Universiteit Utrecht

FaceReader, a Promising Instrument for Measuring Facial Emotion Expression? A Comparison to Facial Electromyography and Self-Reports

Yannick T. Suhr

5947170

July 2017

Master Thesis, Clinical and Health Psychology (201500819)

Supervised by: Marloes Eidhof, MSc

Second evaluator: Prof. dr. Claudi Bockting

Faculty of Behavioural and Social Sciences

Utrecht University

Abstract

Various methods have been developed to examine human facial expressions, including a recent facial coding software known as FaceReader. Although a number of studies that examined its performance in identifying emotions reported promising results, the majority of studies relied on still images as emotion-inducing stimuli and exclusively examined FaceReader peak-values. The current study aimed to further examine FaceReader's emotion recognition by assessing the utility of measurements that were taken over periods of time. Furthermore, this study directly compared FaceReader to facial electromyography (fEMG), a well-established method of measuring facial expressions, as well as to self-reports of emotion experience. In a repeated-measures and within-subject design, the emotions of sadness, disgust and fear were induced using video stimuli. The facial reactions of 26 participants were video recorded, while changes in facial muscle activity associated with the expression of the emotions were recorded using fEMG. Instead of peak values, the current study analysed the average measurements that were recorded during each clip. The video-clips were repeatedly presented, which was expected to result in a decrease in emotional reaction to allow evaluation of the instruments' performances when emotion intensity decreases. The performance of both FaceReader and fEMG was inconsistent for all three emotions. However, FaceReader appeared to have a bias to identify neutral facial states as expressing sadness. Limitations of the current study that prevent from definite conclusions about the performance of the instruments are pointed out and are followed by suggestions for future research.

Keywords: FaceReader; facial electromyography (fEMG); emotion recognition; automated facial coding software (AFC)

Table of Contents

FaceReader, a Promising Instrument for Measuring Facial Emotion Expression? A Comparison to Facial Electromyography and Self-Reports

The phenomenon of human emotion has been a topic of interest in the scientific world since the early dawn of psychological research (Darwin, 1872; James, 1884; Wundt, 1897). However, a collectively agreed upon definition or theory of emotion still fails to exist in the current day, despite several attempts to define the concept (Kleinginna & Kleinginna, 1981; *see also* Scherer, 2000). As such, a clarification of how emotion can be defined and understood remains an important aspect in the research of emotions (Mulligan & Scherer, 2012). Nowadays, emotion is commonly understood as a complex system, consisting of physiological, experiential and behavioural responses (Mauss & Robinson, 2009). Notably, the complexity of the topic has led to the emergence of numerous methods and instruments that are available to assess emotions.

In earlier days, scientists interested in investigating emotions had to rely mostly on self-reports and observations (e.g., Wallbott & Scherer, 1989). However, the range of available methods has since then rapidly expanded. In addition to using interviews or self-report measures to assess the subjective experience of emotions (e.g., Hofmann, Carpenter, & Curtiss, 2016), researchers with an interest in the more unconscious aspect may rely on physiological measures, such as heart rate or skin conductance (Thomas, Leeson, Gonsalvez, & Johnstone, 2014). Furthermore, technological advances in the last decades have introduced instruments that allow to investigate brain activity thought to be associated with the experience or expression of emotions. Electroencephalograms (EEG) can be used to track brain waves linked to the subjective experience of emotions (Ackermann, Kohlschein, Bitsch, Wehrler, & Jeschke, 2016; Jenke, Peer, & Buss, 2014), while functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) are commonly used to explore the functional neuroanatomy of emotions (Phan, Wager, Taylor, & Liberzon, 2002).

Facial electromyography (EMG) is one of the most commonly used methods to study the expressive component of human emotions (Cohn, Schmidt, Gross, & Ekman, 2002; Mavratzakis, Herbert, & Walla, 2016). Small surface electrodes that are placed onto facial muscles associated with the expression of certain emotions measure tiny electrical impulses that are elicited at every instance of muscle activity (Hess, 2009). Although not as costly as fMRI and PET and not requiring medically trained personal, the application of EMG still involves expensive equipment and is typically bound to a laboratory setting (Wolf, 2015). Not only is it likely that the emotions evoked in such a setting may differ from those occurring in real life, but a heightened awareness of being observed may further influence the expression of emotions (Benţa et al., 2009). Despite these limitations that may restrain the generalizability of results, facial EMG remains a popular instrument in research on the expressive component of emotion (e.g., Mavratzakis et al., 2016; Thompson, Mackenzie, Leuthold, & Filik, 2016).

To overcome these limitation, Vicar Vision and Noldus Information Technology developed *FaceReader*, a commercially available automated facial coding software (AFC) that aims to identify human facial expressions (den Uyl & van Kuilenburg, 2005; Noldus, 2017). Among others, FaceReader claims to measure the six basic emotions of happiness, sadness, fear, anger, surprise, and disgust (Ekman, 1982; Ekman & Cordano, 2011). Lewinski, den Uyl, and Butler (2014) argue that the results of a FaceReader analysis can be considered reliable, as the algorithm behind it will always lead to the same outcome when applied to the same data. They further state that FaceReader is built on pre-existing knowledge and theories of emotion of the past decades, starting with the seminal paper on facial displays of emotion by Ekman, Sorenson, and Friesen (1969). As FaceReader does not involve any theoretical interpretations of its own (Lewinski et al., 2014), it can be considered an objective tool that is, arguably, free from subjective bias. Furthermore, AFC is both time

and cost-effective, as it requires only a high-resolution camera and a minimal amount of manual labour. Contrary to facial EMG, which requires being connected to a computer in a laboratory setting, FaceReader can be applied in a non-invasive way in a variety of settings. Therefore, it provides the opportunity to investigate emotions occurring in real-life situations and may additionally lend itself to the use of clinicians in certain therapeutic situations.

Since its first appearance in 2005 (den Uyl & van Kuilenburg, 2005), FaceReader has been applied in a number of studies. Terzis, Moridis, and Economides (2011) compared momentum FaceReader measurements (peak values) to subjective ratings of the experimenters who observed participants' emotional expressions. They found that FaceReader was in accord with the subjective ratings of the experimenters in 87% of the cases. Similarly, Benţa and colleagues (2009) found that the FaceReader scoring of facial expressions was in line with psychologist-evaluations of the same facial expressions. However, certain methodological issues present in the study may explain why the correlations between FaceReader scoring and psychologist ratings were smaller than expected. Firstly, FaceReader's scores were based on short video recordings of participants' facial reactions to stimuli, while the psychologists rated still images of the faces at the moment in which FaceReader recognized a 'peak' emotional expression. Furthermore, the stimuli in the study used to elicit emotions were pictures that were shown for only 4.5 seconds and immediately followed one another. This rapid presentation of the stimuli may have resulted in the first image influencing the emotion elicited by the following image.

A variety of fields have applied FaceReader as a tool for facial expression recognition, including fields such as educational science (Terzis, Moridis, & Economides, 2011; 2012), human-computer interaction (Goldberg, 2014), and consumer behaviour and marketing research (Danner, Sidorkina, Joechl, & Duerrschmid, 2014; Lewinski, Fransen, & Tan, 2014). However, despite the number of applications, the majority of studies that

assessed FaceReader only examined instantaneous emotions, most often evaluating the congruence between FaceReader peak values and subjective observer ratings of images depicting human facial expressions. The utility of FaceReader measurements that are taken over longer periods of time has yet to be examined, and research comparing FaceReader to other technical methods of assessing emotional facial expression is sparse (D'Arcey, 2013). Therefore, the main purpose of the current study was to investigate an alternative to the use of peak values, namely by analysing average facial expression of emotions over time. Furthermore, instead of comparing it to subjective ratings by observers (e.g., Chóliz & Fernández-Abascal, 2012; Lewinski, 2015; Terzis et al., 2013), this study compared FaceReader to facial EMG, allowing a direct comparison to another objective measure of facial expression. Unlike previous studies that used still images as emotion-inducing stimuli (e.g., Benţa et al., 2009; Bernhaupt, Boldt, Mirlacher, Wilfinger, & Tscheligi, 2007; Lewinski et al., 2014), the current study made use of video stimuli, as research suggested it may be superior to image stimuli in inducing strong emotional reactions (Horvat, Kukolja, & Ivanec, 2015). A direct comparison to the subjective experience of participants was provided through the use of self-report measures, which in return also allowed an evaluation of whether the stimuli in fact induced an emotional experience. As one of the incentives of this study was to evaluate whether FaceReader could potentially become a useful tool in clinical practice, it was decided to focus on emotions of negative valence. Our extensive search for emotion-inducing videos that have been used in previous studies and shown to elicit the according emotion discretely yielded in videos for sadness, disgust and fear. Lastly, as facial EMG is regarded as a sensitive measure that is able to detect even subtle facial muscle activity (Wolf, 2015), we aimed to explore in how far this also holds true for FaceReader. Therefore, every video stimulus was presented for a repeated number of times, which was expected to

desensitize participants, and hence result in weaker emotional reactions over repeated presentations.

Methods

The current study employed a repeated-measures and within-subject design that assessed the effectiveness of FaceReader and facial EMG in measuring facial expressions of emotions. Additionally, a self-report measure was included that served as a measure of subjective emotion experience, as well as a manipulation check for the induction of emotion. The proposal of the current study was approved by the local ethical committee.

Participants

Female undergraduate students were recruited at the Faculty of Behavioural and Social Sciences at Utrecht University. Participants provided written informed consent prior to the experiment, and chose between course credit or a 4-Euro compensation for participation. The decision to include only female participants stemmed from two main reasons. Firstly, it has been suggested that women tend to report emotions more intensely (Bradley, Codispoti, Sabatinelli, & Land, 2001) and are generally more emotionally expressive than men (Lench, Flores, & Bench, 2011). Secondly, as proper skin preparation is indispensable for accurate EMG measurements (Hermens, Freriks, Disselhorst-Klug, & Rau, 2000), problems with the attachment of electrodes to participants with excessive facial hair were avoided by excluding male subjects. Having previous knowledge of facial EMG was a further exclusion criterion, as it was suspected to result in unnatural behaviour, such as endorsement or suppression of certain emotional reactions to stimuli. Lastly, requiring glasses was an exclusion criterion, as wearing glasses has reportedly led to complications in FaceReader analyses (Alitalo, 2016; Terzis et al., 2011). This limitation has also been acknowledged in the FaceReader Reference Manual (Loijens, Krips, van Kuilenburg, den Uyl, & Ivan, 2015).

Measures of emotional experience

FaceReader. FaceReader is a fully automated facial expression recognition software that identifies 491 key points within the human face and analyses changes in skin tone and location of the key points according to emotional behaviour prototypes (Ekman, Friesen, & Ellsworth, 2013; Loijens et al., 2015). The analysis takes into account facial muscle activity, gaze direction, head orientation and other subject characteristics, and reports the presence or absence of an emotion on a scale from 0 to 1. The current study used FaceReader™ 7 set to provide measurements every 40ms using the default face model ('General'). For all analyses, standard settings were used, meaning that neither manual nor continuous calibration were applied. A detailed description of FaceReader and the algorithm at hand can be found elsewhere (*viz.,* van Kuilenberg, Wiering, & den Uyl, 2005).

Facial electromyography (facial EMG). Facial EMG is a technique that allows the observation of facial expressions by recording facial muscle activity with surface electrodes. We acquired muscle activity bipolarly through the use of sintered (Ag/AgCl) 4-mm skin electrodes. The attach-spaces were filled with conductive gel, and a 12-mm reference electrode was placed behind the ear near the mastoid bone (De Luca, 2002; Fridlund & Cacioppo, 1986). A 50 Hz notch filter decreased power interference in the raw EMG signals, which were further rectified and smoothed using a 20 Hz low-pass filter with a time interval of 100 ms. EMG signals were logged with MindWare software (EMG 3.0.21; MindWare Technologies, Gahanna, USA). Intending to examine the facial expression of sadness, disgust and fear, electrodes were placed onto the *depressor anguli oris*, the *levator labii superioris,* and the *medial frontalis* (de Groot, Smeets, Kaldewaij, Duijndam, & Semin, 2012; van Boxtel, 2010; Whitton, Henry, Rendell, & Grisham, 2014). Electrodes were applied following standard procedures for surface electrodes as extensively described by Fridlund and Cacioppo (1986).

Subjective emotional experience. Visual analogue scales that ranged from ‘not at all’ to ‘very much’ were used to assess the three emotions of interest. The scales additionally assessed surprise, happiness and anger, which were not of primary interest, but were used as decoys to reduce the potential threat of demand characteristics (Orne, 2009). Inclusion of these items further minimized chances of other response biases, such as that participants would focus too much on a single emotion per video or respond in a socially desirable way (Grimm, 2010).

Materials

Video stimuli. The emotion-eliciting stimuli were four video-clips of approximately 30-seconds that were cut from commercially available movies. Three of the videos sought to elicit one emotion of interest, namely sadness, disgust, or fear, while the fourth video was a ‘neutral’ stimulus that was intended to serve as a control. Scene descriptions as well as studies that used the videos and reported them to reliably and discretely elicit the corresponding emotion can be found in Table 1. The clips were presented with the original sound on speakers. The sadness-evoking video was the only clip that contained spoken language (English), however, the content was not considered to be essential to the inducement of the emotion.

Experimental setup. At a viewing distance of approximately 80 cm, the experiment was presented on a computer screen in a well-lit laboratory (Windows 7 PC with a 19’’ Eizo Color LCD Monitor). A camera mount with a static Sony SRG300H camera was placed behind the computer screen at a height to capture the participants’ faces. The presentation of the experiment was controlled by an E-Prime script (E-Prime 2.0; Schneider, Eschman, & Zuccolotto, 2002).

Table 1

Descriptives of the video-clips used as emotion-inducing stimuli in this study

Movie and scene description	**Target emotion**	**Length (min:sec)**	**Studies that previously used this video**
The Champ (1979). An injured and dying man lies on a table stumbling his last words to a young boy before he passes away. The boy is devastated, starts crying, and yells at the deceived man to wake up.	Sadness	0:30	Gross & Levenson (1995)
Pink Flamingos (1973). A drag queen sits next to a defecating dog, picks up the faeces and starts eating them. Nearly vomiting, she continues to eat, smiles at the camera and shows her teeth which are covered in faeces.	Disgust	0:32	Gross & Levenson (1995)
The Lover (1992). The scene shows a street where a young lady enters the back of a car. The scene ends with the car driving away.	Neutral	0.30	Schaefer, Nils, Sanchez, & Philippot (2010)
Copycat (1995). A woman walks into a public bathroom where a dead police officer lays in the corner. Holding a gun, she searches the bathroom stalls, while the observer can see how the seemingly dead officer rises in the background and chokes her from behind.	Fear	0.34	Schaefer, Nils, Sanchez, & Philippot (2010)

Note. The exact times and frames at which the scenes were cut from the original movies will be provided upon request.

Procedure

The current study was conducted in a laboratory at the Utrecht University campus. Upon arrival, participants were verbally screened based on the exclusion criteria and were given a general description of the experiment. The assessment of emotional expression was not explicitly mentioned, but it was described that the study involved an assessment of 'internal processes'. Participants were told that the video recordings would be transformed into numerical data. To visualize this, a random FaceReader analysis output was presented. EMG electrodes were applied to the left side of the face, as this side has been suggested to express emotions more intensely (Blackburn & Sirillo, 2012; Dimberg & Petterson, 2000).

Participants were then told that they could stop the experiment at any point and that, although the experimenter would remain in the same room, questions should not be asked during the task as it would interfere with measurements. The experimenters remained in the room as literature suggested that emotions may be expressed to a larger degree when in a social context (Fischer, Maanstead, & Zaalberg, 2011).

The starting screen of the experiment explained that videos would be repeatedly shown in a random order. Each video was presented once in a series of five blocks, while the order of presentation was randomized and counterbalanced both between blocks and participants. Between clips, a black screen was displayed for 10 seconds, intended to prevent carry-over effects (Nonyane & Theobald, 2007; Benţa et al., 2009). Upon finishing the video task, a brief description appeared on screen that explained how indications on the following visual analogue scales work. Participants were then required to rate the emotions experienced during the last and first viewing of each video-clip. Assessing the last segment first was intended to prevent a high initial estimate of emotion for the first viewing that would possibly affect the estimate of the final viewing (Mussweiler, Englich, & Strack, 2016). The order in which videos had to be rated was randomized. Participants were then provided with some questions assessing demographic data and previous knowledge of content of the study. Lastly, electrodes were removed and participants received a debriefing. The overall duration from entering the laboratory to leaving again was approximately 35 minutes.

Plan of analyses

Standardized mean scores for the expression of sadness, disgust and fear, as measured by FaceReader (FR), facial EMG (fEMG) and visual analogue scales (VAS), were computed for each video presentation. In the following sections, a single viewing of a video will be referred to as a segment. Helmert contrasts were conducted to assess whether the video-clips evoked the particular emotion they were intended to evoke over and above the other

emotions. Helmert contrasts always compare the first variable versus the mean of the remaining variables in the analysis. Expecting a desensitization over repeated measurements (Campbell, O'Brien, van Boven, Schwarz, & Ubel, 2014), these contrasts were conducted only for the first segment of each clip, assuming initial reactions to be strongest. In order to evaluate whether the neutral video induced any emotion in particular, separate Helmert contrasts were conducted examining each of the three emotions during the first segment.

The main analysis employed a 2 x 3 repeated-measures analysis of variances (ANOVA) per emotion-inducing video to investigate how well the different measurement tools performed when the intensity of the emotion decreased. The ANOVAs included three levels of measurement (FR, fEMG, and VAS), while the factor time had two levels (time point 1 and 5). As VAS ratings were only provided for every first and last segment, FaceReader and fEMG measurements of only the first and last segments were used in the analyses. The emotion of interest in each of the repeated-measures ANOVAs corresponded to the emotion that each video was intended to evoke. Due to the exploratory nature of the study, the alpha level was set to $\alpha = .05$. Pairwise comparisons were used to further explore the performances of the instruments. All analyses were conducted using IBM SPSS 22.0.

Throughout the results section, muscle activity in the depressor anguli oris, the levator labii superioris, and in the medial frontalis is referred to as facial expressions of sadness, disgust and fear, respectively, as measured by facial EMG.

Results

From the initial sample of 33 participants, data of 7 subjects was omitted. FaceReader data was incomplete in the case of four subjects, as FaceReader reported failed fit model during some of the segments, and facial EMG data of three participants was flawed due to measurement error. The final sample consisted of 26 subjects that provided complete data for FaceReader, facial EMG and self-report measures (M_{age} = 23, *SD* = 2.53, *age range:* 19-30).

Helmert contrasts conducted on the VAS ratings suggested that the videos evoked the particular emotion they were intended to evoke discretely (Sadness: $F(1, 25) = 39.48$, $p < .001$, $\eta_p^2 = .612$; Disgust: $F(1, 25) = 218.82$, $p < .001$, $\eta_p^2 = .897$; Fear: $F(1, 25) = 39.72$, $p < .001$, $\eta_p^2 = .614$). The sadness video-clip was the only clip for which FaceReader and facial EMG were in line with the VAS ratings (FR: $F(1, 25) = 208.52$, $p < .001$, $\eta_p^2 = .893$; fEMG: $F(1, 25) = 4.79$, $p = .038$, $\eta_p^2 = .161$). For both the disgust and the fear-inducing clip, FaceReader measurements were significantly higher for sadness than for the other emotions (Sadness in disgust-clip: $F(1, 25) = 92.79$, $p < .001$, $\eta_p^2 = .788$; Sadness in fear-clip: $F(1, 25) = 269.36$, $p < .001$, $\eta_p^2 = .915$). This also held true for facial EMG measurements of sadness during the disgust-evoking video-clip, $F(1, 25) = 6.83$, $p = .015$, $\eta_p^2 = .214$. During the fear clip, EMG did not detect any emotion over and above the other emotions.

Helmert contrasts conducted on the neutral video indicated that FaceReader identified sadness to a significantly larger extent than fear or disgust, $F(1.26, 31.56) = 391.24$, $p < .001$, $\eta_p^2 = .940$. This was also the case for facial Emit G, although the estimated effect size was considerably smaller, $F(1, 25) = 5.54$, $p = .027$, $\eta_p^2 = .181$. These findings were not supported by VAS ratings, with the neutral video being perceived as inducing fear more than either of the other emotions, $F(1, 25) = 8.68$, $p = .007$, $\eta_p^2 = .258$. An overview of which emotions were identified per video by each measurement can be found Table A1 (Appendix A).

The first repeated-measures ANOVA did not reveal a significant interaction effect for time and the expression of sadness as measured by the three measurement modalities during the sadness-inducing clip, $F(2, 50) = 2.56$, $p = .087$. However, a main effect was present across instruments, $F(2, 50) = 35.89$, $p < .001$, $\eta_p^2 = .589$. Subsequent pairwise comparisons indicated that overall facial EMG measurements of sadness ($M = .119$, $SD = .765$) were significantly lower than the overall measurements provided by FaceReader ($M = 1.353$, $SD = .313$), $t(25) = 7.11$, $p < .001$, $d = 2.112$. A visualization of the individual measurement

performances in gauging the emotion of interest per video can be found in Figure 1. A numerical overview of these measurements can be found in Table B1 (Appendix B).

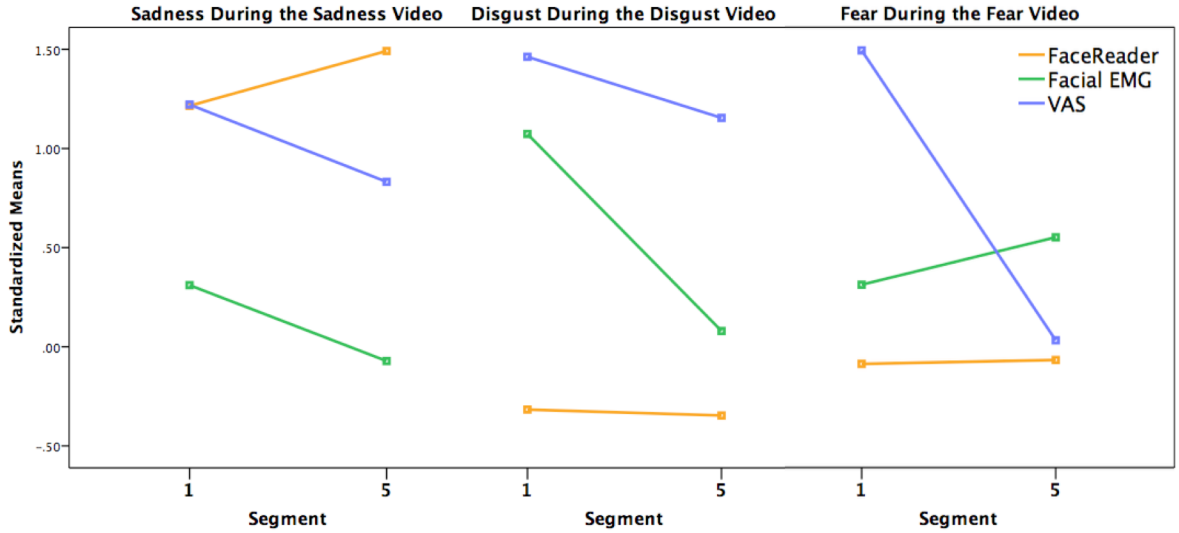

Figure 1. Visualization of the measurements of the emotion of interest per video and instrument.

The second repeated-measures ANOVA investigated disgust during the disgust-inducing clip. The Greenhouse-Geisser correction was applied throughout this analysis, as Mauchly's test indicated a violation of the assumption of sphericity for main effects of measurement instrument, χ^2 (2) = 16.76, $p < .001$, and interaction effects χ^2 (2) = 14.57, $p < .001$. The analysis revealed a significant interaction effect, $F(1.36, 32.67) = 12.69$, $p < .001$, $\eta_p^2 = .346$, and a significant main effect for measurement modality, $F(2, 31.63) = 115.27$, $p < .001$, $\eta_p^2 = .710$. FaceReader measurements of disgust were stable across time, however, it is crucial to point out that the unstandardized disgust scores by FaceReader, which were provided on a scale from 0 to 1, were extremely close to zero for both segments (unstandardized $M_1 = .027$, $SD_1 = .021$; $M_2 = .019$, $SD_2 = .018$). Facial EMG identified high levels of disgust during the first segment, but not during the last (for an overview of means, see Table A1 in the appendix). Although participants reported the video-clip to be less disgust-inducing during the last segment, VAS ratings for the last segment were still high

relative to facial EMG measurements. Specified contrasts indicated this difference in decrease between facial EMG and VAS to be significant, t(24) = 2.74, p = .011, d = .730.

As a Mauchly's test indicated a violation of the assumption of sphericity for measurement modalities during the fear-inducing clip, χ^2 (2) = 21.14, p < .001, the Greenhouse-Geisser correction was applied. A main effect was found across measurement instruments, F(1.26, 31.54) = 15.69, p < .001, η_p^2 = .386. Furthermore, an interaction effect between time and fear measurements was found, F(1.79, 44,73) = 54.13, p < .001, η_p^2 = .684. Among the three measurement modalities, FaceReader provided the overall lowest measures of fear, but as in the case of the disgust-inducing clip, the unstandardized scores for fear provided by FaceReader were close to zero (unstandardized M_1 = .046, SD_1 = .059; M_2 = .049, SD_2 = .057). Participants rated the clip to be fear-inducing when viewed for the first time, but not when presented for a repeated time, t(25) = 8.45, p < .001, d = 2.405. Facial EMG measurements of fear during the first segment were higher than FaceReader, but still significantly lower than VAS ratings, t(25) =-5.40, p < .001, d = 1.492.

Initially planned contrasts intended to examine the performance of FaceReader and facial EMG when the intensity of emotion expression decreases were omitted in this analysis for two reasons. Firstly, we did not observe the decreases in performance that we expected (i.e., FaceReader even observed an increase in sadness over repeated viewings of the sadness-clip). Secondly, FaceReader merely measured any levels of fear and disgust during the corresponding clips. In absence of discernable measurements of emotion, employing analyses that examine the differences in decreases of FaceReader and EMG performance was meaningless.

Discussion

In this multi-method study design, the performance of FaceReader and facial EMG in identifying facial expressions was examined and compared to self-reports. The manipulation

was found to be effective as the subjective ratings of emotions suggested that the video stimuli elicited the emotions they were intended to elicit. This additionally supports the findings of previous studies that made use of these stimuli and suggested that the clips are able to effectively and discretely evoke the specific emotions (Gross & Levenson, 1995; Schaefer et al., 2010). Furthermore, the anticipated habituation effect in response to repeated presentation of the clips was observed for all three videos, although the disgust clip appeared to still induce high levels of disgust at the later presentation. Both FaceReader and facial EMG identified sad facial expressions during the sadness-inducing video, but the intensity detected by facial EMG was much lower. For the first segment, FaceReader observed similar levels of sadness as reported through VAS, however, FaceReader identified an increase in sadness over the course of repeated presentation. This finding contradicts the assumption of desensitization, and further opposes the subjective ratings given by the participants. Benţa and colleagues (2009) suggested that repeated presentation of stimuli with negative emotional valence may lead to emotions being added up. This effect would result in the measurements not exclusively reflecting the reaction to the current stimulus, but in fact (to some extent) represent the experience of all the presented stimuli prior to the current point. However, if this effect would have occurred, one would have expected to find similar increases in the self-report and EMG measures, which was both not the case.

Surprisingly, we observed rather different scenarios for the disgust and fear-inducing videos. FaceReader measurements of disgust and fear were extremely small. This finding seems peculiar, particularly in the case of the disgust clip, which was described by participants as the most intense stimulus, and has been described in literature as highly effective in inducing disgust (Gross & Levenson, 1995). Based on the premise that activity of the levator labii superioris reflects the expression of disgust to a certain extent (e.g., de Groot et al., 2012, van Boxtel, 2010), participants expressed disgust visibly during the first segment.

Being that the algorithm behind FaceReader is based on psychological theories and models of emotion expression (Lewinski et al., 2014), it appears questionable why FaceReader did not identify high levels of levator labii superioris activity as an expression of disgust. It is likely that FaceReader's detection of disgust may not solely depend on activity in this muscle, but that it also takes other muscle areas into account. For example, D'Arcey (2013) found that FaceReader's detection of disgust is highly correlated with activity in the *corrugator supercilii* (a muscle near the eye brow). As the current study did not include EMG measurements of this muscle, one can merely infer that the absence of FaceReader's recognition of disgust may be due to minor activity in this muscle.

Although EMG measures of disgust during the first segment were similar to the subjective ratings, EMG measures of disgust decreased strongly over repeated presentation. This decrease does not correspond to VAS ratings, which indicated that the clip was still highly disgust-inducing when watching it for the last time. Although the mere experience of an emotion does not imply that the emotion is necessarily visibly expressed (Butler, Lee, & Gross, 2007), it was unexpected to find such a large disparity between subjective and EMG measures. It therefore appears that perceived levels of disgust were simply not expressed to the same extent in the last segment in comparison to the first segment.

The standardization of intensity across stimuli has generally been argued to be problematic when using video stimuli (Lench, Flores, & Bench, 2011), an issue that may also have occurred in the current study. There seemed to be an agreement that the disgust stimulus was the most intense one, being rather extreme and inducing disturbing images to one's mind that may be difficult to fade out and forget. Literature has discussed the potential effects of disturbing images on the development of psychological distress (Ahern et al., 2002; Putnam, 2002), and even addressed the likelihood of acquiring forms of posttraumatic stress disorder (PTSD) through exposure to disturbing images in media (Pinchevski, 2016). We do believe

that the disturbance-factor played a role in accounting for the perceived high intensity of the disgust stimulus at the later segment.

The habituation effect appeared to be strongest in the fear-evoking clip, where participants reported the stimulus to be fear-inducing at first, but not at all when watching it for the last time. It seems reasonable to assume that watching a scary movie while knowing exactly what will happen next is unlikely to result in a fear response. Furthermore, considering models of classical conditioning (Fanselow & Sterlace, 2014), this finding is not surprising as repeated presentation of a fearful stimulus in absence of any aversive stimulus is expected to result in fear extinction. At the first presentation of the stimulus, facial EMG appeared to be capable of identifying fear responses. However, according to facial EMG, there was an increase in fear response from the first to the last segment, which seems peculiar in light of theoryies and models of fear (Myers & Davis, 2007), but also in consideration of the subjective ratings.

A number of limitations of the current study need to be addressed. Firstly, there is no existing method to assess the *true* intensity of emotions that were actually induced. Although self-reports are a convenient way to measure people's subjective experience of emotions, one cannot fully rely on these measures to represent the true extent to which emotions were experienced (Feldman Barrett, Quigley, Bliss-Moreau, & Aronson, 2005). Social desirability as well as other factors may play a role in responding and pose a threat to the validity of these measures (Krumpal, 2011). Particularly in clinical samples, self-reports may not always be an appropriate means (Dimaggio & Lysaker, 2014), which could complicate determining the effectiveness FaceReader in clinical investigations. Secondly, there were a number of factors that may have moderated the expression of emotions in this study. Having electrodes attached while simultaneously being filmed may result in an increased awareness of being observed. Previous studies have shown that people tend to modify their behaviour when they are aware

of being observed (e.g., Schwarz, Fischhoff, Krishnamurti, & Sowell, 2013), a phenomenon that is likely to have occurred in this study. Furthermore, being afraid that electrodes fall off when moving facial muscles could be a potential explanation for why perceived emotions were not expressed as visibly as as expected. Considering the relatively young sample and that the clips were selected from rather old and not well-known movies, we do not believe that prior experiences with the stimuli were a confounding factor in this study. Lastly, we do not believe that the attached EMG electrodes interfered with the quality of FaceReader analyses, as the 'deep face model' (incorporated in FaceReader 7) claims to allow analysis even when parts of the face are obstructed (Noldus, 2017).

One of the incentives to examine average values of facial expression instead of peak values stemmed from the fact that previous studies with FaceReader have exclusively examined momentum expressions of emotions, meaning that they used peak values. We suspected peak values of FaceReader and EMG to be prone to be confounded by minor involuntary movements, such as sneezing, coughing, or touching the face. In hindsight, the decision to examine mean values can be seen as the most substantial drawback of this study, as it is likely to have resulted in emotional facial expression to be severely underestimated. Emotional reactions have been described as being as short as a few seconds (Ekman, 1992), and the expressive component may even be of shorter duration. Hence, if individuals showed a strong facial expression, it is likely that this expression would hardly be noticeable when averaging over the full length of the video. The results of this study allow to conclude that one should refrain from using such full-length averages, as EMG measures were inconsistent, and FaceReader clearly failed to measure disgust and fear using this method. Although it performed well in measuring levels of sadness, the fact that FaceReader recognized levels of sadness at basically any point during the experiment gives reason to question the validity of these measurements (see Table A1). A recent study examined FaceReader and facial EMG,

but applied them separately to different samples (viz., Fanti, Kyranides, & Panayiotou, 2017). They reported facial EMG measurements to only inconsistently correspond to FaceReader measurements. However, they used peak values for FaceReader and mean values for EMG, a combination that prevents direct comparisons of the performances. Fanti and colleagues did not provide justification nor explanations for their decision to use such differing measures.

Combining peak and average values in a different manner could, however, be an alternative for both FaceReader and EMG measurements. For example, one could first determine the peak emotional expression and then average short intervals before and after the peak value. The use of such 'peak-averages' could prevent results from being confounded by unintended movements or other third variables, as potentially the case with pure peak values. We would like to point out that this study is far from having made full use of the potential of the current dataset. A secondary analysis could explore the instruments' performance when examining peak values, or the previously presented peak-averages. Furthermore, an analysis similar to the one in the current study could be conducted while employing the continuous calibration function in FaceReader 7. In light of FaceReader's tendency in this study to favour the emotion of sadness, such an analysis would shed light on how well the offered calibration function works and implements the needed adjustments.

Despite the poor performance of FaceReader in this experiment, there is reason to believe that automated facial recognition software has a promising future ahead. The vast majority of studies that made use of FaceReader reported it to perform reliably and accurately in measuring momentum emotion expression (e.g., Fanti et al., 2017; Lewinski, 2014). If facial emotion recognition software becomes even further validated, it could have important implications for both clinicians and researchers. Individuals differ in their ability to differentiate and identify emotions (Feldman Barret, Gross, Conner Christensen, & Benvenuto, 2001), and it is not always possible for a therapist to determine or recognize the

exact emotion a patient is feeling. Psychological treatments, such as exposure therapy, require the activation of the exact fear structure in order to overcome pathological fear (Rauch & Foa, 2006), which illustrates the importance of being able to recognize emotions accurately. An example of a potential form of therapy could rely on webcam recordings and FaceReader analyses in determining *hotspots* in PTSD patients. Hotspots are described as being the parts of traumatic memories that evoke severe levels of emotional distress, and are thought to be a key factor in successful treatment of PTSD (Nijdam, 2013). From a research perspective, validated and flexibly applicable automated facial coding software would mean that researchers no longer have to rely on the artificial induction of emotions in a laboratory setting. The expression of emotions in every-day life situations could be analysed, and the cumbersome use of EMG to measure facial expression would become redundant. However, the current state of evidence for FaceReader is not yet elaborate enough. For instance, there is a need to investigate how prone the commonly used peak values are to be affected by confounding variables. FaceReader's applicability and effectiveness needs to be scrutinized under diverse conditions and settings, as the majority of studies only examined emotions evoked in front of a computer. Furthermore, it appears that researchers affiliated with Noldus Information Technology were most often involved in previous empirical studies of FaceReader. This does not suggest that the results of these studies are not reliable or valid, but implies that future research on FaceReader should also include a certain level of independence of the developers of FaceReader.

The main goal of this study was to extend the existing literature on the utility of FaceReader in measuring facial emotion expression. This was carried out by directly comparing FaceReader to facial EMG measures, and relating these to self-reports. This study did not provide compelling evidence neither for nor against either of the two measurement instruments, which derives from limitations of this study that are likely to have had a

substantial impact on the results. Future research should account for the limitations of the current study, as well consider ideas and concepts suggested for future investigations in this endeavour. Automated facial coding software still has a long way to go to prove fully operational in clinical settings, but as has been discussed in this paper, the potential implications and opportunities it could have in health-care settings and research are worth investigating further.

References

Ackermann, P., Kohlschein, C., Bitsch, J.A., Wehrle, K., & Jeschke, S. (2016). EEG-based automatic emotion recognition: Feature extraction, selection and classification methods. *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. http://doi.org/10.1109/HealthCom.2016.7749447

Ahern, J., Galea, S., Resnick, H., Kilpatrick, D., Bucuvalas, M., Gold, J., & Vlahov, D. (2002). Television Images and Psychologicl Symptoms after the September 11 Terrorist Attacks. *Psychiatry, 65*(4), 289-300.

Benţa, K.I., van Kuilenburg, H. Eligio, U.X., den Uyl, M., Cremene, M., Hoszu, A. & Creţ, O. (2009). *Evaluation of a System for RealTime Valence Assessment of Spontaneous Facial Expressions*. Retrieved from: http://www.researchgate.net/publication/256081706 (06/2017)

Bernhaupt, R., Boldt, A., Mirlacher, T., Wilfinger, D., & Tscheligi, M. (2007). Using emotion in games: emotional flowers. *Proceedings of the international conference on Advances in computer entertainment technology*, 41-48. ACM.

Berri, C. (Producer), & Annaud, J. (Director). *The Lover* [Motion picture]. Saint-Denis, France: Fox Pathe Europa.

Blackburn, K., & Schirillo, J. (2012). Emotive hemispheric differences measured in real-life portraits using pupil diameter and subjective aesthetic preferences. *Experimental Brain Research, 219, 447-455*. http://doi.org/10.1007/s00221-012-3091-y

Bradley, M. M., Codispoti, M., Sabatinelli, D., & Lang, P. J. (2001). Emotion and motivation II: Sex differences in picture processing. *Emotion, 1*(3), 300–319.

Butler, E.A., Lee, T.L., & Gross, J.J. (2007). Emotion Regulation and Culture: Are the Social Consequences of Emotion Suppression Culture-Specific? *Emotion, 7*(1), 30-48. http://doi.org/10.1037/1528-3542.7.1.30

Campbell, T., O'Brien, E., Van Boven, L., Schwarz, N., & Ubel, P. (2014). Too much experience: A desensitization bias in emotional perspective taking. *Journal of Personality and Social Psychology*, *106*(2), 272-285. http://doi.org/10.1037/a0035148

Chóliz, M., & Fernández-Abascal, E.G. (2012). Recognition of emotional facial expressions: The Role of Facial and Contextual Information in the Accuracy of Recognition. Psychological Reports, 110(1), 338-350. http://doi.org/10.2466/07.09.17.PR0.110.1.338-350

Cohn, J.F., Schmidt, K., Gross, R., & Ekman, P. (2002). Individual Differences in Facial Expression: Stability over Time, Relation to Self-Reported Emotion, and Ability to Inform Person Identification. *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, 2002, 491-496. http://doi.org/10.1109/ICMI.2002.1167045

D'Arcey, J.T. (2013). *Assessing the validity of FaceReader using facial EMG* [Master's thesis]. Retrieved from: http://hdl.handle.net/10211.4/610 (06/2017)

Danner, L., Sidorkina, L., Joechl, M., & Duerrschmid, K. (2014). Make a face! Implicit and explicit measurement of facial expressions elicited by orange juices using face reading technology. *Food Quality and Preference*, *32*, 167-172. http://doi.org/10.1016/j.foodqual.2013.01.004

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. United Kingdom: John Murray.

De Groot, J.H.B., Smeets, M.A.M., Kaldewaij, A., Duijndam, M.J.A., & Semin, G.R. (2012). Chemosignals Communicate Human Emotions. *Psychological Science, 23*(11), 1317-1424. http://doi.org/10.1177/0956797612445317

De Luca, C.J. (2002). *Surface Electromyography: Detection and Recording.* Natick, USA: Delsys Corporated. Retrieved from: http://www.delsys.com/Attachments_pdf/WP_SEMGintro.pdf (06/2017)

Den Uyl, M.J., & van Kuilenburg, H. (2005). The FaceReader: Online facial expression recognition. In L.P. Noldus, F. Grieco, L.W. Loijens, & P.H. Zimmerman (Eds.), *Proceedings of Measuring Behavior 2005*, 589-590.

Dimaggio, G., & Lysaker, P.H. (2014). Metacognition and Mentalizing in the Psychotherapy of Patients With Psychosis and Personality Disorders. *Journal of Clinical Psychology, 71*(2), 117-124. http://doi.org/10.1002/jclp.22147

Dimberg, U., & Petterson, M. (2000). Facial reactions to happy and angry facial expressions: Evidence for right hemisphere dominance. *Psychophysiology*, *37*, 693–696.

Ekman, P. (1982). *Emotion in the human face* (2nd ed.). Cambridge: Cambridge University Press.

Ekman, P. (1992). Facial Expression and Emotion. *American Psychologist, 48*(4), 376-379.

Ekman, P., & Cordano, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364-370. http://doi.org/10.1177/1754073911410740

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). *Emotion in the human face: Guidelines for research and an integration of findings*. New York: Pergamon Press.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science, 164*(3875), 86-88. http://doi.org/10.1126/science.164.3875.86

Fanselow, M., & Sterlace, S.R. (2014). Pavlovian Fear Conditioning: Function, Cause, and Treatment. In F.K. McSweeney & E.S. Murphy (Eds.), *The Wiley Blackwell Handbook of Operant and Classical Conditioning*, Hoboken, USA: Wiley Blackwell.

Fanti, K.A., Kyranides, M.N., & Panayiotou, G. (2017). Facial reactions to violent and comedy films: Association with callous–unemotional traits and impulsive aggression. *Cognition and Emotion, 31*(2), 209-224. http://doi.org/10.1080/02699931.2015.1090958

Farnsworth, B. (2016). *Facial Action Coding System: A Visual Guidebook.* Retrieved from: http://imotions.com/blog/facial-action-coding-system/ (05/2017)

Feldman Barret, L., Gross, J.J., Conner Christensen, T., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion, 15* (6), 713–724. http://doi.org/10.1080/02699930143000239

Feldman Barrett, L., Quigley, K.S., Bliss-Moreau, E., & Aronson, K.R. (2005). Interoceptive Sensitivity and Self-Reports of Emotional Experience. *Journal of personal and Social Psychology, 87,*(5). http://doi.org/10.1037/0022-3514.87.5.684

Fischer, A.H., Maanstead, A.S., & Zaalberg, R. (2011). Social influences on the emotion process. *European Review of Social Psychology, 14* (1).

Foa, E.B. (2011). Prolonged exposure therapy: past, present, and future. *Depression and Anxiety, 28*, 1043-1047. http://doi.org/10.1002/da.20907

Fridlund, A.J., & Cacioppo, J.T. (1986). Guidelines for Human Electromyographic Research. *Psychophysiology, 23*(5), 567-589.

Goldberg, J. H. (2014). Measuring software screen complexity: relating eye tracking, emotional valence, and subjective ratings. *International Journal of Human-Computer Interaction, 30*, 518–532. http://doi.org/10.1080/10447318.2014.906156

Grimm, P. (2010). Social Desirability Bias. In J.N. Sheth & N.K. Malhotra (Eds.), *Wiley International Encyclopedia of Marketing*. Hoboken: Jon Wiley and Sons Ltd.

Gross, J.J., & Levenson, R.W. (1995). Emotion Elicitation Using Films. *Cognition and Emotion, 9*(1), 87-108. http://doi.org/10.1037/0021-843X.106.1.95

Hermens, H.J., Freriks, Bart, Disselhorst-Klug, Catherine, & Rau, G. (2000). *Journal of Electromyography and Kinesiology 10*, 361–374.

Hess, U. (2009). Facial EMG. In E. Harmon-Jones & J.S. Beer (Eds.), *Methods in Social Neuroscience* (70-91). New York, NY: Guilford Press.

Hofmann, S.G., Carpenter, J.K., & Curtiss, J. (2016). Interpersonal Emotion Regulation Questionnaire (IERQ): Scale Development and Psychometric Characteristics. *Journal of Cognitive Therapy and Research, 40*(3), 341-356. http://doi.org/10.1007/s10608-016-9756-2

Horvat, M., Kukolja, D., & Ivanec, D. (2015). Comparing affective responses to standardized pictures and videos: A study report. In *Proceedings of 38th International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2015* (1394-1398). http://arxiv.org/abs/1505.07398

James, W. (1884). What is an emotion? *Mind, 9,* 188-205.

Jenke, R., Peer, A., & Buss, M. (2014) Feature Extraction and Selection for Emotion Recognition from EEG. *IEEE Transactions on Affective Computing*, *5*(3), 327-339. http://doi.org/10.1109/TAFFC.2014.2339834

Kleinginna, P.R., & Kleinginna, A.M. (1981). A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition. *Motivation and Emotion, 5*(4), 345-379.

Krumpal, I. (2011). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*(4), 2025-2047.

Lench, H.C., Flores, S.A., & Bench, S.W. (2011). Discrete Emotions Predict Changes in Cognition, Judgment, Experience, Behavior, and Physiology: A Meta-Analysis of Experimental Emotion Elicitations. *Psychological Bulletin, 137*(5), 834-855.

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

Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: Validation of basic emotions and FACS AUs in FaceReader. *Journal of Neuroscience, Psychology, and Economics, 7*(4), 227-236. http://doi.org/10.1037/npe0000028

Lewinski, P., Fransen. M.L., & Tan, E.S. (2014). Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli. *Journal of Neuroscience, Psychology, and Economics, 7*(1). http://dx.doi.org/10.1037/npe0000012

Loijens, L., Krips, O., van Kuilenburg, H., den Uyl, M., Ivan, P., ... & Spink, A. (2011). *FaceReader Tool for automatic analysis of facial expressions. Reference Manual version 4*. Wageningen, The Netherlands: Noldus Information Technology.

Loijens, L., Krips, O., van Kuilenburg, H., den Uyl, M., & Ivan, P. (2015). *FaceReader 6.1. Reference Manual,* Wageningen, The Netherlands: Noldus Information Technology.

Lovell, D. (Producer), & Zeffirelli, F. (Director) (1979). *The Champ* [Motion picture]. Culver City, USA: MGM/Pathe Home Video.

Mauss, I.B., & Robinson, M.D. (2009). Measures of emotion: A review. *Cognition and Emotion 23*(2), 209–237. http://doi.org/10.1080%2F02699930802204677

Mavratzakis, A., Herbert, C., & Walla, P. (2016). Emotional facial expressions evoke faster orienting responses, but weaker emotional responses at neural and behavioural levels

compared to scenes: A simultaneous EEG and facial EMG study. *NeuroImage, 123*, 931-946. http://doi.org/10.1016/j.neuroimage.2015.09.065

Milchan, A., & Tarlov, M. (Producers), & Amiel, J., (Director) (1995). *Copycat* [Motion picture]. West Hollywood, USA: Regency Enterprises.

Mulligan, K., & Scherer, K.R. (2012). Toward a Working Definition of Emotion. *Emotion Review, 4*(4), 345-357. http://doi.org/10.1177/1754073912445818

Mussweiler, T., Englich, B., & Strack, F. (2017). Anchoring effect. In R.F. Pohl, *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking and Memory.* Abingdon, UK: Psychology Press.

Myers, K.M., & Davis, M. (2007). Mechanisms of fear extinction [Feature Review]. *Molecular Psychiatry, 12*, 120-150. http://doi.org/10.1038/sj.mp.4001939

Nijdam, M. J. (2013). Memory traces of trauma: Neurocognitive aspects of and therapeutic approaches for posttraumatic stress disorder. *UvA-DARE* (Digital Academic Repository). 's-Hertogenbosch: Uitgeverij BOXPress

Noldus (2017). *What's new in FaceReader 7.* Retrieved from: http://www.noldus.com/facereader/whats-new-facereader-7 (06/2017)

Nonyane, B.A., & Theobald, C.M. (2007). Design sequences for sensory studies: achieving balance for carry-over and position effects. B*ritish Journal of Mathematical and Statistical Psychology, 60*(2), 339-349. http://doi.org/10.1348/000711006X114568

Orne, M.T. (2009). Demand Characteristics and the Concept of Quasi-Controls. In R. Rosenthal & R. Rosnow (Eds.), *Artifacts in Behavioral Research* (pp. 110-137). Oxford: University Press.

Phan, K.L., Wager, T., Taylor S.F., & Liberzon, I. (2002) Functional Neuroanatomy of Emotion: A Meta-Analysis of Emotion Activation Studies in PET and fMRI. *NeuroImage, 16***,** 331–348. http://doi.org/10.1006/nimg.2002.1087

Pinchevski, A. (2016). Screen Trauma: Visual Media and Post-traumatic Stress Disorder. *Theory, Culture & Society*, *33*(4), 51-75. http://doi.org/10.1177/0263276415619220

Putnam, F.W., (2002). Televised Trauma and Viewer PTSD: Implications for Prevention [Peer commentary on the journal article "Television Images and Psychological Symptoms after the September 11 Terrorist Attacks"]. *Psychiatry, 65*(4), 310-312. http://doi.org/10.1521/psyc.65.4.310.20241

Rauch, S., & Foa, E.B. (2006). Emotional Processing Theory (EPT) and Exposure Therapy for PTSD. *Journal of Contemporary Psychotherapy, 36*, 61–65. http://doi.org/10.1007/s10879-006-9008-y

Scherer, K.R. (2000). Emotion. In M.Hewstone & W.Stroebe (Eds.), *Introduction to Social Psychology: A European perspective* (3rd ed., pp. 151-191). Oxford: Blackwell.

Schneider, W., Eschman, A., Zuccolotto, A. (2002). E-prime 2.0 [Computer software]. Pittsburgh, USA: Psychology SoftwareTools Inc.

Schwartz, D., Fischhoff, B., Krishnamurti, T., & Sowell, F. (2013). The Hawthorne effect and energy awareness. *PNAS, 110*(38), 15242-15246. http://doi.org/doi.org/10.1073/pnas.1301687110

Terzis, V., Moridis, C.N., & Economides, A.A. (2010). Measuring instant emotions during self-assessment test: the use of FaceReader. *Proceedings of Measuring Behavior 2010*, 192-195.

Terzis, V., Moridis, C.N., & Economides, A.A. (2011). Measuring instant emotions based on facial expressions during computer-based assessment. *Personal and Ubiquitious Computing, 17*(1), 43-52. http://doi.org/10.1007/s00779-011-0477-y

Terzis, V., Moridis, C. N., & Economides, A. A. (2012). The effect of emotional feedback on behavioral intention to use computer based assessment. *Computers & Education, 59(2),* 710-721. http://doi.org10.1016/j.compedu.2012.03.003

Thomas, S., Leeson, P., Gonsalvez, C., & Johnstone, S. (2014). Structural equation modelling to assess relationships between event-related potential components, heart rate and skin conductance in the context of emotional stimuli. *International Journal of Psychophysiology, 94*(2), 245-246. http://doi.org/10.1016/j.ijpsycho.2014.08.940

Thompson, D., Mackenzie, I.G., Leuthold, H., & Filik, R. (2016). Emotional responses to irony and emoticons in written language: Evidence from EDA and facial EMG. *Psychophysiology, 53*(7), 1054-1062. http://doi.org/10.1111/psyp.12642

Van Boxtel, A. (2010). Facial EMG as a Tool for Inferring Affective States**.** *Proceedings of Measuring Behavior 2010*, 104-108.

Van Kuilenburg, H., Wiering, M., & den Uyl, M. (2005). A model based method for facial expression recognition. *Science, 3720*, 194-205). Berlin: Springer-Verlag. http://doi.org/10.1007/11564096_22

Wallbott, H.G., & Scherer, K.R. (1989). Assessing emotion by questionnaire. In R. Plutchik & H. Kellerman (Eds.), *The measurement of emotions* (55-82). San Diego, USA: Academic Press.

Waters, J. (Producer/Director) (1973). *Pink flamingos* [Motion picture]. Stamford, USA: Lightning Video.

Whitton, A.E., Henry, J.D., Rendell, P.G., & Grisham, J.R. (2014). Disgust, but not anger provocation, enhances levator labii superioris activity during exposure to moral transgressions. *Biological Psychology, 96, 48-56.* http://doi.org/10.1016/j.biopsycho.2013.11.012

Wolf, K. (2015). Measuring facial expression of emotion. *Dialogues in Clinical Neuroscience, 17*(4), 457-462.

Wundt, W.M. (1897). *Grundzüge der physiologischen Psychology* [Outlines of Psychology]. Leipzig: Verlag von Wilhelm Engelmann.

Appendix A

Table A1

Overview of the individual emotions that each instrument identified per video and segment

	FaceReader		**Facial EMG**		**Visual Analogue Scale**	
Segment	1	5	1	5	1	5
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
Sadness Video						
Sadness	1.22 (0.60)	1.49 (0.56)	0.31 (1.29)	-0.07 (0.75)	1.22 (0.73)	0.83 (0.75)
Fear	-0.49 (0.37)	-0.53 (0.40)	-0.28 (0.85)	0.10 (0.95)	0.51 (0.88)	-0.20 (0.78)
Disgust	-0.75 (0.22)	-0.74 (0.35)	-0.30 (0.80)	-0.3 (0.77)	-0.20 (0.62)	-0.51 (0.56)
Disgust Video						
Sadness	0.97 (0.85)	1.45 (0.49)	1.77 (1.70)	0.19 (0.95)	-0.53 (0.33)	-0.42 (0.48)
Fear	-0.48 (0.39)	-0.57 (0.37)	0.14 (1.59)	0.12 (0.94)	0.09 (0.66)	-0.20 (0.79)
Disgust	-0.73 (0.17)	-0.77 (0.12)	1.38 (2.00)	-0.11 (0.74)	1.99 (0.62)	1.48 (0.85)
Fear Video						
Sadness	1.24 (0.58)	1.25 (0.52)	0.22 (0.94)	0.02 (0.82)	-0.36 (0.50)	-0.56 (0.45)
Fear	-0.59 (0.14)	-0.57 (0.14)	-0.19 (0.82)	0.05 (0.92)	0.99 (0.76)	-0.47 (0.40)
Disgust	-0.76 (0.19)	-0.69 (0.51)	-0.21 (0.85)	-0.21 (0.80)	0.07 (0.72)	-0.29 (0.57)
Neutral Video						
Sadness	1.13 (0.46)	1.27 (0.40)	0.37 (0.84)	-0.04 (0.85)	-0.59 (0.53)	-0.58 (0.47)
Fear	-0.63 (0.14)	-0.56 (0.36)	-0.30 (0.87)	-0.05 (0.79)	-0.19 (0.72)	-0.71 (0.50)
Disgust	-0.76 (0.12)	-0.77 (0.33)	0.01 (0.93)	-0.32 (0.82)	-0.67 (0.44)	-0.75 (0.55)

Note. SD = standard deviation. All means and standard deviations shown in this table reflect standardized values that were computed within participants and per measurement method.

Appendix B

Table B1

Overview of the different instruments' performance in measuring the emotion of interest per video type and segment

	Sadness in Sad-Video		**Disgust in Disgust-Video**		**Fear in Fear-Video**	
Segment	1	5	1	5	1	5
	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)	Mean (*SD*)
FaceReader	1.22 (0.60)	1.49 (0.56)	-0.73 (0.17)	-0.77 (0.12)	-0.59 (0.14)	-0.57 (0.14)
Facial EMG	0.31 (1.29)	-0.07 (0.75)	1.38 (2.00)	-0.11 (0.74)	-0.19 (0.82)	0.05 (0.92)
VAS	1.22 (0.73)	0.83 (0.75)	1.99 (0.62)	1.48 (0.85)	0.99 (0.76)	-0.47 (0.40)

Note. SD = standard deviation. Means and standard deviations are standardized scores that were computed within participant and per measurement method.