuring facial emotion expression? A comparison to facial electromyography and self-reports*. Master thesis, Utrecht University.

[30] Suk, H.-J. 2006. *Color and emotion - a study on the affective judgment across media and in relation to visual stimuli*. Doctoral dissertation, University of Mannheim.

[31] Tze, V.M.C. et al. 2015. Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educational Psychology Review*, 28(1), 119–144.

[32] Vandewaetere, M. et al. 2011. The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behavior*, 27, 118–130. DOI:https://doi.org/10.1016/j.chb.2010.07.038.

[33] Wauters, K. et al. 2010. Adaptive item-based learning environments based on the item response theory: Possibilities and challenges. *Journal of Computer Assisted Learning*, 26(6), 549–562. DOI:https://doi.org/10.1111/j.1365-2729.2010.00368.x.

[34] Zimmermann, P. et al. 2003. Affective computing - A rationale for measuring mood with mouse and keyboard. *International Journal of Occupational Safety and Ergonomic*, 9(4), 539–551. DOI:https://doi.org/10.1080/10803548.2003.11076589.