Education and Information Technologies (2023) 28:7413–7436
https://doi.org/10.1007/s10639-022-11478-7

Investigating the effect of multimodality and sentiments on speaking assessments: a facial emotional analysis

Joey Jia Qi Chong[1] · **Vahid Aryadoust**[1]

Received: 22 August 2022 / Accepted: 16 November 2022 / Published online: 1 December 2022

Abstract

This quasi-experimental study aimed to determine the relationship between (i) oral language ability and emotions represented by facial emotions, and (ii) modality of assessment (audios versus videos) and sentiments embedded in each modality. Sixty university students watched and/or listened to four selected audio-visual stimuli and orally answered follow-up comprehension questions. One stimulus was designed to evoke happiness while the other, sadness. Participants' facial emotions during the answering were measured using the *FaceReader* technology. In addition, four trained raters assessed the responses of the participants. An analysis of the *FaceReader* data showed that there were significant main and interaction effects of sentiment and modality on participants' facial emotional expression. Notably, there was a significant difference in the amount of facial emotions evoked by (i) the happy vs. sad sentiment videos and (ii) video vs. audio modalities. In contrast, sentiments embedded in the stimuli and modalities had no significant effect on the measured speaking performance of the participants. Nevertheless, we found a number of significant correlations between the participants' test scores and some of their facial emotions evoked by the stimuli. Implications of these findings for the assessment of oral communication are discussed.

Keywords Basic emotions · Facial emotional analysis · *FaceReader* · Integrated speaking · Listening · Multimodality · Sentiment analysis

✉ Vahid Aryadoust
Vahid.aryadoust@nie.edu.sg

Joey Jia Qi Chong
NIE18CHON529@e.ntu.edu.sg

[1] National Institute of Education, Nanyang Technological University, Singapore, Singapore

There is a wide array of factors that influence the speaking performance of individuals in assessments, such as task types (e.g., type of response format such as whether tasks are independent or integrated), types of stimuli used (e.g., sentiment and modality of stimuli) and emotions of participants (Butler et al., 2000; Ockey & Li, 2015). Among these, sentiments and modality of stimuli are under-researched in language assessment research, although there is an extensive body of the literature that investigates these factors in other contexts (Broadbent et al., 2017; Schreuder et al., 2016; Taffou et al., 2013; Takagi et al., 2015).

In this study, we investigated the effect of sentiment and modality of stimuli on speaking performance in an integrated speaking (listen-to-speak) assessment. Research shows that sentiments embedded in stimuli (input) during speaking or other language activities have an impact on how efficiently language users communicate (e.g., Barabadi & Khajavy 2020; Karami et al., 2019; Quintelier et al., 2019). On the other hand, the effect of stimuli's modality on speaking and language performance of students remains an open question. Some studies have indicated that visual stimuli can enhance the language performance of students under assessment conditions (e.g., Wagner, 2010). However, it has been suggested that video-mediated stimuli can have an adverse effect on language performance, likely because language users are distracted by the input they receive from several modalities (Pardo-Ballester, 2016). Following previous research (Arramreddy & Krishnan, 2016; Jung et al., 2014; Pardo-Ballester, 2016), we measured sentiments as emotions embedded in auditory-visual stimuli and examined their effects on participants' emotions and oral language performance. In addition, the effect of the modality of stimuli (auditory vs. visual) on oral skills performance was investigated. To achieve the goals of the study, we leverage advanced face-reading technologies and sentiment analysis along with psychometric methods in a quasi-experimental study.

1 Emotions and feelings

Results from neurocognitive research show that there is a close affinity between affect and cognition (Klinger, 1996). Affect is a general term that consists of emotions (unconscious physio-neurological responses to external and internal stimuli) and feelings (subjective and conscious perceptions of emotions) (Tran, 2007). Based on Tran's (2007) study, emotion is episodic and dynamic, of a relatively brief duration, event- or object-specific, and arises based on the cognitive and emotional content of the stimuli. Parkinson et al., (1996) found that both feelings (moods) and emotions can affect cognitive processes like memory and perception, and by extension other behaviours such as verbal interactions. Thus, emotions usually have specific implications for behaviours (Tran, 2007).

Previous studies have shown that emotions can profoundly affect students' academic engagement, learning and hence performance (Hascher, 2010; Linnenbrink-Garcia & Pekrun, 2011; Pekrun, 1992). The role and importance of emotions in academic settings has also been recognized in educational theories like control-value theory of achievement emotions (Pekrun, 2006). However, a study by Pekrun and

Stephens (2010) showed that apart from test anxiety research and attributional studies, these emotions are often under-appreciated in psychological research.

Previous research has employed various methods to measure feelings. Subjective measurements consist of self-reports questionnaires like the achievement emotions questionnaire (AEQ; Pekrun et al., 2011; Peixoto et al., 2015), positive and negative affect schedule (PANAS; Ketonen et al., 2019), and self-assessment manikin (SAM; Geethanjali et al., 2017). However, it is well-known that self-reports may not be accurate in measuring emotions, as people are not conscious about the physiological changes in their bodily chemistry and can overstate or understate their experience of their mood, thus consciously altering measured outcomes (Ciuk et al., 2015). Accordingly, many researchers have applied facial electromyography (EMG; Kulke et al., 2020), face-reading emotion recognition software such as FaceReader (e.g., Hirt et al., 2019), or Affectiva iMotions (e.g., Fasel & Luettin, 2003) to measure emotions expressed through facial emotions. These studies have attested to the utility of these techniques in educational and psychological research, although the techniques are not without limitation. For example, a study by Boxtel (2010) evaluating the strengths and limitations of facial EMG in measuring emotions discovered that it can be an obtrusive technique which requires the fitting of electrodes on face. In addition, the effectiveness of facial EMG can be influenced by nonaffective, behavioural factors such as mental fatigue, which may elicit affective responses in the face.

FaceReader, on the other hand, has been shown to be a relatively reliable automated system for the recognition of a number of specific properties in facial images amongst the major software tools for emotion classification currently available (e.g., Uyl & Kuilenburg 2005). Emotions can be induced through different modalities, such as imagination, film (audio narration with visuals), sound (audio narration), music, images (see Fakhrhosseini & Jeon 2017), and their effect can be partially captured through the measurement of facial emotional expressions. FaceReader utilizes this fundamental principle and, through the use of machine learning, determines emotions of people with a high degree of accuracy.

2 Sentiment analysis

Different emotions can be induced by sentiments embedded in the stimuli; for example, happy stimulus would induce happy emotions in participants, regardless of the modality of the stimuli (Fakhrhosseini & Jeon, 2017). Previous research has shown the effect of sentiment on different aspects of language and educational assessment performance or test scores (Barabadi & Khajavy, 2020; Karami et al., 2019; Quintelier et al., 2019).

The sentiment embedded in a stimulus can be determined through sentiment analysis of content which can identify emotions such as frustration, joy, anger, sadness, excitement, and so on (Mohammad, 2015). A stimulus targeted at inducing a particular emotion may also induce other emotions. In a study conducted by Hewig et al. (2005), sad stimuli evoked other negative emotions such as disgust and rage in addition to the primary sad emotions and an amusement stimulus caused higher intensity of positive emotions than any of the negative emotions. Some studies have

shown that emotions do affect a participant's performance, where better performance (i.e., higher test scores) is observed when the stimuli contained happy rather than sad sentiments (Arramreddy & Krishnan, 2016). In Jung et al.'s study (2014), participants who experienced positive emotions performed better than those experiencing negative emotions, and both groups of participants outperformed participants experiencing neutral emotions. In another study by Lochner (2015), however, the experimentally manipulated emotions did not affect test performance in an online reasoning test.

There is comparatively little research on how induced emotions of people affect performance in integrated speaking tests, as much research only discussed how emotions affect test performance in other areas such as in listening tests (Stientjes, 2012; Wagner, 2010), general academic tests (Gumora & Arsenio, 2002) and logical reasoning tests (Jung et al., 2014). As performance on speaking tests (indicated by scores) is affected by many other factors such as task type and the participant's ability (Tuan & Mai, 2015), it is difficult to generalise whether emotions do affect test performance and whether these effects are significant. In this study, we hypothesized that only a small share of variance in test scores is attributed to the emotions of participants. This assumption is based on the analysis of the first operationalisation of communicative competence (Canale & Swain, 1980) which included no mention of affects and emotions as a dimension that is related to spoken language. Even based on later formulations of communicative competence, it can be inferred that affective schema is not expected to have a significant impact on participants' performance in standardized or formal situations.

3 Modality of Stimuli and oral Language performance

Previous research has examined the effect of modality of stimuli on oral language performance, but there is a dearth of research on the effect of mode on integrated listening test performance. Wagner (2010) discovered that participants who were exposed to video-audio stimuli performed better than (achieved higher scores) when pure audio stimuli was used in an auditory comprehension test and this difference was statistically significant. This suggests that the non-verbal visual information of spoken texts helped participants better comprehend aural information and contributed to the video group's superior performance. Contrastingly, Suvorov (2008) found that participants scored significantly lower for video-mediated passages than for audio-only passages and photo-mediated passages in an auditory comprehension test upon comparing the mean scores.

Further, Pardo-Ballester (2016) reported that learners of different proficiencies performed differently (achieved different scores) when different modalities are used. Learners with lower proficiency levels performed better with only audio, while learners with higher proficiency levels performed better with video-audio stimuli. A plausible reason provided was that the presence of distracting visual elements in video-audio stimuli could have obstructed learning (Pardo-Ballester, 2016).

3.1 The Present Study

While many research studies compare the effects of visual-auditory or auditory stimulus on either emotions (Riviello & Esposito, 2016) or performance (Pardo-Ballester, 2016; Wagner, 2010), there is little empirical research investigating and comparing the use of both auditory and visual stimulus, and their impact on emotions and integrated speaking test performance. Extending the existing body of research, this study investigates whether there is a relationship between two groups of variables: (i) the participants' oral ability and emotions proxied by their facial emotions, and (ii) the modality of assessment (audios vs. videos) and sentiments embedded in each modality.

The research questions of this paper are:

1. Is there any significant effect of the sentiments and modalities of stimuli on the participants' integrated speaking abilities?
2. Is there any significant effect of the sentiments and modalities of stimuli on the participants' facial emotions?
3. What is the relationship between integrated speaking performance and emotions evoked by various types of stimuli?

4 Method

4.1 Participants

Sixty ($\text{male} = 33, \text{female} = 27$) bilingual undergraduate students from a public university in Asia, aged between 18 and 29 years old, participated in the study. English was the first language of the participants. Participants who were under 21 years of age were asked to obtain parental consent before their allocated test session. Each participant was assessed on their English oral proficiency in answering 12 follow-up questions (three questions per stimulus) after watching or listening to four 2-minutes long stimuli. As discussed later, the stimuli and questions were all in English.

5 Raters

Four ($\text{male} = 1, \text{female} = 3$) post-graduate university students from an Asian university between 29 and 38 years of age who were enrolled in a graduate course in applied linguistics participated in the study as raters. The raters had high proficiency levels in English for them to specialise in applied linguistic. Each rater assessed the performance of 18 assigned participants on 4 items for 4 stimuli (approximately 288 data points per rater) using the Internet-Based Test of English as a Foreign (TOEFL iBT) Integrated Speaking Rubrics.

6 Measurement Instruments

6.1 Integrated tests

Participants were exposed to four 2-minutes long stimuli on two main topics (Education-Animals and Education-Earth), where two were designed in video forms and two were in audio forms. They were required to answer follow-up comprehension questions orally and their responses were recorded for coding. Special care was taken to select only 2-minutes portion of the videos that are formal, informational, and educational as studies conducted by Wistia (Fisherman, 2016; Guo et al., 2014) found that the optimal video length for maximum engagement is 2 min.

Transcripts were then generated for selected videos and analysed using the Sentiment Analysis and Social Cognition Engine (SÉANCE) by Crossley et al. (2017) which provides data for further analysis. Accordingly, multiple indices that measure happiness were chosen, as they were the only relevant indices that directly measure happiness and sadness. They are:

(1) Happiness_GALC (Geneva affect label coder): Words with a positive valence which imply or indicate happiness. They were extracted and coded based on the GALC list (see Scherer, 2005). "Happiness_GALC_neg_3" is another index used which indicates vocabulary associated with the feeling of happiness. This index was computed based on the GALC list and includes words such as cheers and delight ("neg" stands for the negative filter).
(2) Joy_GALC: Positive emotion words describing joy based on the GALC list. Similarly, "Joy_EmoLex" (emotion lexicon) refers to positive emotion words describing joy such as tantalizing, loveable, etc. EmoLex is a list of English words annotated for basic emotions (anger, anticipation, surprise, fear, trust, sadness, joy, and disgust) and negative and positive sentiments (Mohammad & Turney, 2013, p. 451). Another related index used is "Joy_EmoLex_neg_3", which also refers to positive emotion words describing joy based on the EmoLex list ("neg" stands for the negative word filter.). Likewise, "Joy_GALC_neg_3" represents the degree of positive emotions describing joy based on the GALC list. Finally, "Joy_component" describes positive emotions describing joy, and is derived from the principal component analysis.
(3) Anticipation_EmoLex: Vocabulary extracted from the EmoLex wordlist, indicating anticipation such as tantalizing or unbeaten. Relatedly, "Anticipation_EmoLex_neg_3" is an anticipation index computed with the negative filter on. The other two indices related to anticipation are "Surprise_EmoLex" and "Surprise_EmoLex_neg_3".
(4) Four indices that measure sadness: "Sadness_GALC", "Sadness_EmoLex", "Sadness_GALC_neg_3", and "Sadness_EmoLex_neg_3".

These indices were used to choose eight videos which were downloaded, and audio versions were generated for each of them.

After each stimulus (video or audio), three follow-up questions would appear on the screen. For each question, the participants were given 30 seconds to construct

their answers and 60 seconds to orally-present their answers to the camera. Each participant answered a total of 12 questions during the test. The sequence of videos was counterbalanced.

6.2 FaceReader

We used *FaceReader* 8.0, a software developed by Noldus Information Technology, to detect and classify the facial emotions of participants into seven categories: happy, sad, angry, surprised, scared, disgusted and neutral (see Fig. 1). These emotional categories were described by Ekman (1970) as basic or universal emotions (Loijens & Krips, 2018), which is sufficient for this research that works with basic emotions.

In this research, video files were processed through *FaceReader* to obtain frame-by-frame analysis. *FaceReader* works by first using the popular Viola-Jones algorithm (Viola & Jones, 2001) to detect the face followed by using the Active Appearance Method (AAM; Cootes & Taylor 2000) to analyse over 500 key points in the face, as well as facial texture to create an accurate 3D model of the face. *FaceReader* then uses an artificial neural network built by trained experts (Bishop, 1995) to classify facial emotion. The Deep Face classification method used in addition to the AAM allows for a better analysis of a face even if part of is hidden (Loijens & Krips, 2018). Additionally, the accuracy of *FaceReader* in measuring emotions has been validated in a recent validation study by Zumhasch (2018) where *FaceReader* showed 100% precision in measuring happy emotions, 97% in measuring neutral emotions and 91% precision in measuring scared and sad emotions, thus providing additional validation for the sentiment analysis carried out in this study.

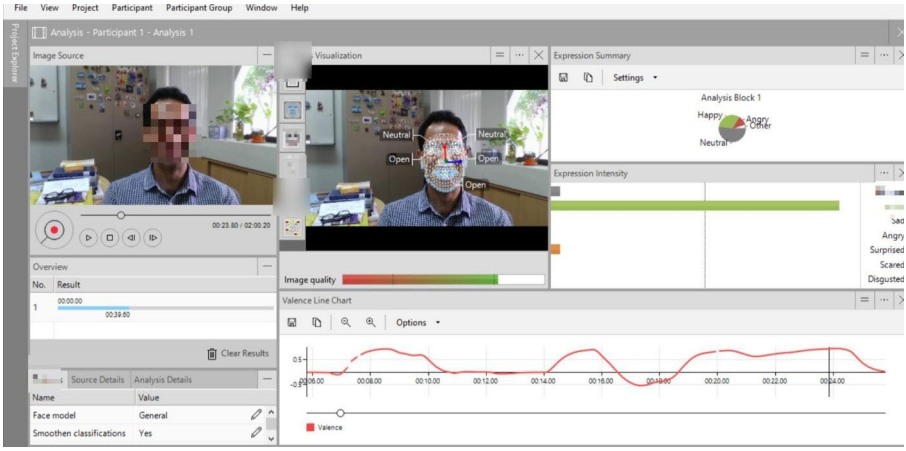

Fig. 1 FaceReader Interface. (*Note:FaceReader* can detect and classify the facial emotions of participants into seven categories: happy, sad, angry, surprised, scared, disgusted, and neutral.)

7 Assessment Rubrics

We used the TOEFL iBT Test Integrated Speaking rubrics to rate participants' performance. The Integrated Speaking test of the TOEFL iBT requires participants to utilise their reading and listening skills in addition to their speaking skills. As the designed test in this research requires participants to answer follow-up questions based on visual and/or auditory stimuli, the Integrated Speaking rubrics was used. There are 4 items for scoring in the Integrated Speaking rubrics (namely General Description, Delivery, Language Use and Topic Development), with scores ranging from 0 to 4. In this research, no score 0 was given to any participants as all participants made attempts to respond and all responses were related to the topic to various degrees.

8 Procedures for participants

Participants were seated comfortably in front of a computer monitor that was connected to researchers' laptop which projected relevant content onto the computer monitor. They were first given a 5-minutes briefing. Next, they were asked to complete the consent form and a background questionnaire. Special care was taken to ensure that the participants were in good illumination, which is important for *FaceReader* to yield reliable results (Loijens & Krips, 2018). The participants were given a pen and some paper to take down notes during the test. Next, the participants were exposed to the four 2-minutes long stimuli on two main topics (Education-Animals and Education-Earth) and answered 12 follow-up questions (3 questions per stimulus). At the end of the session, participants were given $10 in cash. The sessions lasted 25 to 50 min.

9 Procedures for raters

The four raters were trained through a synchronous video chat with the researchers via WebEx due to the CoVID-19 pandemic. Prior to the training session, the TOEFEL Integrated Speaking rubrics, a marking guide and samples from level 1, 2, 3, and 4 were prepared and uploaded onto a folder on Google Drive which is shared with the four raters. The transcripts of the 4 stimuli, their YouTube links and relevant duration to watch, as well as suggested answers for each of the 12 questions were further included in the marking guide.

During the training session, the 4 components of the TOEFEL Integrated Speaking rubrics (namely General Description, Delivery, Language Use and Topic Development) were explained and important points were highlighted. Ratings were given out for each stimulus (4 ratings per stimulus per participant) instead of each question (12 ratings per stimulus per participant) to normalize any variance in performance for the three questions of the same stimulus. Four samples from level 1, 2, 3 and 4 were presented to the raters where researchers explained the reasons for the gradings. It was followed by hands-on rating practices where all raters were asked to view some video snippets of participants uploaded on Google Drive and subsequently rate their

performances based on the TOEFEL Integrated Speaking rubrics. Discrepancies in rating were addressed by directing the raters' attention back to the rubrics as well as the four samples from level 1, 2, 3, and 4.

After the training session, each rater was assigned 18 participants to assess where 4 sets of ratings were given per participant, amounting to a total of 72 sets of ratings per rater. As each set of rating includes 4 items, there was $4\mathrm{X}72 = 288$ data points per rater, thus yielding sufficient information for measuring their ability (Linacre, 2008). A cascading design was employed to assign the participants to raters where each rater had to assess 9 participants that were common to the rest of the raters in addition to the 9 participants that were only assigned to them. One of the researchers also rated a total of 24 participants (9 common participants, 9 exclusive participants and 6 participants) which were used as samples or for practice during the training session.

9.1 Data Analysis

Before answering the research questions of the study, we examined the reliability and psychometric validity of the data. This consisted of examining raters' performance, functionality of the stimuli or assessment instruments, and participants' scores. We chose the multi-faceted Rasch measurement (MFRM; Linacre 1994) for this purpose, as it provides very useful evidence concerning the functionality of the instrument and the performance of raters and participants.

After establishing the reliability and psychometric validity of the data, we answered the research questions by using repeated-measures MANOVA and bivariate correlation analysis. We discuss the analytical procedures next.

10 Reliability and psychometric Validity: multi-faceted Rasch Measurement (MFRM)

We used the Facets computer package (version 3.83.2) (Linacre, 2008) to investigate the psychometric validity of the measurements and the reliability of rater performance. The MFRM model applied is, as follows:

$$\log \frac{p_{nijk}}{p_{nijk-1}} = B_n - D_i - C_j - E_h$$

where P_{nijk} is the probability that participant n will be awarded a rating of k on item i by rater j ; B_n is the ability of participant *n*; D_i is the difficulty level of item i ; C_j is the severity of rater *j*; and E_h is the difficulty of the threshold from $k-1$ to k on the scale unique to item i . We computed (i) infit mean square (MnSq) which is an index sensitive to disturbances near the ability of the participant, such as when a mid-ability student receives extremely high or low score by one or more raters, and (ii) outfit MnSq which is outlier sensitive where it highlights distortions in the data distant from the actual speaking ability of the participant, such as when a low-ability student is given an unexpectedly high score or vice versa (Linacre, 1994). MnSq values of the participants, raters, and stimuli (both audio and visual ones) fell between

0.5 and 1.5 indicating tolerable randomness in the data for these facets. This further lends evidence to the psychometric validity of participants' scores, raters' performance (i.e., their scoring practice), and the stimuli.

In addition, the variance explained by Rasch measures in this analysis was 71.04%, which is a significantly large share of variance. This provides additional supporting evidence for the psychometric validity of the measurements conducted by the raters in the study (Linacre, 1994). The reliability of student scores was 0.79, with a strata coefficient equal to 2.89, indicating high reliability and the detection of roughly 3 (2.89) levels of ability or test performance among the students.

Finally, the last piece of evidence backing up the reliability and psychometric validity of the measurement stimuli and raters' performance was derived from the psychometric functionality of the scoring categories. The TOEFL rubrics adopted in the study comprised four scoring categories (1 to 4). This yields three "thresholds", that is, one threshold between every two adjacent categories, i.e., a threshold between 1 and 2, a threshold between 2 and 3, and a threshold between 3 and 4. In MFRM, a threshold is a point on the ability continuum where the participant starts to have a higher probability of achieving a higher score. Thresholds should be adequately distant from each other, meaning that they should ascend monotonically (Linacre, 1994). We estimated the Rasch-Andrich thresholds in this study, which were −3.38, -1.05, and 4.42, suggesting a monotonic increment in the difficulty level of the thresholds.

In sum, we found the ratings provided by the raters, the instrument, and the stimuli to be psychometrically valid and reliable. Thus, we utilized the scores in follow-up analyses to answer the research questions of the study.

Research Question 1: Is there any significant effect of the sentiments and modalities of stimuli on the participants' integrated speaking abilities?

To address the first research question, we first performed a descriptive statistical test to investigate the normality of the data by examining skewness and kurtosis values. We applied a within-subject design consisting of sentiment in stimuli (sad and happy) and modalities (video vs. audio) as the independent variables, and measures of integrated speaking performance (general description, delivery, language use, & topic development) as the dependent variables. Next, we performed a 2×2 repeated-measures MANOVA on the measured integrated speaking performance (dependent variable), which was the average ratings for each of the 4 individual scoring categories provided by the raters.

Research Question 2: Is there any significant effect of the sentiments and modalities of stimuli on the participants' facial emotions?

To address the second question, we performed a 2×2 repeated-measures MANOVA with participants' facial emotions (neutral, happy, sad, angry, surprised, scared & disgusted measured by *FaceReader*) as the dependent variable and sentiment and modality as independent variables.

Research Question 3: What is the relationship between integrated speaking performance and emotions evoked by various types of stimuli?

To address this question, the data from raters measuring participants' speaking ability and data from facial emotions measured by *FaceReader* were arranged based on the modalities and sentiments of stimuli. The combined data was subjected to bivariate correlation analysis to investigate whether there is a relationship between

the participants' oral ability and emotions proxied by their facial emotions. The effect sizes of data that were significant at either 0.01 or 0.05 alpha levels were calculated by squaring the correlation coefficients.

11 Results

11.1 Descriptive statistics

Table 1 presents the mean, standard deviation, skewness and kurtosis values of participants' measured general performance for the 4 combinations of sentiment and modality of stimuli. As demonstrated, the skewness and kurtosis values fell between the range of -1.26 and 1.98, thus providing evidence for normality. The largest and smallest mean value occurred when the stimulus was in video form and embedded sad emotions ($\text{mean} = 3.32$) and when the stimulus was in audio form and embedded sad emotions ($\text{mean} = 3.19$), respectively. In addition, the mean value of participants' measured general performance was higher when the stimuli embedded happy sentiments ($\text{mean} = 3.26$) rather than sad sentiments ($\text{mean} = 3.25$). The mean value of participants' general performance was higher when the stimulus was in the video modality ($\text{mean} = 3.28$) rather than audio forms ($\text{mean} = 3.24$).

Research Question 1: Is there any significant effect of the sentiments and modalities of stimuli on the participants' integrated speaking abilities?

The results of the repeated-measures MANOVA test conducted for the 4 scoring categories (general description, delivery, language use, & topic development) showed no significant main and interactions effects of sentiments of stimuli (happy vs. sad) and modalities of stimuli (video vs. audio) on participants' measured performance for each of the 4 individual scoring categories ($p > .05$).

Research Question 2: Is there any significant effect of the sentiments and modalities of stimuli on the participants' facial emotions?

The descriptive statistics of the seven emotions of participants proxied by facial emotion showed the data were normally distributed (skewness and kurtosis between −2 and +2). The Pillai's Trace of the repeated-measures MANOVA test showed that the sentiment of stimuli had a significant main effect on the amount of facial emotional expression ($F(7,33) = 59.35$, $p < .001$), with a partial eta squared = 0.926, which indicates a very large effect. Similarly, there was a significant main effect for the modality of stimuli, $F(7, 33) = 72.53$, $p < .001$, partial eta squared = 0.94) as well as a significant interaction effect, $F(7, 33) = 43.91$, $p < .001$, partial eta squared = 0.90).

Table 1 Descriptive Statistics of General Performance in the Integrated Listening Test

Sentiment, mode of stimuli	M	SD	Skewness	Kurtosis
Happy, Video	3.24	0.66	-0.93	0.81
Happy, Audio	3.29	0.68	-1.26	1.98
Sad, Video	3.32	0.64	-0.82	-0.10
Sad, Audio	3.19	0.67	-0.79	0.62

Note: M = Mean; SD = standard deviation. Range of the scale is from 0 to 4.

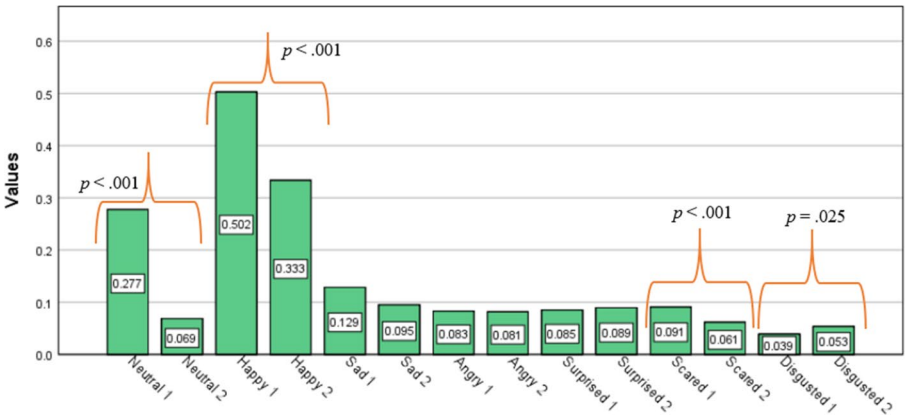

Fig. 2 Pairwise comparisons of participant emotions based on the sentiment of the stimuli (sad vs. happy). (*Note:* 1 = stimulus with happy sentiment; 2 = stimulus with sad sentiment.)

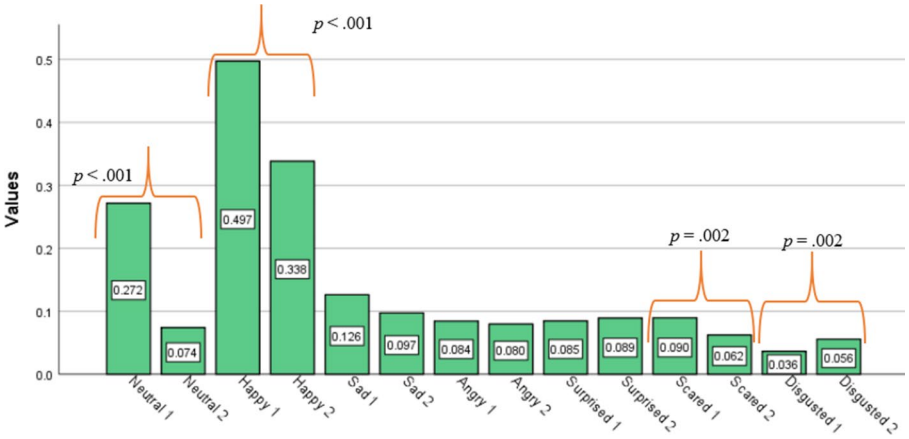

Fig. 3 Pairwise comparisons of participant emotions based on the mode of the stimuli (video vs. audio). (*Note:* 1 = video mode; 2 = audio mode.)

Using a Bonferroni correction, we conducted multiple comparisons on the estimated marginal means. As Fig. 2 presents, there was a significant difference in the amount of facial emotion evoked by the happy vs. sad sentiment videos in four pairwise comparisons: neutral ($p < .001$), happy ($p < .001$), scared ($p < .001$), and disgusted ($p = .025$).

We also performed multiple comparisons on the estimated marginal means of the modality of stimuli and found significance differences in the amount of facial emotion evoked by video vs. audio modalities (Fig. 3). These consisted of neutral ($p < .001$),

Table 2 Sentiment*Mode Interaction Effects on Facial Emotions

Participants' facial emotions measured by FaceReader	Stimuli's sentiment	Mode 1	Mode 2	Mean Difference	*p* value
Neutral	Happy	Video	Audio	0.411*	0.00
	Sad	Video	Audio	-0.016*	0.016
Happy	Happy	Video	Audio	-0.042*	0.003
	Sad	Video	Audio	0.360*	0.00
Sad	Happy	Video	Audio	0.030*	0.01
	Sad	Video	Audio	0.028	0.347
Angry	Happy	Video	Audio	0.008	0.133
	Sad	Video	Audio	0.002	0.924
Surprised	Happy	Video	Audio	-0.004	0.597
	Sad	Video	Audio	-0.005	0.81
Scared	Happy	Video	Audio	-0.003	0.431
	Sad	Video	Audio	0.058*	0.001
Disgusted	Happy	Video	Audio	-0.005*	0.039
	Sad	Video	Audio	-0.033*	0.008

happy ($p < .001$), scared ($p = .002$), and disgusted ($p = .002$). Finally, the interaction analysis revealed 8 significant interactions out of 14 comparisons (Table 2).

Research Question 3: What is the relationship between integrated speaking performance and emotions evoked by various types of stimuli?

To answer research question 3, emotions of participants proxied by facial emotions were correlated with their scores to explore the relationship between the two variables. Table 3 presents statistically significant bivariate correlations (rounded up to two decimal values) and their respective effect sizes in brackets. In this table, the components of participants' integrated speaking scores are presented in the left-hand column. For example, the "Happy Video_General description" represents the participants' general score in the happy video, while the "Sad Audio_Language use" represents their language use score in the sad audio. In addition, the participants' facial emotions captured by *FaceReader* are presented in the top row. For example, the "Happy Video_Neutral Emotions" indicates the amount of neutral facial emotions of the participants as they were watching the happy video, whereas the "Happy Video_Sad Emotion" represents their sad facial emotions (if any) while watching the happy video. Non-significant correlations are not presented.

Overall, the magnitude of effect sizes is low to medium, as indicated by the parenthesized coefficients of determination ranging from 0.08 to 0.34. For example, there was a medium correlation of 0.50 (effect size = 0.25) between the delivery scores in the video with embedded happy emotions and the "Happy Audio_Happy Emotions". The findings indicate that a low-to-medium share of the variance in the participants' test scores is associated with their facial emotions.

Table 3 Correlation Analysis of Average Ratings and Emotions

	Happy Video_ Neutral Emotions	Happy Video_ Happy Emotions	Happy Video_Sad Emotions	Happy Video_Surprised Emotions	Happy Audio_ Neutral Emotions	Happy Audio_ Happy Emotions	Happy Audio_ Scared Emotions	Sad Video_ Happy Emotions	Sad Video_ Scared Emotions	Sad Audio_ Neutral Emotions	Sad Audio_ Happy Emotions	Sad Audio_ Surprised Emotions	Sad Audio_ Scared Emotions
Happy Video_ General description						−0.36* (0.13)		−0.38* (0.14)			−0.38** (0.14)		
Happy Video_ Delivery	0.29* (0.08)	−0.42** (0.17)			0.33* (0.11)	−0.50** (0.25)		−0.52** (0.27)		0.28* (0.08)	−0.53** (0.28)		
Happy Video_ Topic development		−0.30* (0.09)				−0.44** (0.19)		−0.37* (0.14)			−0.33* (0.11)		
Happy Video_ General description	0.33* (0.11)	−0.41** (0.17)				−0.38** (0.15)	0.32* (0.10)	−0.39** (0.15)		0.34* (0.11)	−0.48** (0.23)		
Happy Audio_ Delivery		−0.40* (0.16)			0.33* (0.11)	−0.48** (0.23)	0.30* (0.09)	−0.45** (0.20)		0.35* (0.12)	−0.56** (0.32)		

Table 3 (continued)

	Happy Video_ Neutral Emotions	Happy Video_ Happy Emotions	Happy Video_Sad Emotions	Happy Video_Surprised Emotions	Happy Audio_ Neutral Emotions	Happy Audio_ Happy Emotions	Happy Audio_ Scared Emotions	Sad Video_ Happy Emotions	Sad Video_ Scared Emotions	Sad Audio_ Neutral Emotions	Sad Audio_ Happy Emotions	Sad Audio_ Surprised Emotions	Sad Audio_ Scared Emotions
Happy Audio_ Language use		−0.42** (0.17)				−0.41** (0.17)	0.28* (0.08)	−0.45** (0.20)		0.34* (0.11)	−0.54** (0.29)		
Happy Audio_ Topic development	0.33* (0.11)	−0.43** (0.19)				−0.40** (0.16)	0.39** (0.15)	−0.44** (0.19)	0.27* (0.08)		−0.45** (0.21)		0.38* (0.15)
Sad Video_ General description							0.29* (0.08)						
Sad Video_ Delivery		−0.50** (0.25)	0.33* (0.11)			−0.48** (0.23)	0.36** (0.13)	−0.49** (0.24)			−0.58** (0.34)		
Sad Video_ Language use		−0.41** (0.16)				−0.30* (0.09)		−0.33* (0.11)			−0.38* (0.15)		

Table 3 (continued)

	Happy Video_ Neutral Emotions	Happy Video_ Happy Emotions	Happy Video_Sad Emotions	Happy Video_Surprised Emotions	Happy Audio_ Neutral Emotions	Happy Audio_ Happy Emotions	Happy Audio_ Scared Emotions	Sad Video_ Happy Emotions	Sad Video_ Scared Emotions	Sad Audio_ Neutral Emotions	Sad Audio_ Happy Emotions	Sad Audio_ Surprised Emotions	Sad Audio_ Scared Emotions
Sad Video_ Topic development		−0.34* (0.11)				−0.28* (0.08)		−0.32* (0.10)			−0.36* (0.13)		
Sad Audio_ General description				0.30* (0.09)		−0.33* (0.11)		−0.326* (0.11)			−0.34* (0.12)	0.30* (0.09)	
Sad Audio_ Delivery		−0.33* (0.11)				−0.50** (0.25)		−0.48** (0.23)			−0.43** (0.18)		
Sad Audio_ Language use		−0.34* (0.12)									−0.39* (0.09)		

Note. $^{*}p < .05.$ $^{**}p < .001.$

The correlation coefficients have been rounded up to two decimal values

12 Discussion

This study set out to investigate the relationship between (i) the participants' oral ability and emotions proxied by their facial emotions, and (ii) the modality of assessment (audios vs. videos) and sentiments embedded in each modality. The three research questions are discussed next.

Research Question 1: Is there any significant effect of the sentiments and modalities of stimuli on the participants' integrated speaking abilities?

Using repeated-measures MANOVA, we found that modality (video vs. audio) and sentiments of stimuli (happy vs. sad) had no significant effects on participants' measured performance. Regarding the effects of modality, our observations exhibited a similar (though non-significant) trend with Wagner's (2010) study wherein participants who were exposed to video-audio stimuli did not perform differently from when pure audio stimuli were used in a listening test. This suggests that the non-verbal visual information of spoken texts did not seem to help participants to better respond to the questions concerning the audio-visual modality. While this lends support to the validity of the test (as there was no test method effect), there are several observations that would be worth further investigation—with a larger sample size, the non-significant differences may tend to move towards conventionally accepted statistical significance (Faber & Fonseca, 2014; Mayo & Spanos, 2011). Based on Wagner's (2010) argument, we speculate that some factors such as the nature of stimuli used (dialogue vs. lecturette texts) could have reduced the effectiveness of using video stimuli. An important advantage of videos in the assessment of spoken proficiency is their authenticity or their similarity with real-life situations (Douglas, 2000). In real academic environments, lectures constitute an important component of every subject. Accordingly, the content of spoken proficiency tests, which aim to predict participants' spoken performance in these environments, should exhibit significant similarities with the content of the language use domain the tests aim to emulate.

Regarding the effects of sentiments, the measured performances of participants were better when the stimuli embedded happy rather than sad sentiments, which shows a similar trend to Arramreddy & Krishnan (2016). This could be explained by the findings of Fredrickson (2001) which suggested that positive emotions led to better performance because they encourage exploring and integrating diverse materials, as well as better problem-solving skills. Nevertheless, test developers should note that that, as Valiente et al. (2012) showed, the intensity of positive emotions determines their impact on test performance or scores where high-arousal positive emotions (such as exuberance, excitedness, and elatedness) may instead lead to worse performance. This may not be applicable in this study as it involved happiness (rather than excitedness) which is a low-arousal positive emotion, but it is a point worth investigation in future research.

Research Question 2: Is there any significant effect of the sentiments and modalities of stimuli on the participants' facial emotions?

There were significant main and interaction effects of sentiment and modality of stimuli on the amount of facial emotional expression. In particular, multiple comparisons on the estimated marginal means showed that there was a significant difference in the amount of facial emotion evoked by the happy vs. sad sentiment videos in

four pairwise comparisons: neutral, happy, scared, and disgusted. As expected, the pairwise comparison of the mean values of neutral, happy, scared, and disgusted emotions for video stimuli evoking sad and happy emotions showed that the happy stimuli evoked higher intensities of happy emotions than the sad stimuli, while the sad stimuli evoked higher intensities of disgust emotions. This perhaps provides criterion-based validity evidence for the sentiment analysis that the four chosen stimuli evoked their targeted emotions where the happy stimuli evoked more happy emotions than sad stimuli, while the sad stimuli evoked higher intensity of some negative emotions (e.g., disgust) than the happy stimuli.

Interestingly, the sad stimuli evoked lower intensity of some negative emotions like scared emotions compared to the happy stimuli. This has a theoretical and an empirical implication. Our findings provide evidence that sentiments embedded in auditory texts would have measurable effects on speakers' sentiments expressed as facial emotions. In other words, measurable stimulus-response effects of emotions can be revealed via text-mining and biometric technologies. Based on the study by Ekman (1992) who investigated basic emotions, there is consistence evidence showing that facial emotions provide a good way to distinguish the different emotions, thus supporting the use of biometric technologies such as facial emotional analysis to measure stimulus-response effects of emotions. Just like the study by Hewig et al. (2005) where a sad stimulus evoked other negative emotions such as disgust and rage in addition to the primary sad emotions, our study showed a similar trend where the stimuli targeted at inducing sadness also induced higher intensity of other negative emotions such as disgust emotions as compared to the stimuli that embedded happy sentiments. This is consistent with another study by Schwartz and Weinberger (1980) that investigated the relations between happiness, sadness, anger, fear, depression, and anxiety where they discovered that emotions have similarities and typically interrelate during various affective situations with fear being a particular type of anxiety.

We further found that there was a significance differences in the amount of facial emotions evoked by video vs. audio modalities. These consisted of neutral, happy, scared, and disgusted emotions. It was found that the video stimuli elicited higher intensities of a larger number of emotions (neutral, happy, and scared) compared to audio stimuli which only evoked higher intensities of disgusted emotions. This follows the findings of the meta-analysis of the elicitation techniques by Westermann et al. (1996) which found that videos are the most effective stimuli at eliciting emotions, whether positive or negative, among participants. The study by Murugappan et al., (2009) also confirmed that audio-visual stimulus performs superior in evoking emotions than visual stimulus. An explanation for videos being most effective could be due to additional non-verbal cues in them besides audio, which may help in inducing higher intensity of emotions (Yazdani et al., 2013). The deviation in this study where participants experienced higher intensity of disgust for audio than video stimuli could perhaps be due to the presence of visuals in videos which did not help in eliciting disgust and maybe elicited other emotions in addition to disgust. Further research could perhaps consider analysing tone in addition to facial emotions to gain an even more accurate measurement of emotions.

Research Question 3: What is the relationship between integrated speaking performance and emotions evoked by various types of stimuli?

The amount of variance in test scores attributed to the emotions of participants was low to medium, with an effect size ranging from 0.08 to 0.34. This aligns with our hypothesis that only some part of the observed variance in test scores can or should be associated with the emotions of participants when they are using academic language, while most of the variation should arise from the differences in participants' speaking abilities. Thus, it might be said that similar to the study by Lochner (2015), the experimentally manipulated emotions did not have much significant impact on participants' speaking performance. Lochner (2015) suggested several explanations to account for the deviation from the expected hypothesis such as reasoning tests being less susceptible to the influence of emotions than other types of tests. One reason could be the difficulty to detect the effect of affective state on test performance due to the comparatively unstandardized situations under non-laboratory conditions. Another reason might be the diversity of the sample or the possibility of participants entering a state of flow wherein thoughts or feelings do not affect their ability to perform the task. As the test situation was rather standardized (controlled) in our study, one possible explanation for the small correlation between participants' measured performance and emotions could be the possibility of participants entering a state of flow where feelings do not affect their ability to perform the tasks. Nonetheless, this remains a speculation and further research is needed to examine its plausibility.

Messick's (1996) publication raised construct under-representation and construct-irrelevant variance as causes of invalidly high or low scores but as this study made conscious efforts to consider all the important aspects of what we intended to measure, construct under-representation does not seem to be the cause of low-to-medium correlations between emotions and test scores. A small amount of construct-irrelevant variance, on the other hand, seems to be a plausible explanation for the correlations observed. We suggest that future research on assessing speaking should take into consideration the effect of variables such as affect and modality of presentation of the stimuli on participants performance.

It should be noted that research questions 1 and 3 serve different purposes in this study. While research question 1 examined the overall differences between test scores across sentiments and modalities, research question 3 examined the fine-grained association between test scores and the different types of facial emotions that were captured by *FaceReader* in different modalities. In addition, only the correlation between the facial emotions and test scores on unique stimuli should be considered pertinent to this research question.

12.1 Implications for Assessment and Practice

An implication of this study is that the effect of test takers' affect (emotions and feelings) on the validity of oral proficiency tests is inconclusive. Particularly, as indicated in the MFRM analysis, we found that the test scores were psychometrically valid and reliable. However, there was evidence from research question 3 that some share of variance in some of the scores was associated with the participants' facial emotions. The corollary of these two seemingly contrastive findings is that, as discussed

by Low and Aryadoust (2021), while psychometric analysis is useful and necessary in assessment development, it is not sufficient. Thus, we suggest that assessment designers should exercise caution in developing multimodal assessments with emotionally bound content. Specifically, one should bear in mind that test takers' affective and cognitive *process* cannot be merely represented by test scores. Test scores only represent the end product of the assessment and are completely "imperceptive" to the processes by which test takers answered the test items (Aryadoust, 2023). It is possible to access (some of) the affective and cognitive processes of test takers through adopting modern technologies such as facial emotional analysis in study design. The integration of psychometric analysis (e.g., MFRM) and technology would present a more reliable profile of the validity of assessment instruments.

13 Limitations of the study

This study is not without its limitations. There are three limitations of the study that should be addressed in future research. First, based on Wagner's (2007) study, we suggest that the auditory stimuli should be diversified beyond the lecture style used in the study. As Wagner (2007) suggested, using lecturette texts could lead to participants having less interest and engagement with the stimuli than if dialogue texts were to be used, hence perhaps leading to reduced effectiveness of video stimuli in conveying additional contextual and non-verbal cues.

Second, unlike the study by Jung et al. (2014) wherein participants experiencing positive and negative emotions both outperformed participants experiencing neutral emotions, in our study there were no stimuli exclusively embedding neutral emotions which could serve as a point of reference to determine if positive and negative emotions will lead to better measured performance than neutral emotions. Having a point of reference allows the determination of how emotions affect performance from the norm, regardless of whether it is positive or negative emotions.

Finally, it is important to note that different individuals express emotions differently. Some individuals may show little emotions through facial expressions but their emotions may show clearly in the tone of their voice or in autonomic nervous system reactions, which may be measured using facial electromyography (Barrett, 2006), galvanic skin response, and/or pupil dilation analysis in eye-tracking. Analysing emotions using facial expressions could thus be improved by using biometric measures in speaking assessment research.

14 Conclusion

We investigated integrated speaking performance and found that there were no significant main and interaction effects of modality of stimuli (video vs. audio) and the sentiments of stimuli (happy vs. sad) on participants' measured performance. Another finding was that the sentiment (happy vs. sad sentiment videos) and modality (video vs. audio) of stimuli had a statistically significant effect on the amount of facial emotional expression. It was also found that the video stimuli elicited a larger

number of emotions (neutral, happy, and scared) with a higher intensity compared to the audio stimuli which only evoked higher intensities of disgusted emotions. This resonates with the finding of several studies that show video stimuli are more effective than audio stimuli in eliciting emotions. In addition, there were small to medium correlations among some of the measured facial emotions evoked by the audio and video stimuli and test scores of the participants. Hopefully, these findings will inform and inspire future research on the nature and functionality of multimodal tests of speaking.

Acknowledgements We would like to acknowledge the funding support from Nanyang Technological University – URECA (Undergraduate Research Experience on Campus) for this research project. The IRB number for the study is IRB-2019-10-029.

Data availability The datasets generated during and/or analyzed during the current study are not publicly available but can be made available upon request.

Declarations

Conflict of Interest None.

References

Andralyn Rui Lin, Low Vahid, Aryadoust Investigating Test-Taking Strategies in Listening Assessment: A Comparative Study of Eye-Tracking and Self-Report Questionnaires. *International Journal of Listening*, 1-20 1 https://doi.org/10.1080/10904018.2021.1883433.

Arramreddy, V., & Krishnan, S. (2016). *Effects of emotions on participants*. (California State Science Fair 2016 Project Summary for Project No. J0402). https://csef.usc.edu/History/2016/Projects/J04.pdf

Aryadoust, V. (2023). The vexing problem of validity and the future of second language assessment. *Language Testing*.

Barabadi, E., & Khajavy, G. H. (2020). Perfectionism and foreign language achievement: the mediating role of emotions and achievement goals. *Studies in Educational Evaluation*, *65*, https://doi.org/10.1016/j.stueduc.2020.100874.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28–58. https://doi.org/10.1111%2Fj.1745-6916.2006.00003.x.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. http://home.elka.pw.edu.pl/~ptrojane/books/Bishop%20-%20Neural%20Networks%20for%20Pattern%20Recognition.pdf

Boxtel, A. V. (2010). Facial EMG as a tool for inferring affective states. In A. J. Spink, F. Grieco, O. Krips, L. Loijens, L. Noldus, & P. Zimmerman (Eds.), *Proceedings of measuring behavior 2010* (pp. 104–108). Noldus Information Technology.

Broadbent, H. J., White, H., Mareschal, D., & Kirkham, N. Z. (2017). Incidental learning in a multisensory environment across childhood. *Developmental Science*, *21*(2), https://doi.org/10.1111/desc.12554.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 Speaking framework: A working paper* TOEFL Monograph Series, MS-26. https://www.ets.org/Media/Research/pdf/RM-00-03-Jamieson.pdf

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1.

Ciuk, D., Troy, A. S., & Jones, M. C. (2015). *Measuring emotion: Self-reports vs. physiological indicators*. Paper presented at the 2015 Annual Meeting of the Midwest Political Science Association. https://doi.org/10.2139/ssrn.2595359

Cootes, T., & Taylor, C. J. (2000). *Statistical models of appearance for computer vision* (Technical report). University of Manchester, Wolfson Image Analysis Unit, Imaging Science and Biomedical Engineering.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment analysis and social cognition engine (SEANCE): an automatic tool for sentiment, social cognition, and social order analysis. *Behavior Research Methods*, *49*(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

Ekman, P. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest*, *8*(4), 151–158. https://psycnet.apa.org/record/1972-06605-001.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068.

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. https://dx.doi.org/10.1590%2F2176-9451.19.4.027-029.ebo.

Fakhrhosseini, S. M., & Jeon, M. (2017). Affect/emotion induction methods. *Emotions and Affect in Human Factors and Human-Computer Interaction*, 235–253. https://doi.org/10.1016/B978-0-12-801851-4.00010-0.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. *Pattern Recognition*, *36*(1), 259–275. https://doi.org/10.1016/S0031-3203(02)00052-3.

Fisherman, E. (2016). *How long should your next video be?* [Blog post]. https://wistia.com/learn/marketing/optimal-video-length

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *The American Psychologist*, *56*(3), 218–226. https://doi.org/10.1037//0003-066x.56.3.218.

Geethanjali, B., Adalarasu, K., Hemapraba, A., Kumar, S. P., & Rajasekeran, R. (2017). Emotion analysis using SAM (Self-Assessment Manikin) scale. *Biomedical Research 2017, S18-S24*. https://www.semanticscholar.org/paper/Emotion-analysis-using-SAM-(Self-Assessment-scale-Geethanjali-Adalarasu/f5e12ce391f7598136cb8aae6df638c1a813be27

Gumora, G., & Arsenio, W. F. (2002). Emotionality, emotion regulation, and School Performance in Middle School Children. *Journal of School Psychology*, *40*(5), 395–413. https://doi.org/10.1016/S0022-4405(02)00108-5.

Guo, P. J., Kim, J., & Rubin, R. (2014). *How video production affects student engagement: an empirical study of MOOC videos* Paper presented at the first ACM conference on Learning @ scale conference. https://doi.org/10.1145/2556325.2566239

Hascher, T. (2010). Learning and emotion: perspectives for theory and research. *European Educational Research Journal*, *9*(1), 13–28. https://doi.org/10.2304/eerj.2010.9.1.13.

Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., & Bartussek, D. (2005). A revised film set for the induction of basic emotions. *Cognition and Emotion*, *19*(7), 1095–1109. https://doi.org/10.1080/02699930541000084.

Hirt, F., Werlen, E., Moser, I., & Bergamin, P. (2019). Measuring emotions during learning: lack of coherence between automated facial emotion recognition and emotional experience. *Open Computer Science*, *9*(1), 308–317. https://doi.org/10.1515/comp-2019-0020.

Jung, N., Wranke, C., Hamburger, K., & Knauff, M. (2014). How emotions affect logical reasoning: evidence from experiments with mood-manipulated participants, spider phobics, and people with exam anxiety. *Frontiers in Psychology*, *5*, 570. https://doi.org/10.3389/fpsyg.2014.00570.

Karami, M., Pishghadam, R., & Baghaei, P. (2019). A probe into EFL learners' emotioncy as a source of test bias: insights from differential item functioning analysis. *Studies in Educational Evaluation*, *60*, 170–178. https://doi.org/10.1016/j.stueduc.2019.01.003.

Ketonen, E. E., Malmberg, L. E., Salmela-Aro, K., Muukkonen, H., Tuominen, H., & Lonka, K. (2019). The role of study engagement in university students' daily experiences: a multilevel test of moderation. *Learning and Individual Differences*, *69*, 196–205. https://doi.org/10.1016/j.lindif.2018.11.001.

Klinger, E. (1996). Emotional influences on cognitive processing, with implications for theories of both. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking cognition and motivation to behavior* (pp. 168–189). The Guilford Press. https://psycnet.apa.org/record/1996-98326-008

Kulke, L., Feyerabend, D., & Schacht, A. (2020). A comparison of the Affectiva iMotions facial expression analysis software with EMG for identifying facial expressions of emotion. *Frontiers in Psychology*, *11*, 329. https://doi.org/10.3389/fpsyg.2020.00329.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. https://www.winsteps.com/a/Linacre-MFRM-book.pdf

Linacre, J. M. (2008). *A user guide to facets, rasch-model computer programs*. MESA Press.

Linnenbrink-Garcia, L., & Pekrun, R. (2011). Students' emotions and academic engagement: introduction to the special issue. *Contemporary Educational Psychology*, *36*(1), 1–3. https://doi.org/10.1016/j.cedpsych.2010.11.004.

Lochner, K. (2015). *Emotions and cognitive performance - The impact of participants' emotions on performance in online assessment*. [Doctoral dissertation, Freie Universität Berlin]. https://www.researchgate.net/publication/286174292_Emotions_and_Cognitive_Performance_-_The_Impact_of_Test_Takers'_Emotions_on_Performance_in_Online_Assessment

Loijens, L., & Krips, O. (2018). *FaceReader methodology note*.Behavioral research consultants at Noldus Information Technology. https://www.noldus.com/blog/facial-expressions-emotions-children

Mayo, D. G., & Spanos, A. (2011). Error statistics. *Philosophy of Statistics*, *7*, 153–198. https://doi.org/10.1016/B978-0-444-51862-0.50005-8.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302.

Mohammad, S. M. (2015). Sentiment analysis: Detecting valence, emotions, and other affectual states from text.Emotion Measurement.https://www.saifmohammad.com/WebDocs/emotion-survey.pdf

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x.

Murugappan, M., Juhari, M. R. B. M., Nagarajan, R., & Yaacob, S. (2009). An investigation on visual and audio-visual stimulus based emotion recognition using EEG.International Journal of Medical Engineering and Informatics, *1*(3). https://www.researchgate.net/publication/238158179_An_investigation_on_visual_and_audiovisual_stimulus_based_emotion_recognition_using_EEG

Ockey, G. J., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value*, *7*(1), 1–21. https://doi.org/10.6035/LanguageV.2015.7.2.

Pardo-Ballester, C. (2016). Using video in web-based listening tests. *Journal of New Approaches in Educational Research*, *5*(2), 91–98. https://doi.org/10.7821/naer.2016.7.170.

Parkinson, B., Totterdell, P., Briner, R. B., & Reynolds, S. (1996). *Changing moods: The psychology of mood and mood regulation*https://pdfs.semanticscholar.org/19f9/7eabf17d312e86f763f1ae5b704e303c8455.pdf

Peixoto, F., Mata, L., Monteiro, V., Sanches, C., & Pekrun, R. (2015). The achievement Emotions Questionnaire: validation for pre-adolescent students. *European Journal of Developmental Psychology*, *12*(4), 472–481. https://doi.org/10.1080/17405629.2015.1040757.

Pekrun, R. (1992). The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. *Applied Psychology: An International Review*, *41*(4), 359–376. https://doi.org/10.1111/j.1464-0597.1992.tb00712.x.

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*, 315–341. https://doi.org/10.1007/s10648-006-9029-9.

Pekrun, R., & Stephens, E. J. (2010). Achievement emotions: a control-value approach. *Social and Personality Psychology Compass*, *4*(4), 238–255. https://doi.org/10.1111/j.1751-9004.2010.00259.x.

Pekrun, R., Goetz, T., Frenzel, A. C., Barchfeld, P., & Perry, R. P. (2011). Measuring emotions in students' learning and performance: the achievement Emotions Questionnaire (AEQ). *Contemporary Educational Psychology*, *36*(1), 36–48. https://doi.org/10.1016/j.cedpsych.2010.10.002.

Quintelier, A., Vanhoof, J., & Maeyer, S. D. (2019). A full array of emotions: an exploratory mixed methods study of teachers' emotions during a school inspection visit. *Studies in Educational Evaluation*, *63*, 83–93. https://doi.org/10.1016/j.stueduc.2019.07.006.

Riviello, M. T., & Esposito, A. (2016). *On the perception of dynamic emotional expressions: a cross-cultural comparison*. Springer.

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

Schreuder, E., Erp, J. V., Toet, A., & Kallen, V. L. (2016). Emotional responses to multisensory environmental stimuli: a conceptual framework and literature review. *SAGE Open*, *6*(1), 1–19. https://doi.org/10.1177%2F2158244016630591.

Schwartz, G. E., & Weinberger, D. A. (1980). Patterns of emotional responses to affective situations: relations among happiness, sadness, anger, fear, depression, and anxiety. *Motivation and Emotion*, *4*(2), 175–191. https://doi.org/10.1007/BF00995197.

Stientjes, M. K. (2012). A study of listening Assessment Stimuli and Response Mode Effects. *International Journal of Listening*, *12*(1), 29–39. https://doi.org/10.1080/10904018.1998.10499017.

Suvorov, R. S. (2008). *Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format* [Doctorate thesis, Iowa State University]. https://lib.dr.iastate.edu/rtd/15448

Taffou, M., Guerchouche, R., Drettakis, G., & Viaud-Delmon, I. (2013). Auditory-visual aversive stimuli modulate the conscious experience of fear. *Multisensory Research*, *26*(4), 347–370. https://www.researchgate.net/publication/259250392_Auditory-Visual_Aversive_Stimuli_Modulate_the_Conscious_Experience_of_Fear.

Takagi, S., Hiramatsu, S., Tabei, K., & Tanaka, A. (2015). Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality. *Frontiers in Integrative Neuroscience*, *9*(1), https://dx.doi.org/10.3389%2Ffnint.2015.00001.

Tran, V. (2007). *Functionality, intentionality and morality* Emerald Insight.

Tuan, N. H., & Mai, T. N. (2015). Factors affecting students' speaking performance at LE Thanh Hien High School. *Asian Journal of Educational Research*, *3*(2), 8–23. https://www.semanticscholar.org/paper/FACTORS-AFFECTING-STUDENTS'-SPEAKING-PERFORMANCE-AT-Tuan-Mai/e23c9e4360079b812d924ed1229907098a9dda9d#citing-papers

Uyl, M. J. D., & Kuilenburg, H. V. (2005). *The FaceReader: Online facial expression recognition*. Paper presented at Measuring Behavior 2005, The Netherlands. http://www.vicarvision.nl/pub/fc_denuyl_and_vankuilenburg_2005.pdf

Valiente, C., Swanson, J., & Eisenberg, N. (2012). Linking students' emotions and academic achievement: when and why emotions natter. *Child development perspectives*, *6*(2), 129–135. https://doi.org/10.1111/j.1750-8606.2011.00192.x.

Viola, P., & Jones, M. (2001). *Rapid object detection using a boosted cascade of simple features*. Paper presented at the IEEE Computer SocietyConference on Computer Vision and Pattern Recognition, Kauai, HI, U.S.A. https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf

Wagner, E. (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning & Technology*, *11*(1), 67–86. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44089/11_01_wagner.pdf.

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing*, *27*(4), 493–513. https://doi.org/10.1177/0265532209355668.

Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures. A meta-analysis. *European Journal of Social Psychology*, *26*(4), 557–580. https://doi.org/10.1002/(SICI)1099-0992(199607)26:4%3C557::AID-EJSP769%3E3.0.CO;2-4.

Yazdani, A., Skodras, E., Fakotakis, N., & Ebrahimi, T. (2013). Multimedia content analysis for emotional characterization of music video clips. *EURASIP Journal on Image and Video Processing*, *26*, https://doi.org/10.1186/1687-5281-2013-26.

Zumhasch, J. (2018). *Validation-study: basic emotions and action units detection. [Blog post]*. :text=A.,additionally%20the%20neutral%20facial%20expressions. https://www.noldus.com/blog/validation-study-facereader#:~.