Research Article

Franziska Hirt*, Egon Werlen, Ivan Moser, and Per Bergamin

Measuring emotions during learning: lack of coherence between automated facial emotion recognition and emotional experience**

https://doi.org/10.1515/comp-2019-0020
Received May 29, 2019; accepted Sep 16, 2019

Abstract: Measuring emotions non-intrusively via affective computing provides a promising source of information for adaptive learning and intelligent tutoring systems. Using non-intrusive, simultaneous measures of emotions, such systems could steadily adapt to students emotional states. One drawback, however, is the lack of evidence on how such modern measures of emotions relate to traditional self-reports. The aim of this study was to compare a prominent area of affective computing, facial emotion recognition, to students' self-reports of interest, boredom, and valence. We analyzed different types of aggregation of the simultaneous facial emotion recognition estimates and compared them to self-reports after reading a text. Analyses of 103 students revealed no relationship between the aggregated facial emotion recognition estimates of the software FaceReader and self-reports. Irrespective of different types of aggregation of the facial emotion recognition estimates, neither the epistemic emotions (*i.e.*, boredom and interest), nor the estimates of valence predicted the respective self-report measure. We conclude that assumptions on the subjective experience of emotions cannot necessarily be transferred to other emotional components, such as estimated by affective computing. We advise to wait for more comprehensive evidence on the *predictive validity* of facial emotion recognition for learning before relying on it in educational practice.

Keywords: affective computing, facial emotion recognition, FaceReader, emotional self-reports, epistemic emotions

1 Introduction

1.1 Potential of affective computing for adaptive learning and intelligent tutoring systems

In the context 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 terms adaptive learning systems and intelligent tutoring systems. Such systems modify the learning experience by addressing specific characteristics or behaviours of the learner. However, the promise to deliver a truly personalized learning experience has not been met so far. For such personalization, the system would not only adapt to students' current state of knowledge, but also to their emotions and beliefs, etc. One challenge, which must be met to fulfill this promise of personalization, is to measure students' states such as emotions frequently and without hindering the learning process. The most prominent measures of emotions in learning studies, self-reports, are not a satisfying sensor for measuring emotions during learning. First, self-reports interrupt the learning process and thereby might hamper learning. Second, self-reports can be manipulated (*e.g.*, by answering incorrectly on purpose in order to undermine the system).

Future adaptive learning and intelligent tutoring systems might benefit from affective computing [1, 2]. One aspect of affective computing is affect detection. Affect detection builds upon sensors (such as webcams, wearables, eye-

***Corresponding Author: Franziska Hirt:** Swiss Distance University of Applied Sciences (FFHS), Brig, Switzerland; Email: franziska.hirt@ffhs.ch
Egon Werlen: Swiss Distance University of Applied Sciences (FFHS), Brig, Switzerland; Email: egon.werlen@ffhs.ch
Ivan Moser: Swiss Distance University of Applied Sciences (FFHS), Brig, Switzerland; Email: ivan.moser@ffhs.ch
Per Bergamin: Swiss Distance University of Applied Sciences (FFHS), Brig, Switzerland; Email: per.bergamin@ffhs.ch

** A previous version of this article has been published as Hirt F. S., Moser I., Werlen E., Imhof C., Bergamin P., A comparison of students' emotional self-reports with automated facial emotion recognition in a reading situation, in Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality, ACM, 2018, 320-327 trackers, log files, etc.), which attempt to infer students' emotions by analyzing these big data. In the context of adaptive learning and intelligent tutoring systems, the potential of affective computing is often promised, some prototypes are tested in experimental settings (*e.g.*, [3]), but only rarely are they implemented in broad field settings. Accordingly, it is difficult for practitioners to understand the current as well as future potential and limitations of affective computing for education.

1.2 What are emotions?

A general difficulty in the detection of emotional states is that there is no uniform understanding of their nature. In fact, the term emotion is "notoriously fuzzy, ill-defined, and possibly indeterminate" [4]. In this paper, we, therefore, use the terms affect and emotions interchangeably. One view of emotions is that they only comprise subjectively aware aspects: feelings. Another view on emotions, which is possibly more fruitful for affective computing, is to regard them as multi-component responses. In Scherer's *Component Process Model of Emotions*, for example, emotions comprise five different components [5]:

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.

In this model, subjective feelings are just one component among many. Hereby, the model offers many sources of data for affective computing. The relationship between these components, however, remains unspecified: "The link need not be particularly strong, nor does there need to be strong coherence or synchrony among the various bodily/physiological responses" [6]. In our view, the lack of a well-defined expectation on the relationship between these components delivers a huge challenge for the validation of affective computing, but also a chance for more extensive multimodal measures of emotions.

Two forms of motor expressions, voice, and facial expression are the most common modalities used as indicators for affective computing [7]. The present study focuses on facial expressions of emotions. The first reason for that choice is that empirical findings suggest that facial expression is the most promising modality as a single indicator for affective computing [8]; second, facial expressions can be captured relatively easily via webcams – even in naturalistic reading situations. In contrast to affective computing, classical psychological research traditionally relies on self-reports of emotions (*i.e.*, the subjective experience). As most didactic implications of emotions are based on research on the subjective experience of emotions (*e.g.* [9–11]), it is of interest to investigate whether facial expression based emotion recognition does measure something similar. Without evidence on the coherence between facial expressions and the subjective experience of emotions, the empirical results and implications stemming from studies on the subjective experience of emotions should not be directly transferred to facial expressions of emotions.

One view that supports the measurement of emotions via facial expressions is the concept of Ekman's universal basic emotions advocating a genetic basis for (facial) expressions of some emotions [12]. Another point of view is that facial expressions emerge less from the emotions per se, but rather from their underlying appraisals which are also influenced by the social context in which the emotions are expressed [13]. Respectively, the context might affect not only the expression of emotions but also their detection through automated facial emotion recognition. 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 [14].

1.3 The state of the art of facial emotion recognition

The field of affective computing has developed facial emotion recognition algorithms and software such as OpenFace [15], Affectiva [16], and FaceReader [17]. Often, facial emotion recognition is based on the Facial Action Coding System (FACS; [18]), which was developed through human observation. The FACS allows the observer or an algorithm 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 [18]. 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 [17, 19]. The classification of FaceReaders' basic emotions was trained on human observers' ratings of faces. In version 7.1, the software offers estimations of the affective attitudes 'interest', 'boredom', and 'confusion' which are partly based on Action Units of the FACS [17].

However, it remains yet unclear what facial emotion recognition actually corresponds to. Studies comparing facial expressions with self-reports, for example, yielded mixed results. Soleymani and Mortillaro [8, 20] trained random forest regression models by using student's self-reported interest and curiosity as ground truth. In a study in 2016, Soleymani [20] found no agreement of facial expressions (recorded while watching different images and micro-videos) with participants' self-reports. In 2018, Soleymani and Mortillaro [8] report rank correlations ranging from ρ = 0.23 to 0.33 (*SD* = 0.06 to 0.11) between facial expressions and participants' self-reports of curiosity. In the context of bereavement interviews Bonanno and Keltner [21] found correlations between facial expressions (coded by observers according to the FACS) and self-reports of anger, but non-significant results for sadness and joy.

Facereader's recently available affective attitudes (interest, boredom, and confusion) were labeled as "experimental" in version 7.1. We are not aware of any published information on how valid those new estimates are. Even for FaceReader's more established estimations (*e.g.*, the basic emotions), we are not aware of broad validations in different contexts. Preliminary evidence, however, partly supports FaceReader's validity for the basic emotions. For example, Lewinski and colleagues [22] found that FaceReader recognized 89% of the target basic emotions in two high-quality picture databases. They concluded that "FaceReader is as good at recognizing emotions as humans", which in those observations correctly categorized 85% of the pictures. Furthermore, Harley and colleagues [23] compared FaceReader with self-reports of emotions, finding an agreement of 75.6%, although by using a particular method of comparison[1]. In different contexts, however, FaceReader seems to perform more poorly. Brodny and colleagues [24] found that the accuracy differs between photo and video stimuli. In their study, FaceReader's estimations of basic emotions in good quality video clips, for which participants were asked to express specific emotions, matched human ratings in only 56% of the cases. In his master thesis, Suhr [25] analyzed negative emotions in a video data set with the FaceReader and facial electromyography (fEMG). Comparing both measures revealed that they were inconsistent.

Concluding from these studies, facial emotion recognition performs differently depending on how and in which context it is administered. Validation studies of automated emotion detection – including facial emotion recognition software – are often based on pictures or videos, for which actors were asked to express specific emotions [7]. Videos from naturalistic reading or learning settings differ from this rather artificial and highly emotion-inducing settings [26]. To our knowledge, there is little evidence on FaceReader's validity in the context of factual text reading, which is the context of interest for our research.

1.4 The goals of this study

As we were interested in affective computing in learning settings, we did not focus on the basic emotions, but on emotions particularly relevant during complex learning. We refer to those learning-related emotions as epistemic emotions [27]. In particular, this study investigated two selected epistemic emotions, interest and boredom, which are supposed to have an impact on learning behavior and performance [11, 28]. As epistemic emotions were available by FaceReader only recently, 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 [29]. According to this simple dimensional system, discrete emotions can be classified regarding their level of arousal and positivity/negativity.

We aimed at comparing two measures of emotions: FaceReader's emotional estimates and traditional self-reports. Both are measures of specific emotions, but measure different components of those emotions (*i.e.*, motor expressions versus subjective feelings) and therefore might differ. Nevertheless, we expected some form of coherence between both types of measurement, as both represent components of the same emotion or emotional dimension.

A challenge in this study was that we aimed at comparing self-reports measured after a text with continuous measures during the reading of the text. Hence many different ways to compare those point measures with the continuous measures are conceivable. As the development of facial expressions and subjective experience might differ over the time course [4], it is complicated to compare them. Just aggregating the continuous measures over the whole event in the form of calculating the mean is therefore not the only strategy. We applied different approaches of comparison which will be explained in more detail in the method section.

To conclude, our research question was whether continu-

[1] 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).

ous facial emotion recognition estimates would in some manner match self-reports of discrete emotional states. Specifically, we compared facial expression based recognition of interest, boredom, and valence of affect on students' self-reports of those emotions in the context of the digital reading of factual texts. As most of the evidence on emotions' influence on learning is based on emotional self-reports, understanding the coherence between facial expressions and subjective experience is a first step towards determining the value of facial emotion recognition.

Although there are some studies on the empirical evaluation of affective computing based emotion detection, their conclusions differ. Some argue that their systems are ready for real-world applications in naturalistic educational settings [2], whereas others conclude that further research is necessary in order to do so (*e.g.*, [7]).

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 emotion 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 the 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 incentive.

2.1 Experimental design and material

Each participant read two texts which were selected from a pool of three text pairs (*i.e.*, a pool of six texts). The texts were between 200 and 230 words long. One text pair was of relatively low, one of the medium, and one of high readability (according to the German version of the FLESCH-index [30, 31]). 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). Afterward, 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[2] and boredom was each measured with one item from Pekrun and colleagues [27]. Participants were asked about the intensity of emotions they felt towards the topic of the text (*e.g.*, how interested and bored they were). They rated the intensity on a scale from 1 *(not at all)* to 5 *(very)*. The valence of affect was measured with one item: a modified version of the SAM (Self-Assessment-Manikin; in this form first used by Suk [32]). Participants had to choose between nine icons (ranging from sad to smiling faces) to describe their current emotional state.

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. Participant-to-screen distance (as well as to camera) was around 60 cm. 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 and students' self-reports. The experimental script (stimuli presentation and data collection) was operated by the software OpenSesame [33].

For the facial emotion recognition in the videos, we used Noldus' FaceReader version 7.1. According to Noldus, "FaceReader is used worldwide at more than 700 universities, research institutes, and companies in many markets, such as consumer behavior research, usability studies, psychology, educational research, and market research" [17]. Accordingly, FaceReader seemed appropriate for our purposes. Similar to former studies (*e.g.*, [22]) and as recommended by the manual [19], we used FaceReader with its default settings (*e.g.*, using the general face model) and without continuous calibration. The basis for the computation of valence is the estimates of the basic emotions. The estimates for the basic emotions rely on an artificial neural network trained on image data sets with human observers'

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

ratings ('annotations') [19]. FaceReader computes valence as the intensity of the positive emotion 'happy' minus the intensity of the negative expression with the highest intensity (either 'sad', 'angry', 'scared' or 'disgusted'). A particularity about the estimation of interest, boredom, and confusion is that they are based on action units and additional facial cues such as nodding. The selection of the action units and facial cues is based on studies on computer-based tutoring systems [34–36]. In contrast to the basic emotions, FaceReader's estimates for boredom and interest are not calculated isolated on a frame basis [17] but over a window of 5 seconds for boredom and 2 seconds for interest. In the beginning, there is only data after 2.5 seconds for boredom and 1 second for interest, hence the output at the beginning is based on fewer data. 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 video frame (15-30 per second). We aggregated FaceReader's estimates over each event (one text reading situation) in order to facilitate the comparison with the self-report after the event. We did so by using three different approaches which seemed most reasonable to us:

(1) *Mean:* We calculated the mean value over each event. However, aggregating the estimates might flatten effects and lead to underestimation of differences between events.

(2) *Standard deviation:* We calculated the standard deviation of the mean value over each event as some additional information on the pattern over time (variance during the event).

(3a) *Peaks for boredom and interest:* We calculated the mean over 10% of the highest estimates within each event. Such an approach has been successfully used in a previous study [37]. We assumed that such peak values might be particularly relevant for the subjective (retrospective) experience of emotions. Aggregating the estimates by only using peak values might, however, be strongly influenced by minor movements (*e.g.*, sneezing or coughing [25]).

(3b) *Peaks for positive and negative valence:* We computed the mean of the 10% of the most extreme positive and the mean of most extreme negative values (peaks) per event. Valence has a range from −1 (highly negative valence) to 1 (highly positive valence). Zero hereby represents a neutral state of emotional valence. Due to this two-dimensional nature of valence, we could not directly apply the method described in 3a. Instead, we decided to divide valence into two separate variables: positive and negative valence. We based the computation of the two variables on FaceReader's formula for valence. Accordingly, we used the values of 'happy' as positive valence and the negative expression with the highest intensity per frame (either 'sad', 'angry', 'scared' or 'disgusted') as negative valence. Then we proceeded as in 3a and calculated the mean of the 10% of the most extreme values over each event for both positive and negative valence.

Some descriptive data of the self-reports and the aggregated facial emotion recognition estimates are presented in Table 1.

All aggregated FaceReader-estimates were introduced as level 1 predictors in separate Bayesian generalized linear mixed models predicting the three emotional self-reports. Only the standard deviation of the mean was not included alone as level 1 predictor but in combination with FaceReader's mean. We also included the interaction of the standard deviation and the mean, as we were interested in whether their effects were dependent on each other. For example, FaceReader's mean over the event might predict the emotional experience better when there is more variability (*i.e.*, more expressiveness) in FaceReader's estimates. Random intercepts for the participants as well as for the six different texts were included to account for between-subject and between-text variability in the emotional self-reports. We standardized all predictors (mean = 0, *SD = 1*). As students' self-reported emotions are represented on an ordinal scale, we applied cumulative distribution functions assuming the latent variable of the measured ordinal outcome to be continuous [38]. We first applied the 'probit' link function which assumes the latent variable of the outcome to be normally distributed [38]. As the distributions for self-reported interest and boredom were highly skewed in our study, we also applied complementary log-log link distributions ('cloglog'), which are used for extreme-value distributions and allow for asymmetry [39]. As the cloglog link distribution performed better (according to leave-one-out cross-validation), we only report models based on cloglog link distributions for valence and boredom. As self-reported valence was relatively symmetrically distributed, we kept the probit link function.

Calculations were done using the brms (Bayesian Regression Models Using 'Stan') package [38] within the statistical software R [40]. The parameter estimates were obtained using Markov Chain Monte Carlo sampling (MCMC) with four sampling chains of 4000 iterations with the first 2000 iterations per chain being omitted for warm-up. To ensure convergence, we checked for no divergent transitions and that all R-hat values were ≤ 1.01. We applied weakly informative priors on all estimated parameters. The regression coefficients are reported at the mean of their posterior distribution which is comparable to the frequentist point estimate [41]. The Bayesian framework allows estimating the relative credibility of a parameter value given the data. Accordingly, we report the range which contains the true value of the regression coefficients with a probability of 95% (95% credible intervals: CrIs). R code of the statistical analyses, detailed results, and plots are available on the Open Science Framework (OSF) at https://osf.io/rynse/.

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 after reading the text; furthermore, the mean per event and the mean of the 10% of the highest values per event 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; cf. supplementary plots on OSF).

Table 2 presents the regression coefficients (mean of their posterior distributions) of the Bayesian generalized linear mixed models. None of the regression coefficients clearly differed from zero (no relationship) as assessed from their 95% credible intervals.[3] The results show the lack of a clear indication of a relationship between FaceReader's aggregated estimates and the self-reports. None of the aggregation methods of FaceReader's estimates clearly predicted students' emotional self-reports. This was the case for all three emotional states: interest, boredom, and valence.

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

	Self-report after	FaceReader mean	FaceReader mean of peak values
Interest	3.80 (0.93; 1-5)	0.01 (0.04; 0-1)	0.03 (0.08; 0-1)
Boredom	1.44 (0.79; 1-5)	0.06 (0.15; 0-1)	0.32 (0.28; 0-1)
Valence			pos: 0.14 (0.18; 0-1)
	6.94 (1.13; 1-5)	−0.12 (0.15; 0-5)	neg: −0.28 (0.21; −1-0)

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

4 Discussion

The objective of this study was to compare students' self-reports with automated facial expression based recognition of interest, boredom, and valence in reading situations. We investigated whether aggregated scores of 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 epistemic emotions (*i.e.*, boredom and interest), nor the estimates of valence predicted the respective self-report measure. Most regression coefficients are distributed around zero pointing to the lack of a (linear) relationship between facial emotion recognition and self-reported emotions. The results indicate a lack of a clear indication of a relationship between FaceReader's aggregated estimates with the self-reports and effect sizes are far from what we expected.

The lack of a strong coherence between facial emotion recognition and self-reported emotions in our study may be less surprising to some – arguing that different emotional components do not need to strongly cohere (*e.g.*, [6, 42]). We, however, argue that there should be some form of coherence as the different emotional components relate to the same overall construct and emerge from the same appraisals. The relationship may not be extremely strong or linear, but without any visible coherence (at least not to us), we would be thankful for some further ideas for what relationships to look for. Moreover, if no relationships are found between some emotional components, we would favor some further reasons to why and how they should still relate to the same overall construct.

[3] In the interest models, we detected and omitted two highly influential events – both from one participant. The individual mean and individual peak were more than four standard deviations above the general mean of those estimates.

Table 2: Overview of the regression coefficients of the Bayesian generalized linear mixed models

	Model with aggregation method for FaceReader's estimates	Regression coefficient	Credible interval	Number of observations
Interest	mean as predictor	b = −0.17	CrI = [−0.71, 0.38]	obs = 203
	interaction of mean and SD of mean as predictor	b_mean*sd = −0.05	CrI = [−0.60, 0.58]	
		b_mean = −0.35	CrI = [−2.38, 1.59]	
		b_sd = 0.21	CrI = [−0.44, 0.88]	
	mean over the 10% of the highest values as predictor	b = −0.07	CrI = [−0.44, 0.29]	
Boredom	mean as predictor	b = −0.10	CrI = [−0.47, 0.23]	obs = 204
	interaction of mean and SD of mean as predictor	b_mean*sd = 0.38	CrI = [−0.03, 0.90]	
		b_mean = −0.25	CrI = [−1.02, 0.35]	
		b_sd = −0.03	CrI = [−0.61, 0.58]	
	mean over the 10% of the highest values as predictor	b = −0.16	CrI = [−0.53, 0.16]	
Valence	mean as predictor	b = −0.01	CrI = [−0.30, 0.28]	obs = 193
	interaction of mean and SD of mean as predictor	b_mean*sd = −0.13	CrI = [−0.43, 0.16]	
		b_mean = 0.16	CrI = [−0.31, 0.63]	
		b_sd = 0.03	CrI = [−0.23, 0.30]	
	mean over the 10% of the most extreme positive and negative values as two separate predictors	b_pos = 0.04	CrI = [−0.20, 0.28]	
		b_neg = −0.05	CrI = [−0.32, 0.23]	

Note: Coefficients are based on standardized predictors (FaceReader), but unstandardized outcomes (self-reports). Some missing self-reports reduced the sample size of specific analyses.

4.1 Construct and predictive validity of components of emotions

The lack of a relationship between facial expressions and self-reports does not give us information on which of the two measures are more valid. Although self-reports are traditionally more common, they should not necessarily be considered the 'true' benchmark. One way to further understand the two measures is to regard their distributions. Inspecting the means of interest, boredom, and valence indicates clear differences in intensity compared to FaceReader's estimates. Self-reports all differ from normal distributions in the direction of social desirability (higher interest, lower boredom, more positive valence). 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 skew in direction of social desirability raises the question if self-reports are a valid or useful instrument for measuring emotions at all [43]. Nevertheless, also FaceReader's estimates delivered some noteworthy distributions. For interest and also to some degree for boredom, FaceReader produced continuously many zero values, indicating a prevailing absence of those emotions (cf. supplementary plots on OSF). A possible limitation of this study is that the texts presented to the students were similar. All texts focused on psychological topics which is a possible explanation for the reduced variance found in students' emotional experience (self-reports) and expressions (FaceReader). Further studies are necessary to clearly determine the validity and practical usefulness of emotional self-reports and facial emotion recognition. The lack of an evident gold standard in measuring emotions clearly represents a huge challenge for this study as well as for the validation of affective computing in general.

Although affective computing generally is less involved in theoretical approaches towards emotions, statistical modeling of emotions already involves some assumptions about the concept of emotions. Often, affective computing includes supervised learning techniques for which a ground truth for emotions has to be defined. Some use external observers' ratings, others use people's self-reports, and again others experimental manipulations as ground truth [6]. Combinations of them are merely found. The chosen type of ground truth reflects assumptions on the nature of emotions. When the emotional components are indeed only loosely coupled, then different types of ground truth will result in very different predictions. 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 to understand the *predictive validity* of the emotional components for learning. It is crucial to understand which combination of measurement types of emotions predict learning behavior and performance most accurately – irrespective of what those measures actually represent. Instead of searching for a 'true' explanatory model of emotional components it might at this stage be more fruitful to investigate the predictive power of different emotional measurements. Which combination of, for example, physiological data, facial expression recognition, or self-report predicts learning behavior 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 log files) and algorithms might deliver estimates the most predictive of learning behavior or performance. Such evidence could help to understand in the dependence of which affective sensors learning instructions should be adapted so that learning is enhanced. We argue that the exact informational value of the different components as emotional indicators is still undetermined [4] and will presumably remain so if no evidence on their predictive validity is constructed. Accordingly, we propose to use learning indicators (such as a change in test performance or learning behavior) as ground truth for broad measures of specific epistemic emotions. Theories on the epistemic emotions deliver some indications on the expected learning behavior when intensities of these emotions are high versus low.

4.2 Specific issues in the computation of facial emotion recognition estimates

Testing facial emotion recognition in a reading situation is rather conservative as we expect students to be less expressive while reading alone compared to social learning contexts (see *e.g.*, [14]). 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). Such an approach might help to deal with the large inter-individual variability in the expression of emotions. 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 (*e.g.*, social context, the emotional content of the stimuli) and different formats of recording (*e.g.*, picture/video, quality). As emotional expressions seem to be more context-dependent and variable than commonly assumed [44] the algorithms should additionally be informed by details on the context and accordingly be trained in and for specific contexts.

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 (afterward). 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.

We aggregated FaceReader's estimates in different ways (*i.e.*, mean, standard deviation, mean of peak values). Feldmann Barrett and colleagues [45] proposed to analyse emotional patterns (and not only the overall mean). Therefore, rather than only focusing on the mere magnitude, we included – in a first attempt – the standard deviation as a measure of variance. Of course, more and different approaches are conceivable to compare continuous measures of emotions with point measures (*e.g.*, techniques stemming from growth curve modeling [46]). Basic experimental research might also help to find more sophisticated and creative ways to reveal a potential coherence between continuous measures of emotions and point measures. Such advanced techniques of comparison might promote the understanding of the relationship between the different components of emotions.

4.3 Conclusions for intelligent tutoring and adaptive learning systems

This study did not yield high agreement rates of automated facial emotion recognition with other forms of emotion measurement such as previous studies did [22, 47]. We assume that changes in the experimental setting (reading situation) and comparing the estimates with a different measurement type of emotions (self-reports) have led to a 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 relying on them in intelligent tutoring and adaptive learning systems. We agree with a recent review on emotional expressions [44] that it is premature to use automated facial expression recognition as a single indicator to infer people's emotions. As studies on didactics are traditionally based on self-reports of epistemic emotions, we conclude that assumptions from theories and empirical evidence focusing on the experience of emotions cannot necessarily be transferred to other emotional components such as estimated by affective computing.

For intelligent tutoring and adaptive learning systems, the accuracy and predictive power of the sensors used as the basis for adaptation is crucial for the performance of the whole system. Thus, the importance to develop useful 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 learning systems. Practitioners need tools at hand, which measure variables relevant to learning without impeding the learning process. Otherwise, how can intelligent learning and tutoring systems keep their promise to effectively adapt to a variety of students' needs?

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 (and their combinations) with regard to their predictive validity for different learning parameters in applied settings. We consider this the most fruitful approach to enable future applications of affective computing. Moreover, we argue that applications should be trained and tested in different contexts (*e.g.*, marketing, social interaction, reading). The relevance of the specific context of emotions has probably been underestimated in affective computing and research on emotional expressions in general [44]. Facial expressions of emotions have shown to be much more variable (*e.g.*, situation-dependent) than generally assumed [44]. Informing algorithms of facial emotion recognition with details on the specific context (*e.g.*, the content of speech, social features, personal conditions) seems, therefore, an important next step in developing a satisfactory performance of emotion recognition.

References

[1] Wu C.H., Huang Y.M., Hwang J.P., Review of affective computing in education/learning: Trends and challenges, British Journal of Educational Technology, 47(6), 2016, 1304–1323, 10.1111/bjet.12324

[2] Bosch N., D'Mello S.K., Ocumpaugh J., Baker R.S., Shute V., Using video to automatically detect learner affect in computer- enabled classrooms, 2016, 10.1145/0000000.0000000

[3] Wang C.H., Lin H.C.K., Constructing an Affective Tutoring System for Designing Course Learning and Evaluation, Journal of Educational Computing Research, 55(8), 2018, 1111–1128, 10.1177/0735633117699955

[4] Calvo R.A., D'Mello S., Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications, IEEE Transactions on Affective Computing, 1(1), 2010, 18–37, 10.1109/T-AFFC.2010.1

[5] Scherer K.R., What are emotions? And how can they be measured?, Social Science Information, 44(4), 2005, 695–729, 10.1177/0539018405058216

[6] D'Mello S.K., Kappas A., Gratch J., The affective computing approach to affect measurement, Emotion Review, 10(2), 2018, 174–183

[7] D'mello S.K., Kory J., A Review and Meta-Analysis of Multimodal Affect Detection Systems, ACM Computing Surveys, 47(3), 2015, 1–36, 10.1145/2682899

[8] Soleymani M., Mortillaro M., Behavioral and Physiological Responses to Visual Interest and Appraisals: Multimodal Analysis and Automatic Recognition, Frontiers in ICT, 5(17), 2018, 10.3389/fict.2018.00017

[9] Bosch N., D'Mello S., Mills C., What emotions do novices experience during their first computer programming learning session?, Technical report, 2013, 10.1007/978-3-642-39112-5-2

[10] Trigwell K., Ellis R.A., Han F., Relations between students' approaches to learning, experienced emotions and outcomes of learning, Studies in Higher Education, 37(7), 2012, 811–824, 10.1080/03075079.2010.549220

[11] Tze V.M.C., Daniels L.M., Klassen R.M., Evaluating the Relationship Between Boredom and Academic Outcomes: A Meta-Analysis, Educational Psychology Review, 28(1), 2016, 119–144, 10.1007/s10648-015-9301-y

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

[13] Moors A., Ellsworth P.C., Scherer K., Frijda N., Appraisal theories of emotion: State of the art and future development, Emotion Review, 5(2), 2013, 119–124

[14] Soutschek A., Weinreich A., Schuber T., Facial Electromyography reveals dissociable affective responses in social and non- social cooperation, Motivation and Emotion, 42(1), 2018, 118–125

[15] Amos B., Ludwiczuk Bartosz Satyanarayanan M., Openface: A general-purpose face recognition library with mobile applications, 2016, 10.5281/zenodo.32148

[16] Affectiva Homepage

[17] Noldus, Noldus Homepage

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

[19] Loijens L., Krips O., FaceReader Methodology Note. A white paper by Noldus Information Technology, Technical report, Amsterdam: Noldus, 2018

[20] Soleymani M., Detecting cognitive appraisals from facial expressions for interest recognition, preprint arXiv, 2016, arXiv:1609.09761v2

[21] Bonanno G., Keltner D., Brief Report The coherence of emotion systems: Comparing "on-line" measures of appraisal and facial expressions, and self-report, Cognition & Emotion, 18(3), 2004, 431–444, 10.1080/02699930341000149

[22] Lewinski P., den Uyl T.M., Butler C., Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader., Journal of Neuroscience, Psychology, and Economics, 7(4), 2014, 227–236, 10.1037/npe0000028

[23] Harley J.M., Bouchet F., Azevedo R., Aligning and comparing data on emotions experienced during learning with metatutor, in H. Lane, K. Yacef, J. Mostow, P. Pavlik, eds., Artificial Intelligence

in Education. AIED 2013. Lecture Notes in Computer Science, vol 7926, Springer, Berlin, Heidelberg, 2013, 61–70, 10.1007/978-3-642-39112-5-7

[24] Brodny G., Kolakowska A., Landowska A., Szwoch M., Szwoch W., Wrobel M.R., Comparison of selected off-the-shelf solutions for emotion recognition based on facial expressions, in 29th International Conference on Human System Interactions (HSI), IEEE, 2016, 397–404, 10.1109/HSI.2016.7529664

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

[26] Sneddon I., McRorie M., McKeown G., Hanratty J., The Belfast induced natural emotion database, IEEE Transactions on Affective Computing, 3(1), 2012, 32–41, 10.1109/T-AFFC.2011.26

[27] Pekrun R., Vogl E., Muis K.R., Sinatra G.M., Measuring emotions during epistemic activities: the Epistemically-Related Emotion Scales, Cognition and Emotion, 31(6), 2017, 1268–1276, 10.1080/02699931.2016.1204989

[28] Krapp A., Hidi S., Renninger A.K., Interest, learning, and development, in The role of interest in learning and development, Erlbaum, Hilsdale, NJ, 1991, 3–25

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

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

[31] Amstad T., Wie verständlich sind unsere Zeitungen?, Studenten-Schreib-Service, Zürich, 1978

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

[33] Mathôt S., Schreij D., Theeuwes J., OpenSesame: An open-source, graphical experiment builder for the social sciences, Behavior Research Methods, 44(2), 2012, 314–324, 10.3758/s13428-011-0168-7

[34] Grafsgaard J., Wiggins J.B., Boyer K.E., Wiebe E.N., Lester J., Automatically recognizing facial expression: predicting engagement and frustration, Educational Data Mining, 2013

[35] Kapoor A., Mota S., Picard R.W., Towards a learning companion that recognizes affect, Technical Report 543, 2001

[36] McDaniel B., D'Mello S., King B., Chipman P., Tapp K., Graesser A.C., Facial Features for Affective State Detection in Learning Environments Permalink, in Proceedings of the 29th Annual Cognitive Science Society, 2007, 467–472

[37] Lewinski P., 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), 2015, 1–9, 10.13140/RG.2.1.4653.6409

[38] Bürkner P.C., Vuorre M., Ordinal regression models in psychology: A tutorial, Advances in Methods and Practices in Psychological Science, 2(1), 2019, 251524591882319, 10.1177/2515245918823199

[39] Bürkner P.C., brms: An R package for Bayesian multilevel models using Stan, Journal of Statistical Software, 80(1), 2017, 1–28, 10.18637/jss.v080.i01

[40] R Core Team, R: A language and environment for statistical computing., 2018

[41] Heino M.T.J., Vuorre M., Hankonen N., Bayesian evaluation of behavior change interventions: a brief introduction and a practical example, Health Psychology and Behavioral Medicine, 6(1), 2018, 49–78, 10.1080/21642850.2018.1428102

[42] Scherer K.R., What are emotions? and how can they be measured?, Social Science Information, 44(4), 2005, 695–729, 10.1177/0539018405058216

[43] Zimmermann P., Guttormsen S., Danuser B., Gomez P., Affective computing - A rationale for measuring mood with mouse and keyboard, International Journal of Occupational Safety and Ergonomics, 9(4), 2003, 539–551, 10.1080/10803548.2003.11076589

[44] Feldman Barrett L., Adolphs R., Marsella S., Martinez A.M., Pollak S.D., Emotional expressions reconsidered: challenges to inferring emotion from human facial movements, Psychological Science in the Public Interest, 20(1), 2019, 1–68, 10.1177/1529100619832930

[45] Feldman Barrett L., Quigley K.S., Bliss-Moreau E., Aronson K.R., Interoceptive sensitivity and self-reports of emotional experience, Journal of Personality and Social Psychology, 87(5), 2005, 684–697, 10.1016/j.molcel.2009.10.020.The

[46] Rogosa D., Saner H., Longitudinal Data Analysis Examples with Random Coefficient Models, Journal of Educational and Behavioral Statistics, 20(2), 1995, 149–170, https://doi.org/10.3102/10769986020002149

[47] Lewinski P., Automated facial coding software outperforms people in recognizing neutral faces as neutral from standardized datasets, Frontiers in Psychology, 6, 2015, 1386, 10.3389/fpsyg.2015.01386