DISSERTATION APPROVAL SHEET

Title of Dissertation: DEVELOPING AN APPROACH TO SUPPORT INSTRUCTORS TO PROVIDE EMOTIONAL AND INSTRUCTIONAL SCAFFOLDING FOR ENGLISH LANGUAGE LEARNERS THRO

Name of Candidate: Heera Lee

Doctor of Philosophy, 2022

Graduate Program: Human-Centered Program

Dissertation and Abstract Approved:

Ravi Kuber

Ravi Kuber

Associate Professor

Information Systems, UMBC

4/13/2022 | 9:15:47 AM PDT

Jiaqi Gong

Jiaqi Gong

Associate Professor

Computer Science, University of Alabama

4/13/2022 | 12:22:36 PM EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

ABSTRACT

Title of Document:	**DEVELOPING AN APPROACH TO SUPPORT INSTRUCTORS TO PROVIDE EMOTIONAL AND INSTRUCTIONAL SCAFFOLDING FOR ENGLISH LANGUAGE LEARNERS THROUGH BIOSENSOR-BASED FEEDBACK**
	Heera Lee, PhD, 2022
Directed By:	Dr. Ravi Kuber, UMBC Dr. Jiaqi Gong, University of Alabama

Delivering a presentation has been found to be one of the most anxiety-inducing tasks faced by English language learners (ELLs). Researchers suggest that instructors should be more aware of the emotional state of ELLs to provide the appropriate level of emotional and instructional scaffolding support needed to improve presentation performance. Challenges can be faced by instructors attempting to determine the emotional state of ELLs solely through observation of their facial expressions and behaviors. To address the ambiguity in this process, the research described in this dissertation focuses on identifying the potential of using biosensor-based feedback to support instructors. A novel approach has been adopted to classify the intensity and characteristics of public speaking anxiety and foreign language anxiety among ELLs, with a view to provide tailored feedback to instructors. A focus group interview was conducted to identify instructors' needs for solutions providing emotional and instructional support for ELLs. This was followed by an ideation and design session, where prototypes incorporating biosensing technology were designed to support teaching. Findings informed the design of a more refined prototype system, which was evaluated with 17 English language instructors. The contributions of this research include: (1) the demonstration of the feasibility of using electrodermal activity to measure emotional states of ELLs during a classroom presentation; (2) the development of an algorithm for classifying degree of anxiety and predominant type of speaking anxiety among learners; (3) the development of an approach to investigate methods to measure levels of anxiety among ELLs through the use of a biosensor and to design and evaluate solutions to support educators using this technology; and (4) design guidance for an educational system using EDA data within an ESL/EFL environment.

DEVELOPING AN APPROACH TO SUPPORT INSTRUCTORS TO PROVIDE EMOTIONAL AND INSTRUCTIONAL SCAFFOLDING FOR ENGLISH LANGUAGE LEARNERS THROUGH BIOSENSOR-BASED FEEDBACK

By

Heera Lee

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Acknowledgements

Foremost, I would like to express my deepest gratitude to my co-advisors, Dr. Ravi Kuber and Dr. Jiaqi Gong who have spent a tremendous amount of time and effort to advise me over the past several years. I sincerely appreciate their immediate responses and detailed feedback on my dissertation to guide me in delivering a better quality of work. In addition to my advisors, I would like to thank my committee members: Dr. Andrea Kleinsmith, Dr. Anita Komlodi, and Dr. Timothy Bickmore, for their constructive comments and questions that encouraged me to review my studies with fresh eyes and eventually carried integral justification in my dissertation. I would also like to extend my gratitude to Dr. Ehsan Hoque and Dr. Lujie Karen Chen who were kindly and sincerely providing their interest, encouragement, and insights towards my research, which made me feel motivated to continue my study.

There is a special recognition for Dr. Yuliya Schmaltz who gave me the opportunity to teach in the English Language Institute (ELI) at University of Maryland, Baltimore County (UMBC). Without the teaching experience at the ELI, it would not have been possible for me to discover my research topic and apply for the doctoral program to address the research questions. I was fortunate to work with the instructors and the students in the ELI thanks to their willingness to participate in my studies. By taking this section, I would like to thank them again for allowing me to investigate their data and have a chance to understand them by listening to their stories and coming up with suggestions.

I would also like to thank my fellows who have supported me in this long Ph.D. journey with me: Varun Mandalapu, staying up many nights with me to write papers with red eyes, taught me many valuable lessons including being patient, positive and confident no matter what challenges are around. I must also thank Julie Haney, Anna Smooth, and Jonathan Hunter-Cevera for proofreading my lengthy dissertation by devoting their time. I really appreciate their effort and patience for proofreading and sending me cheerful messages to support me having enough strength to complete this study.

I am also grateful to have Galina Madjaroff Reitz, May Chung, Michele Williams, In-seok Baek, Myung-kyun Sung, Swamp Squad – Abdullah X. Ali, Adegboyega Akinskiu, Erin Buehler, Gabrielle Salib, Patrick Carrington, William Easley, and Garden church Korean youth group who have always been there for me to support whenever I was physically and mentally burned out while studying.

Last but not the least, I would like to thank my family including my parents, brother, sister-in-law, adorable two nieces, and aunt for their continuous and bubbly support whenever I was in a dark time with insomnia and depression from intensive writing. Even though we were in a 12-hour difference in a different time zone, their endless love was crossing over space and time to sustain me this far in this long Ph.D. journey.

Finally, I would like to thank God, for holding me up every time whenever I was unsure about my ability to study and wanted to give up this Ph.D. journey. When I could not share all my difficulties and frustration with my family members, I spiritually felt His guidance and support to be able to overcome the challenges. I truly appreciated His guidance allowing me to finish my degree. To return His love, I will use my degree for helping people in need and keep being humble by remembering the fact that I accomplished this degree by His support and the support from the people around me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

- EEG: Electroencephalogram
- EDA: Electrodermal activity
- EL: English language
- ELI: English Language Institute
- ESL/EFL: English as a Second Language/ English as a Foreign Language
- HR: Heart Rate
- PD: Professional Development
- SC: Skin Conductance
- SCL: Skin Conductance Level
- SCR: Skin Conductance Response
- TESOL: Teaching English to Speakers of Other Languages

- **Writing Style**

  In this dissertation, the third-person singular term 'investigator' has been used instead of the first-person singular pronoun 'I' (Thomas et al., 2007).

Chapter 1. Introduction

1.1 Speaking Anxiety among English Language Learners

Delivering an oral presentation with a strong degree of fluency is deemed an important skill for English language learners (ELLs) to academically and professionally succeed in an era of globalization. However, for ELLs, speaking has been found to provoke greater levels of anxiety compared to other language skills including reading, writing, or listening (Awan & Anwar, 2010; Brantmeier, 2005; Horwitz, & Cope, 1986; MacIntyre & Gardner, 1991; Yaikhong & Usaha, 2012). The speaking anxiety that ELLs often experience has been investigated since the 1960s, focusing on both Public Speaking Anxiety (PSA) and Foreign Language Anxiety (FLA) (Al-Nouh et al., 2015; Al-Saraj, 2014; Brookes & Ross, 1967; Horwitz et al., 1986; Moskowitz, 1965; Radzuan & Kaur, 2011; Yalçın & İnceçay, 2014; Young, 1990). PSA is often referred to as 'stage fright' or the social anxiety that a person experiences while delivering a speech to an audience (Bodie, 2010; Young, 1990), whereas FLA is considered as a negative emotional reaction when a language learner is using a second language or foreign language (Horwitz et al., 1986; MacIntyre & Gardner, 1991; Yeşim & Backus, 2019). Although the triggers that provoke PSA and FLA are slightly different (e.g., pressure of public attention verses negative sense of self-linguistic competence), the terms PSA and FLA have been used interchangeably to describe the general characteristics of speaking anxiety in prior studies (Güvendir et al., 2020; Luo, 2014). The studies described in this dissertation focus on disambiguating PSA and FLA among ELLs, and investigating how speaking anxiety interplays both with behaviors and physiological responses during presentation performance.

While facilitative anxiety can motivate ELLs to work harder and use more complex sentence structure, debilitative anxiety adversely affects learners' engagement in class and limits their capacity of working memory and performance (Eysenck, 1979; Kleinmann, 1977). In fact, the debilitating anxiety faced by students can lead to avoidance behaviors (e.g., "I wish this was over") and negative self-related cognition, including thoughts of failure (Dewaele et al., 2008; Eysenck, 1979; Schwarzer, 2013). Examples of avoidance behaviors among anxious ELLs include absence from class, procrastinating on completing assignments, avoiding eye contact with the audience, and communicating with them by using a foreign language (Horwitz et al., 1986; Kleinmann, 1977). As anxiety arousal interferes with ongoing cognitive activity (Tobias, 1979), this interference reduces the learners' ability to take in information and learn new material (Khan & Zafar, 2010). Eventually, the increased levels of anxiety lead to poorer levels of preparation for presentations (Daly et al., 1995), which can negatively impact performance among learners (Alemi et al., 2011; Daly, 1991).

Although ELLs suffer from negative anxiety when delivering presentations in English, educational systems still mainly focus on the evaluation of presentation skills (e.g., eye-contact) and verbal communication skills (e.g., fluency, pronunciation, voice volume) as a requirement for courses. Assessing the emotions of learners, particularly the anxiety that the learners may experience during the presentation took second place to the evaluation of the presentation performance. For instance, studies and training systems relating to public speaking have often focused on modifying behavioral mannerisms or prosody among speakers to help strengthen presentation performance (Chen et al., 2014; Chollet & Scherer, 2017; Chollet et al., 2015; Damian et al., 2015; Menzel et al., 2000; Schneider et al., 2015; Tanveer et al., 2016). However, it is important for educators and researchers to first understand what triggers learners' anxiety to provide appropriate emotional and instructional support. By identifying the emotional cues and weaknesses that evoke anxiety among ELLs when delivering presentations, tailored feedback and practices can be developed for ELLs, which ultimately lead to improved levels of presentation performance (Gregersen, 2020).

1.1.1 Importance of Emotional and Instructional Support for English Language Learners

To help anxious ELLs to better achieve their academic goals, emotional and instructional support is proposed as an important quality for EL instructors (Benesch, 2013; Brackett & Katulak, 2006; Epp, 2016; Hargreaves, 1998; Lopez, 2012; Rosiek, 2003; Zhang & Pelttari, 2014) . The term 'support' is borrowed from Vygotsky's concept of scaffolding and combined with the way that instructors use emotions and instructions to support students in learning and development (Mahn & John-Steiner, 2002; Meyer & Turner, 2007). According to Vygotsky (1978), a learner's potential ability will be maximized via scaffolding that occurs during interaction with an instructor or a more advanced peer.

The instructor can provide emotional support through expressions of care, concern, affection, and interest, especially when a person is encountering levels of stress or feeling upset (Albrecht & Adelman, 1987). A person who receives emotional support feels comfort, security, and a feeling of being cared for, which can lead to higher levels of academic achievement (Burleson, 2003b, 2003a; Cutrona & Russell, 1990; Hamre & Pianta, 2005; Roeser, Skinner, Beers, & Jennings, 2012; Selvi, 2010; Hagelskamp et al., 2013). In parallel with emotional support, several types of instructional scaffolding (Lopez, 2012) — including modeling, bridging, building schema, and developing metacognition — can be applied to learners (Walqui, 2006). For instance, instructors can present clear examples of which tasks learners are required to perform (modeling), use learners' prior knowledge to learn new concepts and language (bridging), organize ideas as clusters (building schema), and ask learners to monitor their current level of understanding and make a decision by using it (developing metacognition) (Bransford et al., 2000).

To prepare instructors to understand learners' emotional states and provide personalized instruction, instructor training and assessments (e.g., Academic, Social, and Emotional Learning (CASEL), Classroom Assessment Scoring System (CLASS)) have been developed for educators (Allen et al., 2013; Brackett & Katulak, 2006). During the process of training, instructors are encouraged to develop skills identifying positive and negative classroom climates and reading emotional cues from learners based on their non-verbal behaviors, and practice responding to students to address their academic and emotional needs (Allen et al., 2013; Gregersen, 2007; Pianta et al., 2008). These qualities can be defined as *interpersonal or affective sensitivity* referring to "the ability to detect and describe the immediate affective state of another" (Kagan, 1967), or even more broadly, *perceptions of emotional intelligence (EI)* (Mayer et al., 1990) or *interpersonal intelligence* (Gardner, 1999). Instructors who exhibit higher levels of emotional intelligence are able to better monitor their own feelings, as well as those of others.

Being aware of the emotional states of learners should be prioritized by instructors, as it may help the instructors foresee difficult learning situations among learners, and develop instructions to prevent learners' negative emotions from escalating (Brackett & Katulak, 2006; Gregersen, 2020; Pozo-Rico & Sandoval, 2020).

1.1.2 Challenges of Providing Support

As noted, emotional and instructional scaffolding from instructors can positively impact students' academic performance and mental health. However, it is not an easy process for instructors to identify students' internal emotional states solely through observation of their behaviors. This is because emotional expressions "vary substantially across cultures, situations, and even across people within a single situation" (Barrett et al., 2019). It may be even more difficult for instructors to anticipate areas of difficulty or frustration for students when a relationship with a student has yet to be built. This can be challenging particularly at the beginning of the semester. Unless students intentionally disclose their feelings to instructors either verbally, or through their facial expressions, body language or gestures, the process of detecting emotional states poses difficulties for instructors.

A further set of obstacles faced by English language (EL) instructors include language barriers and cultural gaps in communication with ELLs. Even though learners often want to fully express their frustration and ask for help from their instructors, their non-native language skills, combined with their negative sense of self-linguistic competence, may limit ELLs from conveying their feelings of frustration and anxiety (Yeşim & Backus, 2019). In addition to the language barrier, some students are too embarrassed to express their emotions externally due to their own cultural norms (Barrett et al., 2019). In this context, the instructors may misinterpret the learners' behaviors and emotions during class. Years of teaching experience may allow instructors to better sense the emotional states of learners because they can make assumptions about their emotional state based on their experiences of working with students from prior classes. However, solely relying on instructors' personal interpretations could lead to hasty and biased judgments being made. As emotions are very personal and convey a range of information, what one instructor perceives may be different from another. Inconsistent emotional evaluations of learners can lead to different pedagogical beliefs among instructors in prioritizing students who need immediate emotional and instructional support.

To address the ambiguous interpretation of a student's emotional state, understanding physiological changes across learners in real-time could be a supplementary method of supporting instructors when making assessments. Using a tool to monitor these changes, instructors would be able to modify their teaching to better support those encountering high levels of anxiety. Among technologies that have been used to support the identification of emotional states, biosensor-based feedback has shown considerable promise (Chollet & Scherer, 2017; Croft et al., 2004; Giraud et al., 2013; Kimani & Bickmore, 2019; Pertaub et al., 2002; Sanches et al., 2019; Schwerdtfeger, 2004; Yeşim, 2017; Wörtwein et al., 2015). However, biosensor technologies have yet to be broadly utilized within educational settings, particularly in classes with ELLs from varying cultural and linguistic backgrounds.

Thus, biosensing technologies have been used in the research described in this dissertation as a way to objectively measure the emotional states of ELLs in real-time. The aim of the research described in this dissertation is to identify whether EL instructors would be able to use the biosensor-based feedback to compare the degree of emotional states across students, with a view to using this feedback to promote emotional scaffolding.

Furthermore, this indicator can be used as a reference among instructors in conversation with other instructors, which will reduce the inconsistency in emotional evaluation of student state. Inconsistencies may be attributed to instructors' subjective observations and interpretations of learners' emotional states based upon their experiences with former students.

1.2 Purpose of this Research

As described in the prior section, challenges can be faced in identifying emotional states among ELLs, which may lead to uncertainty among instructors as to how to better support their students. The limitations of learners' language skills to express their emotions, and inconsistency among instructors' ambiguous and subjective criteria to read learners' emotional states, can cause difficulties. Although surveys and questionnaires such as the Strengths and Difficulties Questionnaire (SDQ) have been used by instructors to measure pro-social behavior and emotional symptoms of students (Poulou, 2020), the issue of disagreement on the questionnaire was reported (Cheng et al., 2018). Therefore, in this dissertation, the investigator examined whether physiological cues from learners can be used to support instructors' abilities to objectively detect and compare emotional states among ELLs, in addition to relying on using their own personal criteria and through observation of facial gestures and body language.

To address this, the studies described in this dissertation explored the speaking anxiety of ELLs at a fine-grained level by investigating the datasets in cognitive, physiological, and behavioral aspects and identifying the distinctive characteristics of PSA and FLA. These multiple data sources need to be conveyed to the instructors so that they can articulate information to develop personalized emotional and instructional scaffolding strategies and practices for individual learners.

With the purpose of developing a solution that better matches instructors' needs, particularly within an ESL/EFL setting, this research aimed to gather perspectives from instructors. The investigator conducted a design thinking exercise with in-service instructors and explored their needs and suggestions for educational systems which offer insights into the emotional states of students. The investigator applied their ideas to the development of a prototype system. The system informed instructors regarding levels of anxiety among ELLs, alongside their predominant speaking anxiety (PSA and FLA). Target users were invited to interact with the system to determine the efficacy of the feedback for determining the emotional states of ELLs, compared to methods that they would otherwise use (e.g., observation of facial expressions, drawing upon their experience with former students etc.).

Last but not least, the studies were designed to explore both PSA and FLA of ELLs in a classroom setting. Prior studies examining anxiety among learners have often been conducted in highly controlled lab-based settings, which can impact levels of ecological validity. By collaboratively developing an educational system with target users, the aim of this dissertation was to develop and evaluate an aid to support instructors' awareness to detect the emotional state of learners in the context of ESL/EFL teaching and learning.

1.3 Research Questions

In this dissertation, a set of research questions were developed. The main research question addressed in this work was:

- RQ: How could providing awareness of the emotional states of ELLs, through the use of a biosensor, impact teaching?

A set of sub-research questions were developed as part of this work.

Study 1:

- RQ 1-1: Which features extracted from electrodermal activity (EDA) data collected when presenting in class can be used to classify levels of anxiety?
- RQ 1-2: Which non-verbal behaviors made when presenting correspond to EDA arousal (peaks), indicating anxiety?

Study 2:

- RQ 2: Can PSA and FLA be classified using EDA data among ELLs when delivering a presentation in English?

Study 3:

- RQ 3-1: Which traditional methods have EL instructors used to identify the emotional states of learners in class?
- RQ 3-2: How did/would the awareness of learners' emotional states impact instructors' teaching?
- RQ 3-3: How would the experienced EL instructors integrate biosensor-based feedback of learners' speaking anxiety into a prototype system?

Study 4:

- RQ 4-1: Assessment between raters

  Is there agreement between the instructors on ranking the learners' degree of speaking anxiety

    - **without** using the biosensing data presented on the prototype system and solely relying on observation (Evaluation 1)?
    - **with** – referring to the biosensing data on the prototype (Evaluation 2)?
- RQ 4-2: Assessment by rater

  When it comes to ranking the learners' degree of speaking anxiety, are there any similarities or differences between Evaluation 1, Evaluation 2, and biosensing based ranking data on the prototype (ground truth)?
    - Compare Evaluation 1 (observation) and Evaluation 2 (observation + referring to the biosensing data on the prototype system)
    - Compare Evaluation 1 and ground truth (biosensing data on the prototype system)
    - Compare Evaluation 2 and ground truth
- RQ 4-3: How does the instructors' awareness of emotional states of ELLs vary over the experiments?
- RQ 4-4: How would instructors use information presented via a system relating to the emotional states of students, in terms of teaching strategy or class activities?

1.4 Novelty and Contributions

- The investigator demonstrated the feasibility of using electrodermal activity (EDA) to measure the emotional state of ELLs during a classroom presentation. By applying EDA technology to a group of participants hailing from diverse linguistic and cultural backgrounds, in a classroom setting where the presentations counted towards the final grades for the course (i.e., high stake), it enabled the investigator to present empirical evidence, with which educators can recognize the intensity and types of anxiety among ELLs through processed EDA data.

- The investigator developed and validated a machine learning algorithm that can classify PSA and FLA for ELLs using biosensor data. PSA and FLA have historically been used interchangeably in the context of speaking anxiety among ELLs. However, no clear criteria exist for differentiating speaking anxiety into the categories of PSA and FLA. This information would be valuable to instructors, as tailored emotional and instructional support based on different types of learner anxiety could be provided via a system. Thus, the investigator attempted to develop a model for distinguishing PSA and FLA and demonstrating the anxiety dynamics with the aim of assisting instructors with providing personalized emotional and instructional support. The investigator compared descriptions relating to anxiety among ELLs with physiological data and behaviors, to create an input dataset for the algorithm. The algorithm has shown the predictive ability to classify ELLs into either PSA or FLA, which may have been overlooked in the past.
- A novel approach has been developed to investigate methods to measure levels of anxiety among ELLs through the use of a biosensor and to design and evaluate solutions to support educators using this technology. Few studies have made use of physiological data to support foreign language teaching. Driven by advanced sensing technologies and machine learning algorithms, this innovative research approach demonstrated how the speaking anxiety of ELLs can be objectively measured and compared across ELLs. This information contributes to foreign language educators because it offset the educators' subjective interpretation of the

learners' emotional states. The educators are able to use the information provided to objectively track and evaluate learners' emotional states. This information can be used as a datapoint to communicate with other instructors to better support students in the classroom. When a high level of EDA arousal is detected, instructors can then utilize methods such as modifying their teaching style and content, and use terms and grammar which may be more understandable. They could then highlight the value of fluency rather than focusing on accuracy. Examples of how to do this could be to encourage students to repeat the same exercise until their levels of confidence increase or ask students to work in pairs with other ELLs where they can practice without feeling embarrassed about their abilities.

The mixed methods approach used as part of the Design Thinking Process (DTP) included a focus group, sketching session, individual interviews, and ranking surveys. These helped the investigator to both develop a targeted educational prototype system for educators using EDA technology, and to evaluate its potential application for teaching based on the perspectives of educators.

- Findings from the studies resulted in the development of design guidance to better support educators aiming to use biosensor-based feedback to aid their teaching. The investigator designed an interface that displays levels of anxiety, along with the predominant type of speaking anxiety among students. The intended goal is that instructors using the intervention will be able to better perceive the emotional states of learners and utilize this information to support instruction. The outcome of these studies contributed to the development of an objective indicator of PSA and FLA for instructors. Instructors can use the anxiety indicator of ELLs to make a decision in identifying which students exhibiting higher levels of anxiety need urgent support, especially when time is at a premium. Moreover, they can identify anxiety-related challenges in class as they arise in real-time, using this as a tool to reflect upon how further classes are designed.

1.5 Structure of the Dissertation

Chapter 2: Related Work

Chapter 2 offers an overview of anxiety including PSA and FLA, its measurements, and how emotional support has been provided for learners' anxiety.

Chapter 3: Classifying Anxiety of English Language Learners during Presentation Performance in Class

Chapter 3 reviews the studies (Study 1 and Study 2) conducted by the investigator which relate to the main research question. It presents innovative algorithms that include classifying the intensity of anxiety and identifying predominant speaking anxiety between PSA and FLA that ELLs experienced during presentation performance.

Chapter 4: Using Participatory Design to Develop an Interface to Support Instruction

A focus group interview has been described in this chapter, with the aim of identifying the experiences of ESL/EFL educators when aiming to perceive emotional states of ELLs to provide emotional and instructional support in teaching. Participants designed educational prototype systems through hand-drawn sketches integrating a biosensing technology. From this work, the investigator was able to outline the initial design of an educational prototype system that the ESL/EFL stakeholders wanted including the types of platforms, data, functions, interactions, and structure of the interface.

Chapter 5: Developing a Prototype System using Biosensor-Based Information

The needs, perspectives and prototype sketches gained from Chapter 4 facilitated the development of a prototype system using biosensor-based information of ELLs. A low-fidelity prototype was redesigned and developed into a more refined high-fidelity prototype based on the feedback that in-service ESL/EFL instructors and ELLs gave during the individual interview as a pilot study.

Chapter 6: Evaluation of the Efficacy of a Prototype System using EDA Data

The final high-fidelity prototype delivering audio/video recordings of the learners' presentation performance, EDA data, and processed information of EDA data to inform the degree and types of anxiety was evaluated by a total of 17 experienced and novice instructors to determine whether anxious ELLs could be effectively identified.

Chapters 7 and 8: Contributions, Discussion and Conclusion

In Chapter 7, the four main contributions are described, along with the implications of the research. The paradigm of using EDA technology in an educational field and research design is also discussed.

Chapter 2. Related Work

This chapter explores prior work examining anxiety, specifically concerning PSA and FLA. It focuses on examining how researchers have measured levels of anxiety using quantitative and qualitative methods, and identifies the methods and technologies used for emotional and instructional scaffolding developed to address issues attributed to anxiety.

2.1 Anxiety

2.1.1 Defining Anxiety

Anxiety is defined as the "subjective feeling of tension, apprehension, nervousness, and worry associated with an arousal of the autonomic nervous system" (Spielberger, 1983). In the study of anxiety, two concepts are often described. *Trait* anxiety is characterized as stable personality, which is "an individual's likelihood of becoming anxious in any situation" whereas *state* anxiety is temporarily experienced at a particular moment in time responding to a specific anxiety-provoking situation (e.g., "Are you nervous now?") (MacIntyre & Gardner, 1991; Spielberger, 1983; Tanveer, 2007). The intersection of these two approaches has been described as *situation-specific* anxiety. MacIntyre & Gardner (1991) described trait of anxiety in a case of examining anxiety reactions in a "well-defined situation" such as public speaking, during tests, or in a foreign language class. *Situation-specific* anxiety was explored in the studies described in this dissertation, as the studies were conducted in foreign language classes where presentations were delivered. Thus, this research has explored both *state* and *trait* anxiety of ELLs (which is *situation-specific* anxiety) by employing various instruments to measure PSA and FLA. More specifically, the physiological arousal detected by a biosensor and anxious behaviors were used to examine the *state* of anxiety in real time, whereas self-reported data from participants assessing their levels of anxiety relate to *trait* anxiety.

2.1.1.1 Public Speaking Anxiety

Public speaking anxiety (PSA) is a subtype of social anxiety where negative self-focused cognition (e.g., "I never feel quite sure of myself when I am speaking in my class"), behavioral concomitants (e.g., avoiding eye contact), or physiological arousal (e.g., increased heart rate) occur when individuals give an expected or actual presentation in public (Ayres & Heuett, 1997; Daly, 1978; McCroskey, 1970). Various labels — including speech fear, stage fright, speech anxiety, performance anxiety, audience anxiety, and "the threat of unsatisfactory evaluations from audiences" (Schlenker & Leary, 1982a, 1982b) — have been used in the previous PSA studies (Daly, 1978). The labels fright, fear, threat, and anxiety were used interchangeably in the past, but the studies described in this dissertation consistently use the term 'anxiety' as an umbrella term for PSA labels. In fact, fear is usually derived from a real danger in the external environment, but anxiety is raised by a vague and unknown fear indirectly associated with an object (Scovel, 1978). The fears of humiliation and embarrassment in specific social situations were brought together under the term 'PSA' (Blöte et al., 2009). In the studies described in this dissertation, PSA has been defined as a subtype of **social** anxiety that ELLs may experience during an oral presentation in front of an audience of a little more than 15 students in a class. The ratio of eye contact with the audience during the presentation performance was identified to predict the students who have predominant PSA (Chapter 3).

2.1.1.2 Foreign Language Anxiety

The concept of Foreign Language Anxiety (FLA), known as language learning anxiety or language anxiety, was first conceptualized by Horwitz et al. (1986). It is defined as "a distinct complex of self-perceptions, feelings, and behaviors related to classroom learning arising from the **uniqueness of the language learning process**." In particular, the feelings of inadequacy and fear of failure are associated with class activities in a second or foreign language learning context (MacIntyre & Gardner, 1994; Wang, 2005). FLA is a subtype of performance anxiety experienced within a specific academic context. Horwitz et al. (1986) identified three related performance anxieties arising from the unique context of foreign language learning: communication apprehension; test anxiety; and fear of negative evaluation.

- *Communication Apprehension*

  Horwitz et al. (1986) stated that communication apprehension is manifested when ELLs feel difficulty in speaking in front of the class ('stage fright') or in listening to a spoken message (receiver anxiety). This apprehension is attributed to "a disparity between learners' mature thoughts and their immature foreign language proficiency" (Gregersen & Horwitz, 2002; Horwitz et al., 1986), and their performance is constantly monitored by both their teacher and peers (Daly, 1991; Onwuegbuzie et al., 1999).

  Thus, the inability to fully express oneself in speaking or to understand others by listening to foreign language content often causes frustration in communication among ELLs. Even though stage fright and performance anxiety overlap with attributes of PSA, the source of FLA is highlighted on the apprehension derived from limited foreign language speaking and listening skills rather than from the presence of the audience.

- *Test Anxiety*

  ELLs who encounter test anxiety tend to experience stress due to putting pressure on themselves to succeed when utilizing their foreign language skills. Even though making errors is natural in the language learning process, learners can feel a sense of failure when this happens (Aida, 1994; Horwitz et al., 1986). The anxiety examined in the studies described in this dissertation involved test anxiety, as the presentation performance served as a graded deliverable for a class.

- *Fear of Negative Evaluation*

  In contrast to test anxiety, fear of negative evaluation can be widely applied to any social or evaluative situation in a psychological construct (Watson & Friend, 1969). It relates to "apprehension about others' evaluations, avoidance of evaluative situations, and the expectations that others would evaluate oneself negatively." Horwitz et al. (1986) characterized that language learners who are highly concerned about the impressions of others tend to minimize the possibility of negative evaluations by acting passively. For example, they are likely to avoid attending a class or withdraw from classroom activities to avoid situations where anxiety may be experienced. This fear of negative evaluation is similar to PSA in terms of the fear of being introduced to others. In this dissertation, the investigator has explored the fear of negative evaluation within the specific context of foreign language learning. Unlike the social anxiety associated with PSA, the investigator has placed more weight on the fear of negative evaluation that is attributed to the negative sense of self-linguistic competence within a specific context of foreign language learning (Yeşim & Backus, 2019). For example, grammatical mistakes or disfluency in verbal presentation are common concerns reported by ELLs in studies relating to anxiety (Horwitz et al., 1986). Overall, the investigator has applied the term ‘PSA’ to social anxiety corresponding with the audience, whereas the term ‘FLA’ has been applied to the apprehension triggered by novice foreign language skills.

2.2 Measurements of Anxiety

Researchers have focused on three aspects of anxiety and fear of communication: cognitive, behavioral, and physiological (Beidel et al., 1985). Each aspect has been investigated with the measurements conducted through self-report (Gilkinson, 1942; Lomas, 1934), observer ratings on behaviors (Bell, 1853; Henning, 1934), and physiological arousal (Bagchi & Greenwald, 1937; Neumann & Blanton, 1970). The combination of these methods has been used to validate levels of experienced anxiety (Chollet & Scherer, 2017; Giraud et al., 2013; Gregersen et al., 2014). Oral presentations have been used by researchers as one of the common tasks to provoke situation-specific anxiety from participants. However, the majority of studies have been conducted in a laboratory setting where a speaker often spontaneously delivers presentations based on researcher-prescribed topics alone in the room, or in front of a monitor, video cameras (Fung et al., 2015; Giraud et al., 2013; Schneider et al., 2015; Tanveer et al., 2016), or in front of a virtual audience (Chollet et al., 2018; Kimani & Bickmore, 2019; Pertaub et al., 2002; Takac et al., 2019; Trinh et al., 2017). Although the feelings of anxiety are real, these settings are far from a high-stake testing environment (e.g., delivering a presentation based on a topic selected by the presenter/instructor, taking enough time to prepare and practice a presentation, delivered face-to-face in front of a physical audience, or graded as an activity which would count towards the class). In a foreign language context, instructors believe using oral presentations as a tool for assessment provides an opportunity for learners to practice using English to communicate with others in a natural way (Brooks & Wilson, 2014), even though ELLs reported that it is one of the most anxiety-provoking tasks they have performed (Horwitz et al., 1986). Thus, the studies described in this dissertation were conducted in an English language learning class where natural anxiety exists among ELLs. To preserve levels of ecological validity, presentation tasks were not prescribed by the investigator, but were created by course instructors as part of the curriculum. While prior studies have used simple and static stimuli that insufficiently presents aspects of real-world activities and interactions (Parsons, 2015), the research described in this dissertation is valuable to satisfy a growing interest in investigating contextually embedded stimuli via experiments conducted within a real world setting.

2.2.1 Self-assessment

2.2.1.1 Questionnaires

To measure levels of speaking anxiety among presenters, self-report questionnaires have been traditionally employed in many psychological and educational studies. One of the common self-report assessments of PSA is *Personal Report of Public Speaking Anxiety (PRPSA)* (McCroskey, 1970). It consists of 30 items with a 5-point Likert scale (ranging from 1 = strongly disagree to 5 = strongly agree) with higher total scores representing higher anxiety.

When it comes to exploring the subtle effects of anxiety on a specific situation such as foreign language learning, Horwitz et al. (1986) developed *Foreign Language Classroom Anxiety Scales (FLCAS)* targeting foreign language learners' anxiety. It has 34 items with a 5-point Likert scale the same as PRPSA, and is reflective of communication apprehension, text anxiety, and fear of negative evaluation in the foreign language classroom. The items were endorsed as indicative of speech anxiety (Sarason, 1984). The FLCAS questionnaire has been frequently used in empirical studies of foreign language education to identify correlations between anxiety reactions and variables including gender, socio-cultural background, language proficiency ratings, and language performance.

For instance, the results of FLCAS administered to university students in Korea (Park & French, 2013) and Turkey (Çağatay, 2015) show that females seem to be more anxious than male counterparts when speaking. On the other hand, conflicting findings that male ELLs in the U.S. or in Singapore had higher levels of anxiety than females were reported (Campbell & Shaw, 1994; Zhang, 2001). With regards to socio-cultural backgrounds, Zhiping et al. (2013) demonstrated that Nigerians generally are less anxious of speaking, whereas Iranians and Algerians struggle with the anxiety attributed to communication apprehension and fear of negative evaluation. The proficiency level of a foreign language also affects anxiety responses in that lower proficiency level students tend to be more anxious than those students at a higher level of proficiency (Young, 1986). Other studies also found that there is either a positive or negative relationship between the anxiety level and the quality of foreign language speech. The anxiety termed 'debilitating anxiety' leads to a negative relationship with performance (Phillips, 1992) because it motivates learners to avoid anxiety-provoking situations. On the other hand, the anxiety that triggers the positive performance (Aida, 1994; Elkhafaifi, 2005) is labeled as 'facilitating anxiety', which motivates learners to actively perform the task (Scovel, 1978).

Although many variables are associated with the level of anxiety identified, the conflicting findings from opposing variables (male vs. female, high vs. low proficiency level, positive vs. negative relationship with performance) have made it difficult to ascertain the factors contributing to the anxiety. Thus, the studies described in this dissertation are open to the possibility that those variables could be reinterpreted. More importantly, the research described in this dissertation focuses on the individual differences across ELLs, rather than making a generalized profile of ELLs in relation to the level of investigated anxiety, with the aim of providing personalized support in educational instruction.

Questionnaires are widely used by researchers to gather data as they are arguably fast and cost-efficient to administer. While questionnaires can be used to directly gather information from participants, the process is disruptive and subjective because the participants can suffer from information recall bias (Hernandez et al., 2014). Moreover, it may be inappropriate for certain populations, such as ELLs, to complete questionnaires as they may not be able to understand the intended meaning of questions written in English. Moreover, cultural differences may require the revision of questionnaires for participants to meaningfully answer each item (Al-Saraj, 2014). Thus, the self-report questionnaires in the studies described in this dissertation were triangulated with other measurements including in-depth individual interviews, physiological arousal (i.e., increased EDA) and observable behaviors (i.e., eye-contact, filler words and pauses) to cross-check qualitative data and increase the credibility and internal validity of the research (Patton, 2014).

2.2.1.2 Interviews

Interviews and focus groups have been widely used by many researchers to better understand experiences and feelings of ELLs delivering a presentation in a foreign language. In several interview studies, students who were found to experience high levels of anxiety, had a tendency to be perfectionistic (O'Connor et al., 2010; Price, 1991), competitive (Bailey, 1983), and excellent in using an unfamiliar language (Hilleson, 1996). In another study undertaken in Turkey, Yalçin et al. (2014) asked 14 students to complete FLCAS and answer open-ended essay questions. Findings showed that students felt less anxious when they could prepare their talk using topics familiar to them and wait for the instructor to call on them. In the focus group, four participants representing the most and the least anxious learners reported that they felt a sense of safety in group-based activities over individual activities because group members help each other by explaining themselves in a foreign language. Radzuan & Kaur (2011) also conducted four focus groups to explore the main sources of anxiety in delivering oral technical presentations in English. The results revealed that demanding evaluation panels, immature content knowledge, and obstacles in students' English language proficiency contributed to the anxiety experienced.

Overall, PSA and FLA were further examined through exploring interview statements (Bailey, 1983; Hilleson, 1996; Price, 1991; Radzuan & Kaur, 2011; Yalçın & İnceçay, 2014). PSA (e.g., concerns of panels, preference for group activity) and FLA (e.g., English proficiency) were intertwined in the interview statements. However, variables highlighted from these studies were investigated in Chapter 3.

Even though self-assessments like questionnaires and interviews have demonstrated some interesting relationships between anxiety and other variables, the retrospective, self-report methods show a bias regarding recalling the negative features of events. Self-reported responses are likely to be exaggerated in a negative direction compared to the average ratings of concurrent moods (Sato & Kawahara, 2011). Thus, the following measurement of behavioral observation by others has been explored as a way to supplement reliability issues of self-assessments.

2.2.2 Behavioral Observation

Behavioral observation refers to collected data representing a firsthand encounter with the phenomenon of interest. This observation is natural and produces trustworthy results when conducted by trained observers (Merriam & Tisdell, 2015). To evaluate how nervous the speaker appears during the presentation, several descriptive indices of observable behaviors were drawn from prior scholarly reviews and discussions with experienced instructors of public speaking as described below.

One of the most common non-verbal behaviors exhibited in PSA studies is eye contact with the audience. Daly et al. (1989) developed descriptive items to ask observers to report the following while speakers were presenting: the frequency with which the speaker looked at the audience, the number of times the speaker notably read or looked at their notes, and the number of times the speaker looked at the floor. The findings showed that speakers with higher levels of anxiety showed lower levels of eye contact with the audience (including the instructor) and focused more on their notes as compared to the speakers with lower levels of anxiety, which also found in the ELLs in this chapter. This is because they may "focus their attention excessively on themselves when speaking, reducing in turn, their attention toward the audience and the setting" and are hesitant to interact with the audience (Daly & Stafford, 1984; Gregersen, 2005). In addition to levels of eye contact, the number of nervous gestures made by the speaker (e.g., fidgets) and the amount of time the speaker moved around have been also explored as related to the level of speaking anxiety (Daly et al., 1989).

The affective state of a person can be objectively measured by analyzing features from speech. The number of stutters that a speaker makes, and the number of unnatural pauses has been used to determine the level of anxiety experienced (Daly et al., 1989; Na, 2007). Anxious presenters have been found to more frequently stutter and use more filler words (e.g., "um.." and "uh..") while taking longer pauses compared to less anxious presenters (Duvall et al., 2014). In particular, in the context of foreign language learning, the learners were found to use longer pauses, word retrieval pauses, and filler words (Gregersen et al., 2014).

The presenters' behaviors can be manually annotated by a human annotator, which can be more accurate than automated behavioral annotation via a system because it allows synchronization of the annotated behaviors with the verbal content of presenters. Even though this approach allows for the investigation into the context of which behaviors are triggered, it requires a substantial amount of time to annotate each video frame. It is also expensive to hire experts to identify and interpret behaviors. As an alternative, less-expensive, micro-level annotations in the cloud using crowdsourcing were introduced

(Park et al., 2014). However, this solution presented privacy issues as people view videos and judge the performance and body language of presenters.

As technology advances, the annotation of behaviors that were once manually annotated by human coders has been automated. Many systems have embedded behaviors — including eye gaze (Chollet et al., 2015; Echeverría et al., 2014), gestures (Chen et al., 2014; Fung et al., 2015; Tanveer et al., 2016), stutters (Park et al., 2014; Schneider et al., 2015; Trinh et al., 2017), and pauses (Chollet et al., 2016; Park et al., 2014; Schneider et al., 2015) — as a ground truth to annotate presenters' behaviors. However, the main purpose of using these automated behavioral annotations is mostly to assess presentation performance and not to recognize the learners' emotional state.

Bearing in mind the studies covered in this review, the investigator adapted observable behaviors of ELLs as a part of indicator to measure the ELLs' levels of anxiety when presenting. The ratio of eye contact with the audience and reading or looking at notes/slides (e.g., eye gaze) was selected with the aim of measuring the social anxiety of PSA. Since eye contact is likely to increase the communication attempts and interaction between people (Richmond et al., 2008), the frequency and duration of eye contact can be considered as measuring elements of social anxiety, namely PSA. On the other hand, the duration of pauses and frequency of filler words that ELLs exhibit during an oral presentation were selected to measure FLA, as the disfluency and pauses can be attributed to the following factors: being afraid of making grammatical mistakes and fear of negative evaluation while using a foreign language (Horwitz et al., 1986).

In this dissertation, the investigator has triangulated these behaviors with the data from other measurements, such as self-assessment by ELLs along with their levels of physiological arousal. Triangulation is necessary because negative emotional facial expressions and behaviors can be controlled during social communications (Shu et al., 2018) or masked due to different emotional display rules based on cultural norms (e.g., Japanese culture) (Ekman, 1973; Keltner & Anderson, 2000). For example, people can smile at a formal social event even if they are experiencing a negative emotional state. Thus, the investigator has measured the internal physiological arousal that detects the emotional states of ELLs in real time as an objective surrogate measurement to offset the weakness of subjective self-report assessments and controlled behavioral observation (Meehan et al., 2002).

2.2.3 Physiological Assessment

A wearable biosensor has been proposed as a useful and objective medium to assess the affective status of a subject, as it allows the measurement of the internal physiological variables in real-time. Physiological cues which can be gathered include electroencephalogram (EEG), heart rate (HR), galvanic skin response (GSR) (also known as skin conductance (SC)), or electrodermal activity (EDA) (Kreibig, 2010; Meehan et al., 2002; Shu et al., 2018).

EEG detects electrical activity from the human brain. Sensors are placed across the person's head to detect signals, gathering information from the brain. Researchers have investigated working memory capacity (cognitive load) and stress of people through the use of EEG (McEvoy et al., 2000; Mills et al., 2017). Findings have shown that EEG can be a viable source of data to model learners' mental states. However, the experimental setting needs to be controlled (more similar to a laboratory environment than to a field setting) as participants are required to wear an EEG headset with between 8 to 512 electrodes when collecting data. Since the EEG electrodes attached to a participant's scalp are used to analyze brain activity, the number of electrodes used, along with their placement on the scalp are important (Usakli, 2010). Even though a single portable electrode EEG system has been developed, further testing is still needed to determine its efficacy (Cheemalapati et al., 2013). On the other hand, a wristband sensor is easier to wear and minimally interferes with real-world activities (Riedl et al., 2017). As a result, many studies on PSA have adopted a wristband sensor to measure HR or SC from participants performing cognitive tasks (Chen et al., 2016; Kye et al., 2017; Qu et al., 2020) or anxiety-inducing public speaking (Bodie, 2010; Croft et al., 2004; Giraud et al., 2013; Kothgassner et al., 2016; Schwerdtfeger, 2004). The findings from prior studies showed that increased levels of HR and SC activity were identified when performing tasks presented, as an indicator of anxiety.

More specifically, SC known as EDA is commonly considered as one of the most reliable indicators of sympathetic arousal and more sensitive to emotion related variations in arousal, as opposed to stress and anxiety (Boucsein, 2012). EDA is composed of two types of GSR activity, skin conductance level (SCL) and skin conductance response (SCR). The tonic SCL stage shows the slowly changing EDA levels, whereas the phasic SCR stage presents the rapidly changing peaks (Poh et al., 2010). These segments of SC allow researchers to refer to the mean changes in tonic SCL as a possible standardization of personal variability (Dawson et al., 2017) and the phasic changes in SCR to determine a stimulus by quick inspection of peaks and amplitudes.

Regarding PSA, prior studies explored SC to investigate speech anxiety in response to pronounced differential facial expressions. Pictures of angry and happy faces (Dimberg & Thunberg, 2007) and a virtual audience manifesting positive to negative facial expressions (Chollet et al., 2018; Pertaub et al., 2002) were used to provoke social stimuli (e.g., stares from an audience). Furthermore, the scenario of delivering a talk in public was presented to participants and physiological signals were recorded when delivering a presentation in a laboratory setting or in class. With regard to FLA, a few studies have been conducted to assess the efficacy of EDA or HR in the context of using a foreign language among immigrants (Gregersen et al., 2014; Yeşim, 2017). However, further work is needed to offer a deeper insight into how this data can help support users.

As described in prior work, the autonomic arousal associated with PSA or FLA has been studied in clinical settings or real-world contexts, and researchers identified that physiological arousal can be used as an objective marker to assess one's anxiety. However, few studies have been conducted to investigate both PSA and FLA with tasks presented in less controlled environments. Thus, the investigator conducted her studies within a classroom at the university's English Language Institute (ELI) where ELLs may experience testing anxiety during an English presentation performance. This is due to the high-stake nature of the presentation, counting significantly towards the final grade. The presentation tasks were created by ELI instructors. The instructors decided on the theme and the format of each presentation task, and the ELLs selected presentation topics based on personal preference and/or familiarity with the area. The attention from an audience in class or fear of using a foreign language can be a stimulus for elevated anxiety in experimental settings of this study. Moreover, these stimuli allow the investigator to unobtrusively and continuously monitor participants' EDA responses to speaking anxiety in real time as part of the class.

Despite the advantage of using objective physiological data in assessing a person's anxiety, some limitations can be faced. The changes in EDA do not occur in isolation (Dawson et al., 2017), so the context and variables (e.g., environment, events) that would impact autonomic arousal should be considered. Moreover, EDA signals are easily vulnerable to the artifacts that involve body gestures and movements, which can skew the analysis of the signal, leading to misinterpretation (Taylor et al., 2015). Manual inspection to decide which parts are too noisy to retain is available to address the problems of artifacts, but this approach cannot apply to large-scale EDA studies. Currently, many researchers employ exponential smoothing (Hernandez et al., 2014) or a low-pass filter (Sano & Picard, 2013) to remove artifacts and noise, but these techniques can still affect small variations in the signal.

Overall, the studies described in this proposal extend the scope to assess PSA and FLA of ELLs with objective physiological arousal in addition to self-report questionnaires, interviews, and behavioral observations.

2.3 Emotional and Instructional Support for English Language Learners

2.3.1 Krashen's Affective Filter Hypothesis

Many educational professionals have highlighted the need for additional emotional and instructional support in learning environments, particularly for those learners who struggle with negative emotions (Brackett & Katulak, 2006; Gregersen, 2020; Hamre & Pianta, 2005; Lopez, 2012; Walqui, 2006). Most notably, Krashen (1982) noted that the affective variables of learners play an important role in second/foreign language acquisition. According to his *Affective Filter Hypothesis*, "the negative affective variables including low levels of motivation and self-esteem, and high levels of debilitating anxiety 'raise' the affective filter and form a 'mental block' that prevents comprehensible input form used for language acquisition" (Schütz, 2007). Therefore, the instructors' role is to create a welcoming learning environment where the learners can feel safe and comfortable to learn and perform in a foreign language.

2.3.2 Support for Anxious Learners in Educational Environment

In order to provide differentiated instructional scaffolding based on the current emotional states of learners, instructors need to perceive learners' various moods and potentially be able to act on this knowledge. Gardner (1999) emphasized that instructors need to have the capacity to understand other people's intentions, motivation, and desires to effectively interact with them (termed: *'interpersonal intelligence'*). Gregersen et al. (2014) also claims that instructors need to identify their students' emotional triggers to help emotionally vulnerable ELLs by considering their individual variances.

As sensing technology has advanced, many interventions have been developed to help users provide improved presentation performance through regulation of their emotions. For example, a false cue of a slow heart rate (Costa et al., 2016) and real-time visual and haptic feedback on users' public speech (Bubel et al., 2016) were presented to users to increase their awareness of emotional states and public speaking skills. Moreover, visual cues to minimize levels of cognitive load were provided for users to assist them in being less pressured when undertaking public speaking activities (Dermody, 2016). However, these interventions were aimed at assisting participant presenters who may feel anxious while delivering a presentation in public. In fact, an instructor needs to be aware of learners' emotional states to provide tailored emotional and instructional support for them, so that the learners can achieve their academic goals in a comfortable learning environment with step-by-step instruction.

Recently, studies have been conducted to help instructors increase their capacity to perceive the learners' emotional states. For instance, a range of studies have used physiological synchrony between learners and instructors (Di et al., 2018) and among learners (Gashi et al., 2018) to help them increase emotional engagement in class and create a positive classroom climate. However, a need still remains to develop educational tools aiding foreign language instructors in a specific learning situation. The different mother tongue and cultural background between instructors and learners may hinder instructors from identifying and understanding their students' emotional states. Thus, the investigator has conducted studies to address the limitations found in responsive teaching among EL instructors (Gay, 2018; Jiang et al., 2016; Santamaria, 2009), with a view to enhancing their ability to detect learners' emotional states by employing triangulation of multiple anxiety measurements introduced in this chapter (e.g., questionnaire, interview, physiological arousal, and observable behaviors) and assess the following main research question "How does providing awareness of the emotional states of ELLs, through the use of a biosensor, impact teaching?" in the remainder of this dissertation.

2.3.3 Privacy and Ethical Concerns

As awareness of the importance of student mental health has gained more prominence among educators, interest has grown in using technologies such as wearable biosensors to recognize and assess emotions to support educational activities. However, the ethics of affective computing are still in their infancy, and researchers have begun constructing ethical application guidelines for when capturing and working with data relating to emotions (Hernandez et al., 2021; McStay & Pavliscak, 2019; Steinert & Friedrich, 2020). The researchers have described the challenges associated with emotional recognition. These include the concern about the potential misuse of data, and the invasion of privacy for information that could be considered personal. The process of emotional recognition relates to the judgment under uncertainty from evolving human emotions (Barrett et al., 2019; Richardson, 2020; Russell, 1994). Therefore, it is critical to consider a way of describing and labeling of human emotions, and set a boundary between what should be private and public on the system (Hernandez et al., 2021).

The research described in this dissertation has taken into account the checklist of ethical application guidelines proposed by Hernandez et al. (2021) into account when developing the prototype system proposed in Chapter 5 (see Table 1). This system was used as a prompt to aid discussion relating to the ethical perspectives which EL instructors held relating to biosensing technology for emotional recognition (Chapter 6). To minimize the likelihood of making inaccurate and presuming judgments about the degree of students' anxiety (G1), numeric physiological data was used as the 'ground truth'. For purposes of transparency, the investigator shared the purpose of the system's use, limitations of bioinformation, and described scenarios of how the system could be used with participants

(G3). All participants including the users of the biosensor and bystanders (EL instructors) were allowed to opt-in prior to any measurements that were collected, or drop out at any point during the study. Participants were not forced to use the affective data or information presented on the prototype system. They were given the freedom to choose the information needed to make a decision of levels of anxiety faced by students (G4 to G6). Moreover, the participants had opportunities to provide feedback on performance and efficacy of the system and to describe further modifications needed (G9).

Table 1 Guidelines for emotion recognition applications adopted by Hernandez et al. (2021)

Responsible Communication
G1. Predictions are not handled as ground truth
G2. System descriptions should be described with granularity
G3. Technology should be described with transparency Informed Consent
Informed Consent
G4. Opt-in is facilitated before measurements are performed
G5. Data handling is described to facilitate comprehension
G6. Consent facilitates freedom of choice without consequences Contextual Calibration
Contextual Calibration
G7. Training data is representative of real-life data
G8. Sources of variance are accounted by the models
G9. Users can customize the system by providing feedback Comprehensive Contingency
Comprehensive Contingency
G10. Personal data can be deleted by the user
G11. Feedback channels are provided to the users
G12. Shifts in data distribution are detected to ensure robustness

Chapter 3. Classifying Anxiety of English Language Learners when Presenting in Class

3.1 Introduction

The potential offered by using EDA to detect emotional state was described in Chapter 2. It is important to present physiological data (EDA) collected in a manner that enables instructors to effectively identify learners who need immediate and tailored support in class. In this chapter, two studies are described which have led to two of the contributions described in Chapter 1. Namely, the demonstration of the feasibility of using EDA to measure emotional states of ELLs during a classroom presentation (Study 1), and the development of an algorithm for classifying degree of anxiety and predominant type of speaking anxiety among learners (Study 2).

3.2 Study 1: Classification of High and Low Anxious English Language Learners

The aim of the exploratory study (Study 1) was to examine the feasibility of using EDA and non-verbal behavior information of learners to understand their emotional state (primarily ‘speaking anxiety’) when delivering a presentation. The investigator used a wireless biosensor, which resembles a wristwatch, to collect physiological data from participants. As the wireless device was easier to don compared to a wired biosensor prototype, users were able to move their wrists freely during the study. Moreover, E4 is also considered “more robust” in handling noise introduced from movement, so it is appropriate for a real word setting. Among physiological signals including heart rate or blood volume pressure, EDA was selected as it is known to offer promise measuring both levels of anxiety and frustration among users (Dawson et al., 2017; Kreibig, 2010). To the best of the investigator's knowledge, researchers have yet to use a biosensor in a classroom setting with English language learners (ELLs) from diverse cultures to measure their emotional states. The investigator developed the following research questions to examine the feasibility of using EDA and nonverbal behavior information together to explore how students experience anxiety when delivering a presentation. EDA has been used to gather data about emotional states. However, it remains unknown which features of EDA would be useful to extract to better understand speaker anxiety. This study was the first step to identify the efficacy of EDA to classify emotions, which could in turn lead to the development of a system for purposes of supporting instruction.

3.2.1 Research Questions

- RQ 1-1: Which features extracted from electrodermal activity (EDA) data collected when presenting in class can be used to classify levels of anxiety?
- RQ 1-2: Which non-verbal behaviors made when presenting correspond to EDA arousal (peaks), indicating anxiety?

3.2.2 Methods

Participants were recruited from Speaking and Listening classes at the UMBC ELI with the purpose of collecting data in-situ. In these classes, instructors require students to deliver presentations in a classroom setting every two weeks. The investigator did not modify the assignment set by the instructors to promote ecological validity. 15 students (10 male, 5 female) ranging in age from 19-40 ($M = 24$, $SD = 6.81$) participated in the study. Prior to delivering their presentations, the investigator helped the learners don a biosensor to collect physiological data, primarily EDA. The investigator also positioned a tablet at the back of the classroom to record the non-verbal behaviors of each presenter (e.g., head direction towards audience, self-grooming, hand gestures, reading or looking at notes, etc.). By positioning it at the back of the classroom, it was out of sight for participants, enabling them to focus on the task rather than getting distracted by the presence of the tablet. Students were asked to prepare for the presentation task a week in advance of the class. The task was a talk summarizing a news article and describing their reason for selecting the article. Students were given between 3-5 minutes to present, and were asked to use phrases and vocabulary learned in class. They were allowed to use notes while delivering presentations. This study was approved by the Institutional Review Board (IRB) at UMBC.

3.2.3 Analysis and Results

Prior to analyzing the EDA data, the physiological data needed to be smoothed to remove rare occasions noise/artifacts that can be too excessive and distort the signal shape (Braithwaite et al., 2013). Ledalab (Benedek & Kaernbach, 2010) was used to remove artifacts caused by body gestures and movements. The data was then min-max normalized to reduce inter-individual variance to enable comparison across learners. As shown in Table 2, five EDA features were extracted, including the mean and standard deviation of the normalized EDA, the number of peaks, mean peak amplitude, and max peak amplitude. A peak occurs in reaction to a single stimulus in the EDA signal, and the EDA peak height (amplitude) carries information about the stress level of a person (Setz et al., 2009). These features have been typically explored in the studies of psychophysiology to measure one’s level of mental stress.

Table 2. Results of *k*-means clustering. The EDA features are listed in the leftmost column. Final cluster centroids of each feature are listed in the remaining columns.

EDA features	**Cluster 1** (n = 7)	**Cluster 2** (n = 8)	**Full Data** (n = 15)
Number of Peaks	96.71 ± 19.72	50.38 ± 11.72	72 ± 28.43
Mean Peak Amplitude	0.11 ± 0.24	0.08 ± 0.26	0.10 ± 0.34
Max Peak Amplitude	0.51 ± 0.21	0.36 ± 0.19	0.43 ± 0.21
Mean EDA	0.46 ± 0.1	0.05 ± 0.10	0.48 ± 0.10
SD EDA	0.21 ± 0.04	0.26 ± 0.03	0.24 ± 0.04

The investigator employed *k*-means clustering to investigate whether the EDA features can automatically classify students into two groups ($k = 2$), i.e., higher and lower levels of anxiety. The k-means clustering technique is usually used for unlabeled data with the aim of finding the number of groups represented by variable *k*. As Table 2 illustrates, Cluster 1 (n = 7) has a greater number of peaks with a higher mean amplitude, higher max amplitude, a lower mean and standard deviation of EDA compared to Cluster 2 (n = 8). In this context, Cluster 1 can be described as demonstrating higher levels of anxiety compared to Cluster 2. To statistically validate the difference between two clusters, Mann-Whitney U-tests were conducted. As a result, two of the EDA features showed significant differences as follows: number of peaks ($U = 0.000$, $z = -3.24$, $p = 0.001$) and standard deviation of normalized EDA ($U = 5.0$, $z = -2.66$, $p = 0.008$). This is shown in a 2D plot in Figure 1 where ELLs could be distinguished into two groups. The two data ("O") on the x-axis could be considered outliers.

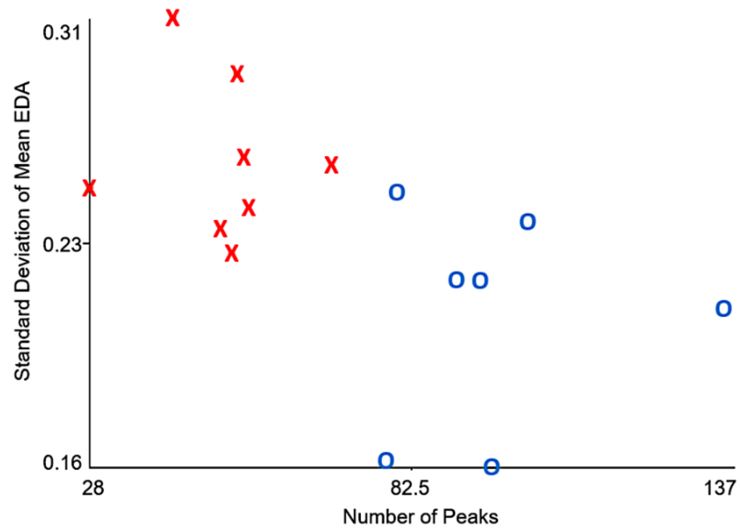


Figure 1 Plot of the two most discriminating features resulting from k-means clustering.

In addition to the EDA features, the investigator explored the non-verbal behaviors that can be indicative of a presenter's anxiety, such as avoiding eye contact and reading or looking at notes (Chollet & Scherer, 2017; Daly et al., 1989; Giraud et al., 2013; Schneider et al., 2015). The investigator hypothesized that these nonverbal behaviors (e.g., eye contact, self-grooming, etc.) may correspond to the onset or offset of a peak in the EDA data. So, the timestamps in a local time were added to each EDA value and synchronized with the recorded video according to the formula instruction introduced on Empatica website (https://support.empatica.com/hc/en-us/articles/202800715-Session-start-time-format-and-synchronization-). To develop a coding scheme, the investigator asked three experienced ELI instructors to watch the audio-video recorded presentations and provide examples of nonverbal behaviors seemingly indicative of anxiety.

As shown in Table 3, the final four main categories of the coding scheme were developed based on instructors' comments after watching videos of students presenting. These included avoiding eye contact, staring, touching their bodies, fidgeting, and reading or looking at notes: 1) head direction (straight, turned left or right); 2) self-grooming; 3) repetitive hand gestures/movements; and 4) reading and looking at notes.

Table 3 The nonverbal behaviors (NVBs) identified by instructors as indicative of public speaking anxiety and used for coding the presentations.

NVB categories	**Specific NVBs**
Head direction: *straight, left, right*	Staring at wall, avoiding eye contact
Self-grooming	Touching: nose, hair, hat, glasses, neck, wrists
Hand gestures	Fidgeting, wringing hands, repetitive motions
Notes	Reading/looking at notes, looking down

Among the instructors' comments, the investigator selected the final comments that were also mentioned in Daly's study as an indicator of anxious behavior (Daly et al., 1989). According to this coding scheme, the investigator coded non-verbal behaviors using BORIS annotation software (Friard & Gamba, 2016). Examples of the nonverbal behaviors (NVBs) data, including movement of the participant's head, were annotated. Touching parts of the body such as the nose, neck, wrist, hand, hair, or accessories (e.g., hat, glasses) were annotated as self-grooming. Hand gestures including fidgeting, wringing hands, and repetitive motions were also annotated. The times when the participant looked down to read his or her notes (e.g., script, index card etc.) were annotated.

The investigator explored the annotated nonverbal behaviors in relation to EDA peaks or peak amplitude. As Figure 2 shows, a student from the higher anxiety cluster demonstrates a greater number of self-grooming behaviors as peak amplitude increases. A greater number of peaks (80) is observed, which can be attributed to the frequent hand gestures and eye contact with the instructor throughout the presentation.

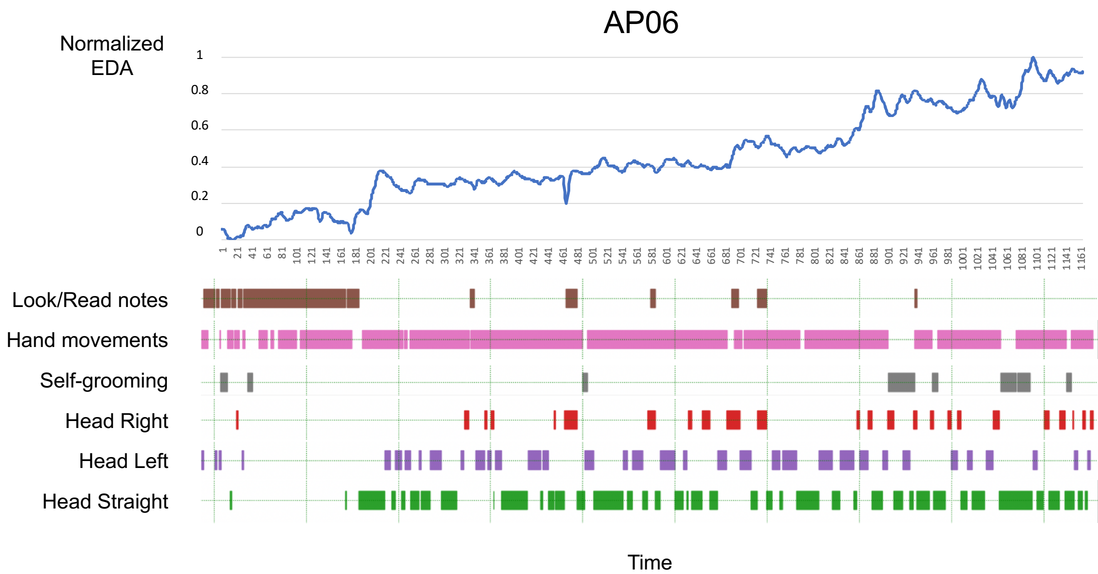


Figure 2 Graph of EDA and corresponding nonverbal behaviors of one student in cluster 1

In contrast, a student from the lower anxiety cluster (illustrated in Figure 3) shows relatively fewer peaks (52) on the EDA graph. When it comes to the non-verbal behaviors, instances of hand gestures and self-grooming were rare, whereas looking at/reading notes occurred more frequently during the presentation. These behaviors appear to correspond to the EDA segments consisting of amplitude peaks.

The annotated behaviors associated with high peaks in the EDA provide the ELLs and EL instructors with valuable insights into how speaking anxiety manifests itself in the behaviors made. Moreover, the behaviors performed immediately before or after a peak could indicate how the learners cope with their anxiety subconsciously. Thus, this information can offer valuable insights to both ELLs and instructors about the levels of anxiety experienced by presenters in class. For example, the ELLs can develop emotional regulation skills through the reviews of their behaviors in the video corresponding to EDA data. On the other hand, instructors can identify the moments where the learners feel anxious and provide instructional support (e.g., personalized practice to appraise the anxious moment and maintain mindfulness), which may alleviate negative emotions among learners, and support them during the process of presenting.

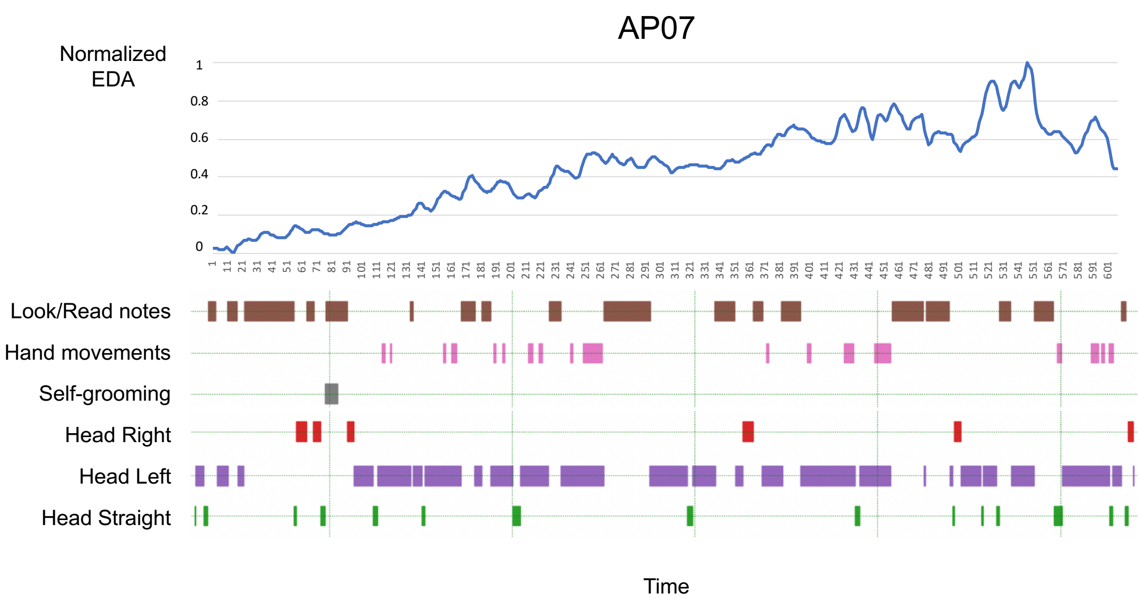


Figure 3 Graph of EDA and corresponding nonverbal behaviors of one student in cluster 2

As changes in behavior appeared to correspond to the EDA segments with higher amplitude peaks, a review of behaviors presented in the recorded presentation videos and corresponding EDA data were conducted in Study 2. This review provided instructors with an understanding of emotional states of the ELLs by associating those two modalities (EDA and behaviors). This finding was used when developing a prototype system in Chapter 5 to help instructors to monitor students' biometric data and reflect on their emotional states when presenting.

3.2.4 Summary

Although additional EDA features need to be extracted and further statistical tests are required to evaluate the correlation between the EDA and nonverbal behaviors, Study 1 demonstrated the potential of using EDA and non-verbal behaviors to understand how student's levels of anxiety can vary when delivering presentations. Further information about this study has been presented in the investigator's paper (Lee & Kleinsmith, 2019). To gain deeper and more personal insight into speaking anxiety, with a view to extending the classification of data, a qualitative study was conducted (Study 2).

3.3 Study 2: Identification of the Predominant Type of Speaking Anxiety among English Language Learners

For ELLs, PSA known as social anxiety (e.g., being afraid of attention from audience members) and FLA (e.g., fear of making mistakes in using a foreign language) are accompanied particularly by the presentation performance (Radzuan & Kaur, 2011). Even though ELLs struggle with these subtypes of speaking anxieties, many studies and educators focus on external properties in training (Chen et al., 2014; Damian et al., 2015; Schneider et al., 2015; Tanveer et al., 2016) rather than careful examination of discrete anxieties (Bosch et al., 2015; Dixon et al., 2012) which ultimately impacts the quality of presentation performance. To improve performance, ELLs need emotional clarity, which refers to the ability to identify the origins of emotions (Butler et al., 2018). By clearly identifying and distinguishing subtypes of speaking anxieties as an initial step, they can determine personalized emotional regulation strategies such as adapting or changing conditions to cope with each subtype of speaking anxiety. However, few studies discuss which subtypes of speaking anxiety predominantly affect ELLs. In fact, the causes of each anxiety are explained differently in psychological and foreign language studies. For example, PSA has been linked to environmental factors, such as the size or type of audience

(Bodie, 2010; Jackson & Latané, 1981) whereas FLA is attributed to the language learning context where the foreign language is not used as a mother tongue. To be more specific, a low level of foreign language proficiency causes concern regarding making grammatical mistakes or mispronouncing words, inability to find suitable words for expressing ideas, or difficulties remembering what to say (Horwitz, 2001; Liu, 2006; MacIntyre & Gardner, 1991). Therefore, the investigator noted the distinction between PSA and FLA as English language instructors can refer to this information to develop personalized class activities and teaching strategies for ELLs. To identify the predominant source of anxiety between PSA and FLA, the investigator examined EDA and behaviors of the ELLs, and conducted individual interviews with the ELLs to understand the deeper context of how ELLs may experience subtypes of speaking anxiety when delivering an oral presentation in an English language class.

3.3.1 Research Question

- RQ 2: Can PSA and FLA be classified using EDA data among ELLs when delivering a presentation in English?

3.3.2 Method

Fifteen participants (4 male, 11 female) were recruited in addition to the sample gained from Study 1. The reason for recruiting further participants was to obtain a gender-balanced sample and to help validate findings from Study 1. Additional measures including two questionnaires and individual interviews were adopted. Presentations from a total of 33 students (16 male, 17 female) from four different Speaking and Listening classes (henceforth referred to as A, C, D, E) at the ELI at UMBC were analyzed in Study 2. The participants hailed from a range of different countries and ranged in age from 19-43 (M = 23.58, SD = 5.67). Levels of English proficiency among participants were 'intermediate', which indicates that participants could produce sentence-level language to talk about familiar topics related to their daily life using English (ACTFL, 2012). The presentation task was to report a news story, similar to the task in Study 1. The only exception was for students in class E who were asked to deliver a talk related to a dream job. Even though the presentation topics differed, the general format of the presentation was kept consistent across the classes. For instance, students needed to deliver a summary of the topic and offer an opinion relating to the topic of the presentation. Additionally, they were permitted to use note cards and presentation slides if necessary, during their presentation.

Two self-report questionnaires were adapted from prior studies (Horwitz et al., 1986; McCroskey, 1970) and deployed to the additionally recruited participants (n = 15). The aim of the questionnaires was to gather data on anxiety to offer an insight to the difference between PSA and FLA. The Personal Report of Public Speaking Anxiety (PRPSA) questionnaire (Appendix A-1) has 30 items with a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree), assessing anxiety in a setting where public speaking is required. On the other hand, Foreign Language Classroom Anxiety Scales (FLCAS) questionnaire (Appendix A-2) has 34 items with a 5-point Likert scale as PRPSA, but the items refer to anxiety that arises from a foreign language learning environment. Higher scores in total represent higher levels of anxiety reported by participants.

Semi-structured individual interviews were conducted with the 15 additional students to gain insights into anxiety experienced when delivering a presentation in a foreign language. Each interview took between 15-30 minutes to conduct. All interviews were audio recorded using software on the investigator's laptop. The interview questions (Appendix B) were neutral in nature to allow the participants to freely express their emotional states based on their own experiences. The interviewer is a non-native English speaker, but she had four years of English teaching experience at the ELI. The investigator aimed at gaining a better understanding of the presenters' personal experiences and feelings of delivering a presentation in a foreign language through the responses from interviewees.

The procedure of collecting EDA and nonverbal behavioral data was the same as Study 1 (3.2.2).

3.3.3 Analysis and Results

3.3.3.1 Self-report questionnaires

After tabulating data, the Wilcoxon signed-rank test was conducted to validate the difference between the two responses between PRPSA and FLCAS. The results showed no significant differences between them ($z = -7.9$, $p = 0.426$). However, the overall responses on the questionnaire were intensively distributed in the moderate PSA and FLA scales.

3.3.3.2 Individual interviews

The investigator took notes during each interview, documenting responses. Interviewees shared experiences when presenting, along with their feelings about differences between delivering a talk in their native language and a foreign language. For instance, the interviewees reported that they feel less anxious when they present using their mother tongue. Concerns were also expressed regarding presenting in front of others and being the center of attention to an audience. These responses inspired the investigator to develop an affinity diagram with four categories of anxiety, shown in Figure 4. If students described positive experiences of delivering a presentation in English in class, their responses were allocated to the 'No anxiety (N)' category as a confident student. If their responses mainly accounted for the concern of the interaction with the audience as social anxiety (i.e., avoid eye-contact, hesitate to stand in front of people), the data was categorized as 'PSA more (P).' If the responses were associated with the fear of linguistic properties including grammatical mistakes, correct pronunciation, searching for a word (i.e., retrieval), or translation from a native language to a foreign language, it was more commonly labeled as 'FLA more (F).' If both PSA and FLA were mentioned in equal amounts by the interviewees, the data was counted as 'Both PSA and FLA (B).'

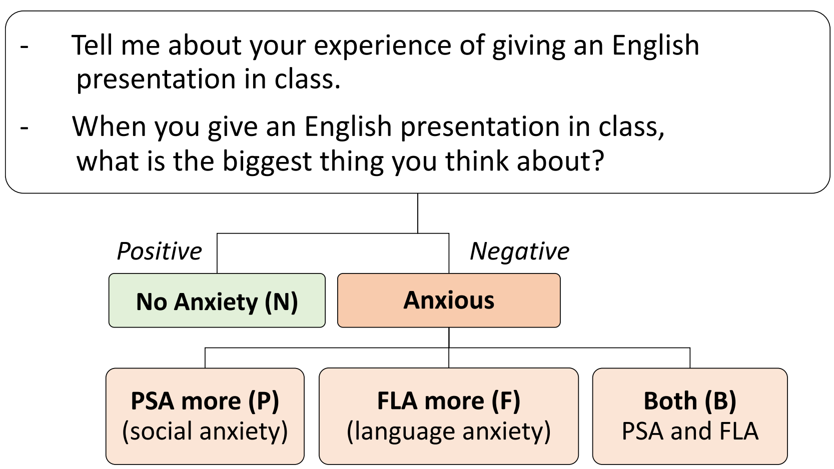


Figure 4 An affinity diagram with four subcategories of speaking anxiety:

Each interview was transcribed manually by the investigator. A thematic analysis (Braun & Clarke, 2006) was then conducted using this data. The repeated themes or words describing each type of anxiety were thematically organized within the affinity diagram as shown in Figure 4. Finally, some conceptual elements were grouped based on the four main anxiety categories. These are presented below.

- *No Anxiety (N): "I like to deliver my thoughts."* - The interviewees in this theme explained that they were not hesitant or afraid of delivering their thoughts to classmates using a foreign language (English). Moreover, they had accepted that they would make mistakes when presenting in English, but this would not cause anxiety, as they knew that the audience were all learning English. Thus, having an audience of ELLs would be supportive of them even if mistakes were made.
- *PSA more (P): "Everyone is looking at me."*- The interviewees in this group were very conscious of the audience and more likely to underestimate their public speaking skills. They knew how to deliver a presentation effectively in theory, but they felt that they could not utilize these skills when giving a talk while the audience was looking at them. They often expressed disappointment at their presentation performance in class.
- *FLA more (F): "I need to memorize my script not to make mistakes in English."* - The interviewees who reported FLA described that they heavily relied on the written script and memorized them when they deliver an oral presentation in English. They worried they might forget what they wanted to say, so were forced to memorize the written script.
- *Both PSA and FLA (B): "I translate from L1 (First Language, a person's first acquired language such as mother tongue or native language), \to English and look at my close friends whenever I feel nervous."* - The interviewees in the Both PSA and FLA (B) group presented a combination of PSA and FLA in their responses.

> When they were struggling with making eye contact with the audience, they tried to look at their close friends, often those from similar ethnic and cultural backgrounds. Moreover, they were likely to translate what they wanted to deliver directly from their native language (L1) to a foreign language, which could take time.

3.3.3.3 Behavioral annotations

To investigate the links between four subtypes of speaking anxiety and their corresponding behaviors stated in the interview responses, two annotating codes were retrieved from the interviewees' responses: 1) eye-contact linked to PSA as a social anxiety and 2) the number of pauses and filler words (i.e., "um.." and "ah..") linked to FLA as a factor of dysfluency. The coding schemes of eye contact, pauses, and filler words were adapted from previous studies (Chollet et al., 2015; Gregersen et al., 2014; Wörtwein et al., 2015). Two coders and the investigator manually annotated the behaviors recorded in the video using an annotation tool BORIS (Friard & Gamba, 2016). For purposes of annotating, eye contact was divided into a binary code, either 'Look' or 'Not look'. The number of pauses was counted when silence occurred for more than two seconds (Siegman & Pope, 1965) and the number of filler words were marked whenever they were voiced during the presentation. The number of pauses and filler words were normalized by the duration of the presentation in seconds, due to each participant presenting for varying amounts of time in class. The investigator manually annotated 33 audio-video recorded presentations, and two annotators randomly selected two audio-video recorded presentations to annotate the behavioral coding. The inter-rater reliability (IRR) among the three coders as measured by Cohen's

Kappa value was: eye contact = 0.812, pauses = 0.635, and filler words = 0.49 respectively. According to the interpretation of Kappa value (Fleiss, Levin, & Paik, 1981), there was almost perfect agreement on eye contact annotation (0.81-1.00) and substantial to moderate agreement (0.61-0.80, 0.41-0.60) on the annotations of pauses and filler words. In comparison to the behaviors of eye contact and pauses, the ELLs used filler words intermittently in very short seconds. This may have affected the agreement value of the filler words annotations because the annotation system counts milliseconds (e.g., to the third decimal place) during the Kappa test. In other words, there was a higher possibility to have lower agreement between the filler words that were annotated with a very short time than behavioral annotation lasting a longer time.

		Eye Contact	
		Look	Not Look
Pauses and Filler words	High	**FLA more (F)** • A2 (90.68%, 59.09%) • A3 (61.77%, 40.56%) • A4 (62.46%, 41.89%) • A6 (70.02%, 70.97%) • A7 (66.06%, 39.34%) • A8 (63.86%, 37.06%) • C4 (65.73%, 39.44%) • D2 (52.27%, 29.33%) • E2 (60.08%, 30.48%)	**Both PSA and FLA (B)** • A1 (43.40%, 26.79%) • C1 (12.17%, 38.3%) • C5 (24.13%, 26.83%) • C6 (43.85%, 32.58%) • C7 (18.08%, 43.84%) • C9 (27.74%, 37.5%) • D1 (17.18%, 27.52%) • D3 (20.16%, 53.1%) • D9 (13.65%, 60%) • E5 (42.27%, 45.71%)
	Low	**No anxiety (N)** • A5 (76.52 %, 25.35%) • D4 (52.76%, 12.4%) • D5 (62.19%, 12.59%) • E1 (65.60%, 21.74%) • E3 (69.47%, 22.86%)	**PSA more (P)** • A10 (32.47%, 10%) • C2 (15.86%, 4.76%) • C3 (36.52%, 25.93%) • C8 (37.29%, 24.04%) • D6 (18.84%, 13.94%) • D7 (37.42%, 6.98%) • D8 (38.93%, 4.24%) • D10 (33.08%, 9.09%) • E4 (27.77%, 6.90%)

Figure 5 Four types of anxiety framework referring to behavioral annotation data

The investigator developed a framework to categorize behaviors based on four types of anxiety – No anxiety (N), PSA more (P), FLA more (F), and both anxieties (B) as shown in Figure 5. The students were divided into two groups named Look (low PSA) and Not Look (high PSA) based on a 50% ratio of eye contact with the audience in the annotation. These two groups were divided into two subgroups again based on accumulated behavioral annotation on the number of pauses and filler words. These groups were labeled high pauses/filler words (high FLA) and low pauses/filler words (low FLA). The threshold of dividing these coordinates was 25% (e.g., greater than 25% is high FLA and less than 25% is low FLA), which corresponded with the interviewees' statements.

As a result, each quadrant represents different types of anxiety that ELLs may predominantly experience throughout the presentation. The investigator then extracted multiple EDA features to evaluate whether the EDA features can predict these different subtypes of speaking anxiety.

3.3.3.4 EDA dataset and features

The EDA data collected from 33 participants underwent multiple feature extraction steps. The Hanning window algorithm with a 1 second window time (Adamson et al., 2010) was used to reduce the artifacts in the EDA data. Then the tonic and phasic components in the EDA signal were extracted by using a non-negative deconvolution algorithm (Benedek & Kaernbach, 2010). A tonic component represents slow changes in the absences of an external stimulus, whereas a phasic component shows the variations that change quickly over time. These two components allowed the classification algorithms to learn efficiently. In addition, the Hilbert Huang Transformation (HHT) method (Li et al., 2011) was used to extract time-frequency and energy distribution from the EDA signal data. This method shows higher efficiency in determining ‘bases’ based on input signal instead of using predefined bases such as Fourier transforms (FFT) and wavelet decomposition.

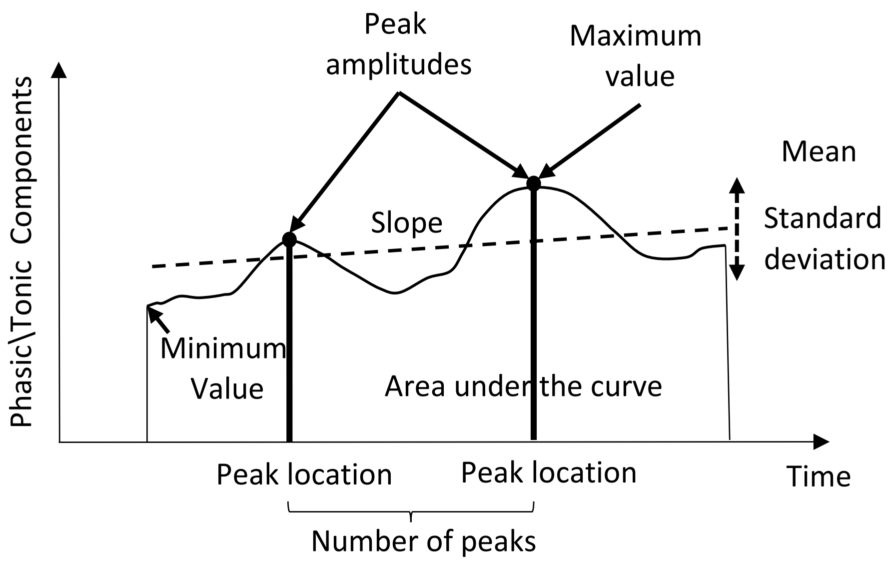


Figure 6 Ten features for both tonic and phasic component of EDA signal

Finally, the investigator extracted the following ten features for both tonic and phasic components of EDA signals as shown in Figure 6: mean, standard deviation, minimum and maximum values in a component, locations of minimum and maximum values, mean peak amplitudes, number of peaks, slope, and area under the curve (Hernandez et al., 2014), based on a sliding window of 10 seconds with an overlap of five seconds for each ELL individually. The time-frequency (TF) features for EDA signals were extracted from the Hilbert spectrum that is obtained by decomposing raw EDA signals using HHT based on the same sliding window as the phasic and tonic features.

3.3.3.5 Classification and feature importance

Three subsets of data were used to classify ELLs into one of the four types of anxiety categories illustrated in Figure 5: 1) all features from tonic and phasic components of EDA signal and time-frequency features from HHT, and either 2) tonic-phasic or 3) HHT features. Five machine learning algorithms including Decision Tree (DT), Auto Multilayer Perceptron (AMLP), Gradient Boosted Tree (GBT), Random Forest (RF) and Support Vector Machine (SVM) were adopted and developed on the RapidMiner data science platform (Mierswa & Klinkenberg, 2019). As hyperparameter tuning is important in terms of classification algorithms, grid search was used to optimize the model. To validate the model, a ten-fold cross-validation method was used to train nine subsets of data and one subset for testing and was iterated until the algorithm predicts for all samples in a dataset. In this analysis, the investigator focused on the features that influenced the top classification algorithm predictions. Thus, a Locally Interpretable Model Explanations (LIME) method was adopted (Ribeiro et al., 2016), and global weights of each feature were calculated to find out the top three features that support algorithm making predictions with high-performance metrics. Four metrics such as Accuracy, Cohen's Kappa, Recall, and Precision were used to evaluate each classifier's performance.

As a result, Gradient Boosted Tree (GBT) with phasic-tonic features outperformed other classifiers in assigning all ELLs to one of the four anxiety categories. Furthermore, the binary classifiers classified 18 ELLs with 2622 samples that belong to either PSA or FLA anxiety types. The GBT classifier performed well in predicting ELL anxiety type based on different input feature sets. Table 4 shows that the performance of GBT classifier with all features (HHT + Phasic-Tonic) has the highest performance.

Table 4 The performance of multi-class and binary class gradient boosting classifiers on different feature inputs.

Input Features	**Accuracy**		**Kappa**		**Recall**		**Precision**	
Class	**Multi**	**Binary**	**Multi**	**Binary**	**Multi**	**Binary**	**Multi**	**Binary**
PhasicTonic + HHT	60.61	100.00	0.45	1.00	61.67	100.00	85.87	100.00
PhasicTonic	75.76	94.44	0.67	0.89	75.56	94.44	75.56	95.00
HHT	57.58	88.89	0.41	0.78	54.55	88.89	60.42	90.91

Finally, we also extracted the feature importance of both multiclass and binary class from GBT model predictions with varying inputs based on a LIME method mentioned in the earlier section. Table 5 shows the top three supporting features of each classifier. It presents which features played a vital role in the case of classifying students into the four types of anxiety (No anxiety, PSA more, FLA more, or both anxiety) or binary types of anxiety (PSA more or FLA more).

Table 5 Top three supporting features of a GBT algorithm on different data subsets based on a LIME method.

Class_ Support Attribute	**Multi_ PhasicTonic + HHT**	**Binary_ PhasicTonic + HHT**	**Multi_ PhasicTonic**	**Binary_ PhasicTonic**	**Multi_ HHT**	**Binary_ HHT**
1	HHT Feature 0-0.1 Hz	HHT Feature 0-0.1 Hz	Min.Tonic component value	Sd. of Tonic data	HHT Feature 0-0.1 Hz	HHT Feature 0-0.1 Hz
2	Min. Tonic component value	Slope of Tonic data	Min. Phasic component value	Sd. of Phasic data	HHT Feature 1.8-1.9 Hz)	HHT Feature 1.8-1.9 Hz
3	Max. Phasic component value	Sd. of Phasic data	Max. Phasic component value	Max. Phasic component value	HHT Feature 1.5-1.6 Hz	HHT Feature 1.5-1.6 Hz

Further details regarding the study design and outcomes can be found in the investigator's paper (Lee et al., 2020).

This study focused on distinguishing types of speaking anxiety among ELLs in an English language learning context using physiological data gathered using wearable sensors. The resulting models showed promising results in classifying ELLs into one of the four anxiety types (e.g., PSA, FLA, Both, None). The features that played a prominent role in model decision making importance can be used to provide biofeedback interventions that support the mitigation of anxiety (MacLean et al., 2013; Schoenberg & David, 2014). As technology is improving rapidly, the adaptation of wearables is becoming easier than in prior decades (Zangróniz et al., 2017). Additionally, the use of wearable sensors to predict anxiety is less intrusive and reduces the bias associated with methods including adopting self-reported questionnaires to detect levels of anxiety (Hernandez et al., 2014).

In terms of implications, findings from this empirical study will help both ELLs and instructors to understand the presenters' learning process when delivering presentations. Firstly, four distinctive anxiety types were detected among ELLs. These can indicate that learners and educators can visually observe the predominant state of anxiety faced. This identification of the dominant anxiety state can encourage ELLs and instructors to reflect, and focus on teaching methods to address challenges experienced by the learner that triggered the identification of negative affective state.

3.3.4 Limitations and future work

While the results of Study 1 and Study 2 were promising, additional data collection is required to achieve adequate power in statistics and find significant differences between the anxiety levels and types. With the sufficient amount of data, the classifiers need to be used with a new class to estimate its performance. Also, a longitudinal study would be helpful to explore the shifts in anxiety type that could help precisely identify triggers of speaking anxiety. Personalized guidance could then be developed to support the learning process. In addition to increasing the size of quantitative data, more qualitative data is needed to strengthen the results from the analysis. Since most participants returned to their home countries on the last day of the class, it was hard to collect the qualitative data as in-person post-task interviews were not possible. Even though the investigator was able to schedule remote interviews, slow Internet speeds and the time differences between countries made the process more complex. In addition, the participants' limited English language skills including participants who identified as Deaf/Hard of Hearing who used American Sign Language (ASL) to communicate, hindered the participants from expressing their thoughts in English/ASL unless there was translation in a native language or interpretation by ASL interpreters. In terms of future work, data from other physiological responses such as heart rate and skin temperature can be identified. This data can be studied to determine if features can help to detect anxiety with higher levels of accuracy.

To address these limitations, administrative cooperation from the school and language institute will be required to conduct further quantitative and qualitative studies with the ELLs with low English proficiency skills in future studies.

3.3.5 Summary

The studies described in this have led in part to two contributions described in Chapter 1. Study 1 was an exploratory study to determine the feasibility of using EDA features and exploring nonverbal behaviors to classify the ELL's degree of anxiety into higher and lower anxiety groups. The analysis revealed that the number of peaks and standard deviation of EDA can be prominent features to distinguish the ELLs between high- and low-level anxiety groups. Most foreign language instructors could benefit from using this physiological data to objectively and comparatively measure the levels of anxiety across the ELLs. On the other hand, Study 2 delved into the interview responses from the ELLs in parallel with analyzing extended EDA features and computed non-verbal behaviors. This enabled the investigator to identify the predominant subtypes of speaking anxiety that ELLs may be struggling with during the oral presentation. No anxiety (N), PSA more, (P), FLA more (F), and Both PSA and FLA (B) were explored, and a GBT algorithm with the top three input features was found as a well-performed algorithm to predict the predominant anxiety among four types. These findings lay the foundation for differentiating speaking anxiety into subtypes of PSA and FLA, which may assist EL instructors to apply it to provide differentiated emotional and instructional support to ELLs.

The findings also opened up the possibility for the investigator to develop an intervention for English language instructors by using biosensor-based feedback to target teaching. The intervention can supplement the instructors' ability to perceive the emotional states of learners, as it provides objective and comparative physiological information across students, in addition to behavioral observations. This can be used as additional context to understand the learners' learning process and provide emotional and instructional scaffolding in a timely manner. Thus, the investigator conducted further studies. First, Chapter 4 examined feasibility of the intervention system through a lens of instructors from an ELI in the United States and encouraged them to ideate how the system can be designed for ESL/EFL environments. Participant perspectives in focus groups and hand-drawn sketches of the intervention prototype system analyzed in Chapter 4 aided development of a prototype intervention system in Chapter 5. The system enabled users to examine the physiological states of the ELLs when presenting in class. This intervention system was used in Study 4 (Chapter 6) to determine whether EL instructors value gaining the support from the intervention system to identify highly anxious ELLs who are susceptible to PSA and FLA. This was achieved through conducting individual interviews and comparison evaluations with/without the intervention system with the instructors.

Chapter 4. Using Participatory Design to Develop an Interface to Support Instruction

4.1 Introduction

In this chapter, a design approach is described. Participants discussed the ways in which biofeedback could support teaching, and developed prototypes to assist instructors in classroom settings. The design approach adopted in Chapter 4 (and continued in Chapter 5 and 6) can be used by instructors and researchers as a first step to developing and evaluating technologies to determine levels of anxiety among students when teaching, with a view to tailoring instruction based upon levels of anxiety detected.

4.2 Background and Motivation

In an attempt to ease international students into an environment different from the one in their country of origin, educational institutions have provided various supports including new student orientation, counseling, workshops, tutoring, and supplemental instruction. However, prior studies reported that international students are less likely to use the services provided by their institution, including counseling services, when compared to their domestic counterparts (Ebert et al., 2019; Tang et al., 2012). This is in part attributed to concerns over their limited levels of English proficiency to communicate, and take advantage of services, the stigma associated with interacting with these services, and a more general sense of embarrassment (Gulliver et al., 2010). Also, Dewaele and Pavlenko (2002) found that ELLs used fewer emotion-related words compared to native speakers. This posed a challenge for English language educators to perceive the emotional states of

ELLs. Although native English-speaking teachers (NESTs) in the ESL/EFL setting have attempted to develop and promote Intercultural Communication Competence (ICC), which is the ability to acknowledge differences in a new culture and communicate effectively and appropriately with ELLs from various cultural contexts (Beamer, 1992), the limitation in understanding students' learning difficulties is still reported in prior studies (Shim, 2019; Shim & Shur, 2018). One of the reasons is due to the lack of shared cultural contexts between NESTs and ELLs (Ma, 2012; Mairi et al., 2018).

Sensing technology to detect physiological information has been proposed as an innovative method to avoid the ambiguous and subjective interpretation of emotions (Di Lascio et al., 2018). As it also unobtrusively monitors the physiological changes across learners in real-time, researchers have used sensors to non-invasively gather information relating to emotions from learners (Qu et al., 2020; Umematsu et al., 2019).

Despite the potential benefits of using biosensing technology, it has yet to be extensively applied and evaluated within an ESL/EFL environment, where levels of anxiety can run high when performing speaking tasks in a non-native language (English). A further challenge relates to the paucity of research relating to identifying perceptions of biosensing technology among ESL/EFL educators, and the ways in which this technology can be incorporated into their teaching practices (Farley et al., 2018). In fact, instructors need to be data literate to effectively utilize biosensing technologies, to ensure that they can understand what the physiological information collected does and doesn't tell them about their students, and how to use this data to support their teaching (Beck & Nunnaley, 2020; Mandinach et al., 2015). In order to address ways to design a solution to aid teaching, this chapter focuses on the five stages of the Design Thinking Process (DTP) (Hasso

Plattner Institute of Design at Stanford) – Empathize, Define (the problem), Ideate, Prototype, and Test (Plattner et al., 2009), in order to understand instructors' needs, define problems in ESL/EFL environment, let the instructors create possible ideas, strengthen their ideas through a prototype sketching, and test it.

To address the problems of subjective interpretation of observation explored in Chapter 2, the study described in this chapter explored how biosensing technology could be integrated to support teaching. In this chapter, the first three stages of the DTP were conducted (e.g., Empathize, Define and Ideate) through a professional development workshop with nine in-service stakeholders at the ELI. The findings from this process led to the development of a prototype system, which was pilot-tested with six in-service instructors and three ELLs in Chapter 5. In the final stage of the DTP, the prototype system was tested with 17 ESL/EFL instructors (described in Chapter 6). Insights from the study are discussed in Chapter 7.

The study described in Chapter 4 began with asking instructors to share their teaching experiences (e.g., pedagogical beliefs, teaching methods, standards of assessment, interaction with students, etc.) through a focus group to better understand and empathize with problems that ESL/EFL instructors encountered when teaching ELLs. Listening to the instructors' personal teaching experiences prior to designing a system enabled the investigator to set aside her own assumptions that the sensing technology should be designed and used in certain ways. If the prototype system was designed without understanding the actual users' needs and the environment in which the technology would be used, this would impact adoption. The discussion among the instructors during the focus group helped the investigator gather information to define the problems discussed in

Chapter 2, and to identify ways to address the problems using biosensing technology. The investigator introduced the biosensing technology with the intention to gain their perspectives on an unexplored technology that could provide solutions to challenges experienced determining levels of anxiety among students.

In addition to the focus group, the investigator employed a design charrette (design sketching activity), where members of a team quickly sketch designs (Roggema, 2013) to explore ways of designing an interface and integrating functionality. Six in-service English language instructors, two tutors, and the director of the ELI were invited to participate in this study. All participants had experiences teaching ELLs. Findings from the work have aimed to address the following sub-research questions in this chapter.

Study 3

- RQ 3-1: Which traditional methods have EL instructors used to identify the emotional states of learners in class?
- RQ 3-2: How did/would the awareness of learners' emotional states impact instructors' teaching?
- RQ 3-3: How would the experienced EL instructors integrate biosensor-based feedback of learners' speaking anxiety into a prototype system?

4.3 Method

4.3.1 Participants

Participants were recruited with assistance from the ELI at UMBC. The Director of the ELI invited the investigator to an annual Professional Development (PD) workshop to describe her research to attendees as she had already conducted prior studies with students at the ELI (Chapter 3). Attendees included six in-service English language (EL) instructors, two tutors, and a director who had English language teaching experience. The PD workshop usually enables ELI stakeholders to discuss issues relating to the ESL/EFL learning environment, and offers an opportunity for attendees to strengthen awareness of pedagogical techniques, and develop skills to support student achievement. The investigator had four years of teaching experience in the ELI as an EL instructor, so there were already pre-established relationships with the attendees. This facilitated a comfortable environment where all interested parties were able to freely discuss topics during the studies (Roulston, 2014; Stewart et al., 2007).

Table 6 Demographic information of participants in PD workshop

ID	Age	Gen-der	Teaching Year	Teaching country (Year)	Language other than English
P1	47	F	27	US (20), Hungary (7)	Hungarian, German, Russian
P2	40	F	10	US (10)	Spanish
P3	28	M	2.5	US (1), Burkina Faso (1), France (0.5)	French, Spanish, Moore
P4	35	M	2	US (2), Korean (2)	Spanish, Korean
P5	43	F	3	US (3)	Korean
P6	32	M	0.3	US (3 months)	Japanese, Chinese, Korean
P7	41	F	20	US (10), Madagascar (13), India (0.5)	French, Malagasy
P8	64	F	10	US (10)	Chinese
P9	56	M	16	US (16)	French, Spanish

Demographic information is presented in Table 6. Nine stakeholders (male 4, female 5) participated in the PD workshop in November 2019 (referred to as P1 to P9). Half of the participants had an M.A. degree or a teaching certification of Teaching English to Speakers of Other Language (TESOL) and all participants had experience teaching ELLs in a range of countries.

The PD workshop was conducted prior to the lockdown associated with the COVID-19 pandemic. The data was collected in a classroom on campus. The classroom was equipped with a document camera with a beam projector that displays an object such as a textbook or handout materials to an audience. The investigator installed two audio-video recorders in the classroom to record the participants' responses to the questions in the focus group and their presentation of hand-drawn prototype sketches (artifacts) during the PD workshop. The workshop lasted for a period of two and a half hours.

Participants attended the PD workshop to learn about the topic and to discuss. The investigator took this opportunity to inform the ELI stakeholders of the findings from her prior studies relating to biosensor technologies and described the intended aim for use within a classroom setting. Participants were asked to discuss and critique the use of proposed biosensing technologies for the classroom, taking into account the practicality and fidelity of the technologies. The steps undertaken for the workshop are presented in Figure 7. This study was approved by the Institutional Review Board (IRB) at UMBC.

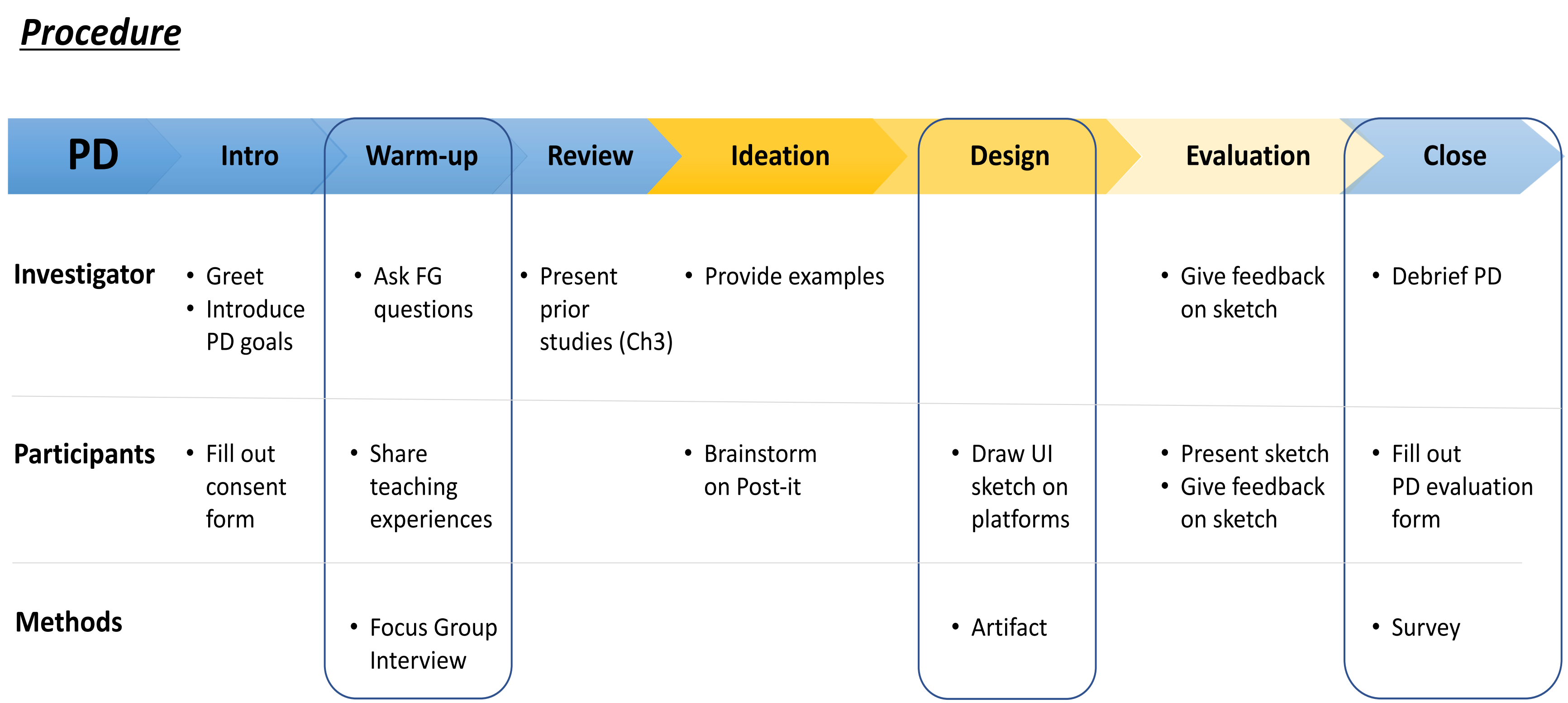


Figure 7 Procedure of Study 3: Professional Development (PD) workshop and methods used in data collection.

4.3.2 Procedure

At the beginning of the workshop, the investigator described her area of research (Appendix C), and participants were asked to complete the informed consent form. After collecting the consent forms, the investigator described the three objectives of the workshop:

- Sharing experiences of approaching anxious students in class.
- Reviewing three studies conducted by the investigator and considering the practical implications arising from these studies from an educational perspective.
- Sketching affective educational interfaces integrated with sensing technology to support anxious ELLs in speaking classes.

The investigator conducted a focus group with ELI stakeholders by posing questions in the scenario presented below:

If there was a student who appeared anxious in speaking in class (a scenario),

- FG Q1. Could you share your experience or how would you feel?
- FG Q2. How do/did you identify students who seem worried about speaking in front of others? What do/did you think is the main source of their anxiety?
    - Public Speaking Anxiety (PSA)? Foreign Language Anxiety (FLA)? Or others?
- FG Q3. What teaching methods or strategies have you used to help ELLs to reduce/address their fears/anxiety and improve speaking performance?

The focus group was selected to encourage the ELI participants to discuss issues relating to determining the emotional state of students in a classroom setting, suggesting ideas for designing solutions, and building upon ideas of others who have knowledge and teaching experience of the topic. During the group discussion, the ELI stakeholders shared their perspectives, listened to the views of others, and offered feedback. Discussion generated through focus groups was thought to lead to the development of ideas which may not have been otherwise identified through individual interviews.

After conducting the focus group, the investigator introduced her prior studies (Appendix C) to the ELI stakeholders. The findings from those studies informed the participants of the potential benefits when accounting for anxiety and the predominant subtype of anxiety from ELLs. When describing findings from the prior studies, the investigator asked the participants to consider how the findings could be applied to their own classes and teaching.

After taking a 10-minute break, the investigator reminded the participants of the last objective of the PD workshop; sketching affective educational interfaces that integrated with sensing technology to support anxious ELI students in classes where presentations would be delivered. The interface would be able to present details of learners' emotional states, which would offer references to instructors to be able to identify anxious students and approach them with evidence-based emotional and instructional scaffolding. The investigator handed sticky notes out to the participants and asked them to jot down one or two ideas for a possible educational technology using the information relating to emotional states from learners. Fifteen minutes were allocated for this brainstorming task. After that, the sticky notes were affixed to a whiteboard, and the investigator asked each participant to share their ideas with the audience.

To help the participants convey their ideas to design a prototype, the investigator provided examples of user interfaces from other studies (Fung et al., 2015; Takac et al., 2019; Tanveer et al., 2016) and a commercial application (Equity Maps) (Appendix C). The examples included visualizations of classroom activity and real-time feedback on a presenter's performance when delivering a talk. Then the investigator provided participants with four paper-based screen layout artifacts (i.e., paper-based representations of a desktop screen, tablet screen, mobile phone screen, smartwatch screen) as shown in Figure 8. Participants were asked to illustrate their ideas using the artifact of their choice. The use of hand-drawn sketches and low-fidelity prototyping has been widely used by researchers, as these methods offer several advantages in the early stage of product development (Buxton, 2007; Rudd et al., 1996; Sefelin et al., 2003). For example, hand-drawn sketches can be powerful and persuasive representations of ideas, sequences, systems, and objects (Baskinger, 2008). Designers usually explore the sketches drawn by participants to evaluate the needs of users at the initial phase of design processing (De et al., 2013). Thus, instructors were asked to select the platform of a paper-based prototype design that they thought would be most valuable in a classroom setting (Figure 8).

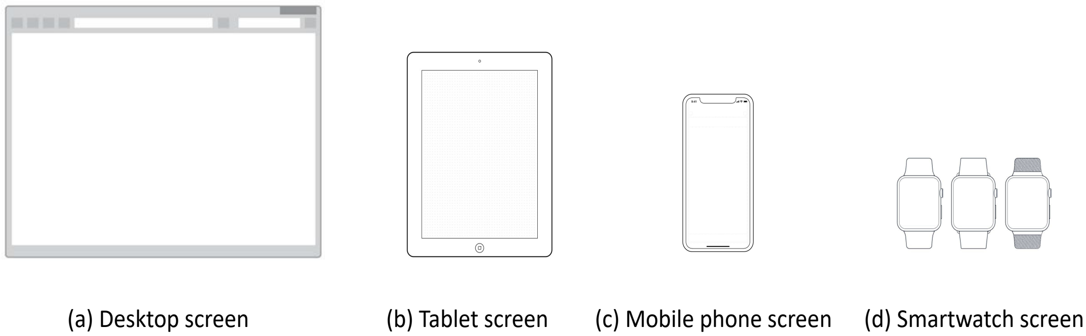


Figure 8 Four paper-based platforms used for hand-drawn sketches for a prototype system

For those who struggled with the process of design, the investigator presented an example of a hand-drawn sketch of an iPad user interface on a PowerPoint slide (Appendix C). After 30 minutes of hands-on sketching prototyping, the investigator projected each sketch on the screen, and asked each participant to explain their design and rationale behind it. Additionally, the investigator asked the rest of the participants to provide feedback. Evaluation forms were then handed out to the participants (Appendix D).

4.4 Data Analysis

The steps involved in the protocol relate to the process of user-centered design described by Norman & Draper (1986), focusing on the instructors' needs in the design process to develop highly usable and supportive educational systems which they could use. Teaching experiences shared by instructors offered a deeper understanding of the context of how they may use a system for supporting ELLs in class. Insights were also gained regarding how instruction can be personalized based upon detecting levels of anxiety among students. The investigator analyzed content from the focus group, along with the instructors' hand-drawn sketches.

4.4.1 Focus Group

The investigator used thematic analysis, as described by Braun & Clarke (2006) to analyze transcripts from the focus group (Joffe, 2012; Neale, 2016). Figure 9 shows the entire coding process of the data gathered from the focus group, which addressed the first two research questions of this study:

- RQ 3-1: Which traditional methods have EL instructors used to identify the emotional states of learners in class?
- RQ 3-2: How did/would the awareness of learners' emotional states impact instructors' teaching?

Pre-processing

The recording of the entire PD workshop was automatically transcribed using an online transcription service (https://sonix.ai). This service converted speech to text and provided a time-stamped transcript. The investigator read through the auto-transcribed texts while listening to the audio recording to verify their accuracy. Instances of laughter, sighs, pauses, and filler words such as um, well, uh, etc. from the workshop were also included in the transcript. The investigator used transcripts to infer the context around statements that were potentially important to understand the instructors' motivation or frustration when having anxious ELLs in class.

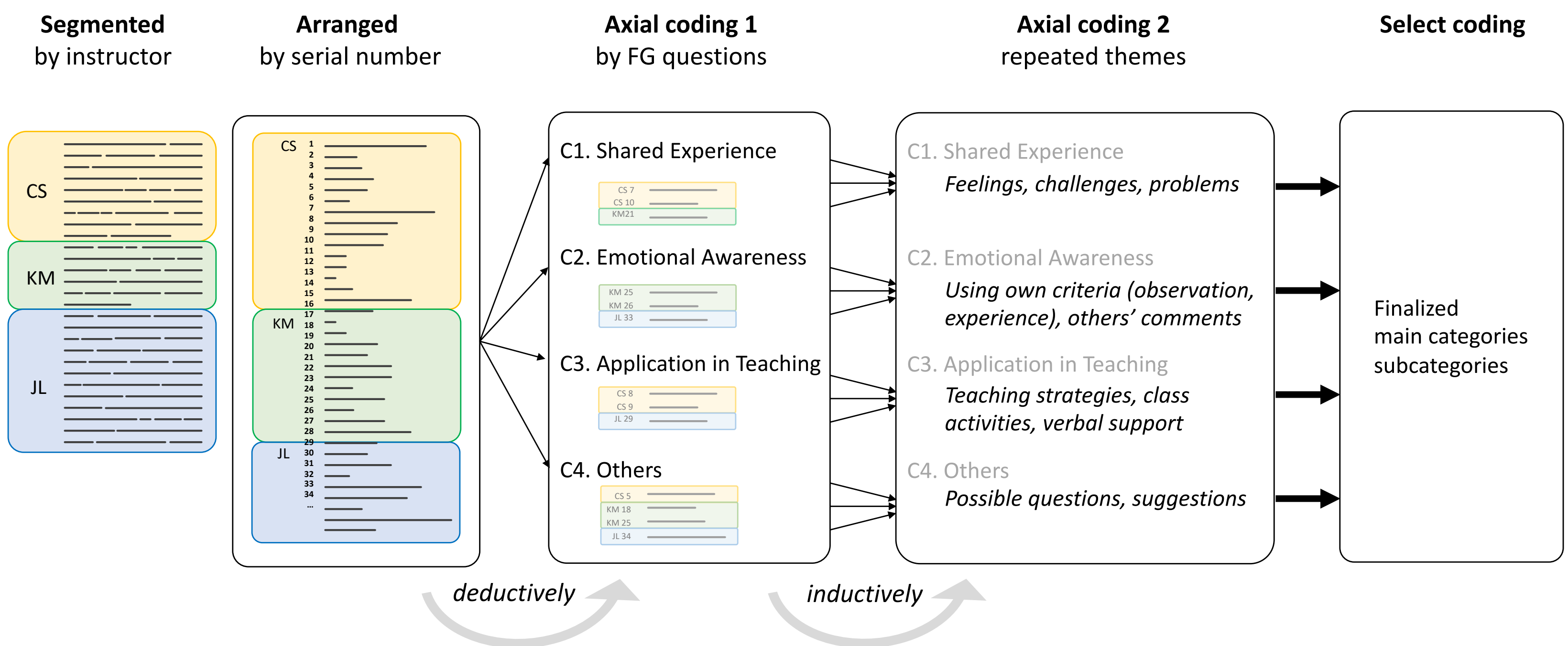


Figure 9 Entire coding process of focus group interview

The investigator divided transcripts by each participant (Figure 9. *Segmented*) in the order in which it was delivered. Each segment relating to a statement made was assigned a serial number (e.g. JL1, JL2, JL3) (Figure 9. *Arranged*) and entered into a spreadsheet. This was used for the purpose of conducting an inter-rater reliability (IRR) test (Richards, 1999; Welsh, 2002) with two other trained coders. Fifty-five transcribed strips were examined in total. The investigator and each coder independently categorized segments into four categories, which correspond to the focus group interview questions (FQ1 to FQ4). The entire development of a coding scheme is outlined in Figure 9.

Axial Coding 1

Transcript codes were categorized by questions posed to the focus group. The same transcript strip could apply to more than one category. The strips not relating to questions posed in the focus group were categorized into Category 4.

- Category 1: Shared Experience

  FQ1. If there was a student who appeared anxious in speaking class, could you share your experience? How would you feel?

- Category 2: Emotional Awareness

  FQ2. How do/did you identify students who seem worried about speaking in front of others? What do/did you think is the main source of their anxiety? Public Speaking Anxiety (PSA)? Foreign Language Anxiety (FLA)? Or others?

- Category 3: Application to Teaching

  FQ3. What teaching methods or strategies have you used to help ELLs reduce/address their fears/anxiety and improve speaking performance?

- Category 4: Other questions (FQ4) that do not fit in Categories 1 to 3

Inter-rater Agreement (IRA) Test

After the coding process was completed, as shown in Figure 10, the percentage of absolute agreement (Altman, 1991; Chaturvedi & Shweta, 2015) was calculated to examine agreement among coding by investigator and two coders regarding the grouping of units of transcript into four categories (Category 1 to Category 4). This measure is known to be easier to interpret and calculate when the number of raters is small, and rating levels are less than 5-7 (Chaturvedi & Shweta, 2015). The serial numbers attached to each coded snippet in pre-processing were used to run the inter-rater reliability test. The number of concordant serial numbers was divided by the total number of serial numbers. An absolute agreement level close to 75%-90% demonstrates an acceptable level of agreement of IRA (Stemler, 2004). When the test result was below 75%, the investigator and each coder met again in pairs, and reviewed the definitions of terms and discussed the rationale used in their axial coding process. After that, the process of axial coding was repeated, by reflecting on what participants discussed during the meeting in order to reach above the 75% threshold.

$$\text{Percent agreement} = \frac{\text{Number of concordant responses}}{\text{Total number of responses}} \times 100$$

Figure 10 Formula of the percentage of absolute agreement

Axial Coding 2

Upon achieving a high level of inter-rater reliability, the two coders and investigator highlighted key words or statements of each coded snippet for the four categories described above. After highlighting the important statements, the transcribed strips in each category were further subdivided (*Axial coding 2*) to identify recurring ideas from each category. The subcategories from each category were:

- Category 1: Shared Experience
    - Subcategory 1: Feelings of instructors and problems or challenges of teaching experience described by instructors.
- Category 2: Emotional Awareness
    - Subcategory 2: Using own criteria (i.e., observation, past experiences) or others (i.e., being informed by other co-workers, American Sign Language (ASL) interpreters, student service administrators, etc.) to perceive and identify the emotional states of students.
- Category 3: Application to Teaching
    - Subcategory 3: Teaching strategies, class activities, and verbal support.
- Category 4: Others
    - Subcategory 4: Further questions and comments discussed in the focus group.

The transcript in the first category was a case of how instructors feel having an anxious student in their class. The investigator and two coders investigated emotional expressions (e.g., frustrated, surprised, hesitated, etc.) that instructors used to describe their teaching experiences with an anxious student and the impact resulting from this. In the second category, the transcript was investigated to determine whether instructors actively (e.g., direct observation) or passively (e.g., report by student services department or co-workers) identify anxious students in class. To review how the instructors actively perceive emotionally anxious students, the transcript describing the instructors' own personal criteria including observation of facial expressions and behaviors among the ELLs, and their prior teaching experience with former students exhibiting anxiety were examined. In contrast to the active way to identify anxious students, other informative resources from another instructor, a member of staff from the ELI, and an ASL interpreter, or a student services department who may have interacted with anxious students were explored as a passive way that the instructors identify anxious students in class. The third category was grouped by how instructors have modified instruction relating to identification of learners' anxiety. In this subcategory, the teaching strategies were classified into teaching methods, class activities or verbal support provided by the instructors. Finally, a fourth subcategory was investigated to see if there was any emergent subcategory in addition to the three subcategories already identified (termed: 'others'). The value of other instances and questions was highlighted by (Guba & Lincoln, 1981) to better understand the perspectives of participants.

Select Coding

Once each entire transcript was categorized, three coders compared their transcript strips with one another to identify similarities. When two out of the three coders identified themes in the same subcategory within the transcript, it was classified as an important instance, as the frequency indicates a significant dimension of common ground (Merriam & Tisdell, 2015). Aside from those selected transcripts, the rest of the transcripts were examined to understand the context where instructors encounter anxious ELLs in class. The questions and comments raised by the instructors were reviewed. Through this table, the investigator and two coders reviewed whether all categories and subcategories were matched based on each category and checked if these met the level of theoretical saturation needed to address the first two research questions of this study:

- RQ 3-1: Which traditional methods have EL instructors used to identify the emotional states of learners in class?
- RQ 3-2: How did/would the awareness of learners' emotional states impact instructors' teaching?

4.4.2 Artifact – User-generated Sketches of User Interface

The investigator adapted five filtering dimensions (Table 7) to examine the core aspects of a design idea in interactive systems design (Lim et al., 2008). Hand-drawn sketches are usually adapted in the early stages of final design to gain meaningful knowledge and design ideas from users prior to investing funds into developing actual designs (De et al., 2013). With hand-drawn sketches, researchers have taken advantage of using filtering dimensions to screen out unnecessary aspects of the design and envision the designers' goal explicitly (Berger et al., 2015; Dove et al., 2016). Thus, the investigator applied a set of filtering dimensions to the artifacts to explore design ideas from ELI stakeholders, along with their rationale for these ideas.

Pre-processing

The investigator asked the participants to present their ideas and rationale behind the designs developed, and encouraged the audience to offer feedback on designs. Later, the investigator and two coders compared transcripts with the hand-drawn user interface sketches to infer further details or hidden functions of each sketch.

Table 7 Example variables of each filtering dimension adapted by (Lim et al., 2008)

Filtering	Example variables
Appearance	Platform/device, size, color, shape, margin, etc.
Data	Data type (number, string, media), data use, organization, hierarchy
Functionality	Functional components, system function, users' functionality need (buttons, menu bar, dropdown boxes, text-field)
Interactivity	Input behavior (click, scroll), output behavior, feedback behavior
Spatial structure	Arrangement of interface or information elements (frame division)

Filtering Dimensions

The filtering dimensions (Table 7) served as the standard for classifying design components of the artifacts and transcripts created in the pre-processing stage. The investigator and two coders completed a form (Appendix E) that is identical to Table 7. This form was used as a filter for an in-depth investigation into the UI design components and user interaction expressed by each participant's low-fidelity design sketch. This ended up being used as an initial outline to develop an interactive system design (see Chapter 5).

For the *appearance dimension*, the investigator and two coders identified the platform/device was selected by each participant to deliver their design ideas. As the type of platform/device constrains space to visualize the amount of information presented on the screen, it was critical to explore how the participants condensed their designs into the limited space based on the platform they selected. The size of components or labels drawn on the sketch were not examined because they could have been influenced by personal drawing skills.

In the analysis of the *data dimension*, the investigator and two coders highlighted the components of data input or data visualization in each artifact. In addition to the artifact, the transcripts that describe data usage or delivery to the system were explored, and attached to the form. In particular, the investigator and two coders identified what type of data needs to be entered using those components. These examples can serve as a reference for the information architecture relating to a design, as it allows researchers and system designers to estimate the size of data, the amount of visible data presented, and the ways of labeling and naming components (Lim et al., 2008).

To analyze the *functionality* dimension in the participants user interface sketches, the design components, particularly related to the user interface (such as clickable buttons with icons, menus, and dropdown boxes) were investigated. These components were screen captured and attached to the form (Appendix E). The participants' explanation (shown in the transcript) was examined at the same time as participants were unable to fully illustrate the features hidden behind a button, icon, or link in their sketches. As participants verbally described extended interactions after clicking a component on a sketch, the investigator and two coders referred to the transcripts accordingly.

For the fourth filtering dimension, *Interactivity*, the investigator and two coders examined how the participants described interacting with their user interface prototype. The actional verbs to describe input behaviors, output behaviors, feedback behaviors, and operation behaviors were frequently mentioned by participants. By looking through the transcripts, statements describing these behaviors were highlighted and attached to the form. This analysis was helpful to understand the causal relationship between the design components and the interaction that users expect.

In the final filtering dimension, *Spatial structure*, the arrangement of the interface used in each artifact was investigated. Basically, the main focus for analysis was to investigate how the space of a platform/device was divided and framed. The investigator and two coders drew a simple wireframe on the form to show how each participant laid out their interface.

The investigator addressed the final research question in this section and the objective of this study, which was to co-design a system with experienced instructors, sharing their teaching experiences in class, to integrate a notification informing levels of anxiety experienced by students in class.

- RQ 3-3: How would the EL instructors integrate biosensor-based feedback of learners' speaking anxiety into a prototype system?

Inter-rater Agreement Test

After the investigator and two coders filled out the form (Appendix E), the percentage of absolute agreement (Chaturvedi & Shweta, 2015) was calculated. Instead of using an axial coding scheme, each design component and highlighted statements on the transcript were categorized by five filter dimensions. These were used to run the reliability test. The number of concordant design components and transcripts were manually counted, and it was divided by the total number of design components and transcripts.

Once a finding of 75% or higher was returned from the inter-rater agreement test, the investigator adapted the design components and transcripts to form a middle ground prototyping option to reflect both perspectives from the end-users and investigator in a similar way to the study described by Michael (2012). These design components were applied to the initial prototyping stage in Chapter 5 when the investigator developed a clickable high fidelity prototype system that informed instructors regarding the emotional states of learners in class.

Feedback was gained from six instructors and three ELLs in a pilot study (Chapter 5) after interacting with the prototype. The aim was to determine whether the prototype still needed to be refined in terms of supporting instructors to gauge the level of ELLs' speaking anxiety. As a result of their feedback, the design components that caused confusion among the participants were excluded. For instance, the commentary functions were removed to reduce confusion by the user. The high-fidelity prototype was developed by applying the findings from Chapter 5. It was then tested with 17 EL instructors ranging from novice to experienced instructors in Chapter 6. The study focused on exploring the efficacy of using the biosensor-based system in perceiving the emotional states of ELLs. Participants also discussed the applicability of the biosensing technology in the context of ESL/EFL teaching and learning in Chapter 6.

4.5 Findings

A summary of the inter-rater agreement (IRA) values is presented in Table 8. By comparing the coding work between the investigator and each coder in pairs (A x B and A x C), findings of 50.45% and 46.79% (average 48.62%) were identified for the focus group. Findings of 62.45% and 63.35% (average 62.90%) were identified for artifacts. After discussion with the coders regarding the definition of each coding scheme and the process of coding work, the IRA values from the second set of coding were 83.75% and 75.15% (average 79.45%) for the focus group, and 82.41% and 87.12% (average 84.77%) respectively. As findings from the IRA in the second set of coding were above 75%, further IRA testing was not needed.

Table 8 The summary of the inter-rater agreement (IRA) values between investigator (A) and each coder (B and C) for focus group (a) and artifacts (b).

A x B	**1st coding**		**2nd coding**	
Theme	# of strips in agreement	IRA values	# of strips in agreement	IRA values
FQ1	1/8	12.50%	3/5	60.00%
FQ2	4/7	57.14%	6/8	75.00%
FQ3	4/7	57.14%	7/7	100.00%
FQ4	3/4	75.00%	4/4	100.00%
	average	50.45%	average	83.75%

A x C	**1st coding**		**2nd coding**	
Theme	# of strips in agreement	IRA values	# of strips in agreement	IRA values
FQ1	4/7	57.14%	7/8	87.50%
FQ2	4/6	66.67%	5/7	71.43%
FQ3	3/10	30.00%	6/9	66.67%
FQ4	1/3	33.33%	¾	75.00%
	average	46.79%	Average	75.15%

(a) Summary of focus group IRA

A x B	**1st coding**		**2nd coding**	
ID	**# of design components**	**IRA values**	**# of design components**	**IRA values**
P2 (a)	10/19	52.63%	14/19	73.68%
P2 (b)	5/8	62.50%	7/8	87.50%
P3	10/14	71.43%	12/14	85.71%
P4	5/13	38.46%	10/13	76.92%
P5(a)	7/15	46.67%	12/15	80.00%
P5(b)	14/17	82.35%	16/17	94.12%
P7	5/8	62.50%	7/8	87.50%
P8(a)	5/8	62.50%	6/8	75.00%
P8(b)	9/11	81.82%	9/11	81.82%
P9	14/22	63.64%	18/22	81.82%
	average	62.45%	average	82.41%

(b) Summary of artifacts IRA

A x C	1st coding		2nd coding	
ID	# of design components	IRA values	# of design components	IRA values
P2 (a)	18/27	66.67%	25/27	92.59%
P2 (b)	5/7	71.43%	6/7	85.71%
P3	11/15	73.33%	14/15	93.33%
P4	10/18	55.56%	16/18	88.89%
P5(a)	11/16	68.75%	14/16	87.50%
P5(b)	10/15	66.67%	13/15	86.67%
P7	2/5	40.00%	5/6	83.33%
P8(a)	5/8	62.50%	7/8	87.50%
P8(b)	10/15	66.67%	12/15	80.00%
P9	13/21	61.90%	18/21	85.71%
	average	63.35%	average	87.12%

(b) Summary of artifacts IRA

4.5.1 Findings from the Focus Group

4.5.1.1 Challenges of identifying anxious ELLs

Although participants had been presented with a predefined scenario in the focus group, the instructors described their own experiences of having anxious students in their speaking classes. Terms such as ‘surprised’, ‘scary’, and ‘challenging’ were used when describing those anxious students in class because it was difficult to identify levels of anxiety through observation alone (e.g., *“it actually surprised me because she did not seem to have any kind of depressive… she didn't seem so depressed or anxious or…arrogant as she was described to me. I think what's scarier is that you… Even so, she did not exhibit any kind of. she was a super outgrowing center of the center of attention all the time. (P1)”, “It was so challenging… How do we differentiate between students in distress and students who just need a little bit of help? (P8)”*). One instructor (P1) shared that she did not notice an anxious student throughout the semester until the student requested accommodations for her anxiety through a third-party (student services department). Since the student was outgoing in the class and did not exhibit any explicit signs of anxiety, it was very surprising for the instructor, as the student's behavior contradicted what the instructor would ordinarily expect.

Other instructors (P8, P10, P11) expressed concerns regarding anxious students in their class because of their actions or body language that they would exhibit (e.g., shortness of breath, shaking, fidgeting), which could negatively affect their performance when presenting in class. The instances highlighted by the instructors appeared more serious compared to the pre-defined scenario.

The most common way that instructors identified anxious students in class was through observing their body language and prosody (P9-P12) including avoiding eye contact with the audience, fidgeting, grooming, or sighing, pause, stutters, or speed of speech. These behaviors were in line with prior studies (Daly et al., 1989; Gregersen et al., 2014; Park et al., 2014; Schneider et al., 2015). Moreover, anxious students were easily identified as they often would not volunteer to present first to the class (P8, P10). However, it was noticeable that the instructors felt it was challenging to identify anxious students in the beginning of the semester (P1, P11, P12) unless the student exhibited signs of anxiety or intentionally disclosed their feelings to the instructors. This in itself would pose difficulties, as the student would need to verbally express their anxiety, which can be challenging to do in a language unfamiliar to the student.

The responses of the instructors in the focus group supported the investigator's assumption that using biosensor-based feedback has potential to offer one datapoint to instructors to identify extremely anxious students in class. This can be helpful in cases where observation of anxiety in students may be challenging, as signs of emotions can be masked. Moreover, when the facial movements look the same to the naked eye (Barrett et al., 2019), the degree of anxiety between individuals cannot easily be perceived through observation.

4.5.1.2 A variety of teaching strategies to support anxious students

To help address periods of high anxiety, instructors described recommending that anxious students step out of the room to relax, spend time breathing deeply, drink water, and leave to use the bathroom (P9, P11, P12) (e.g., *"giving encouragement, letting them know it's okay. Well, maybe a breathing exercise like that. And that's what most students need. (P11)*". They also described providing encouraging comments to students (P3, P8 to P12) (e.g., *"I've had encouraging words for students' different time, but certainly not have thought about changing the way I teach it. I mean...I don't know. Maybe on a more micro level, individual students helping them individually. (P11)")*. The instructors described using a warm-up exercise at the beginning of class. Students pair up and discuss a small topic related to the lecture (P10 to P12) (e.g., *"you have like one student who go up and either you pick a topic for them or they pick a topic and they speak for a minute. The minutes over, stop. Then you give them 30 seconds, condense it, make it quicker. Thirty seconds. Stop. You have fifteen seconds. Tell us everything you just told us and they're done. So it's this idea that they have to think on their feet. They have to think that it's like they don't have time to be anxious, especially for 15 seconds. The idea was just getting them talking and getting them more comfortable talking about something. (P9)"*). Moreover, the instructors found it was advantageous to let the students select the order of who presents within the class (P1, P10) so that they can better prepare themselves prior to speaking (P1, P2, P12) (e.g., *"one time I had a class that was smaller and speaking and listening and I noticed that they did have a lot of anxiety. So I had them prepare just a brief presentation about their weekend and they presented it in their native language and then they presented it in English. (P2)"*).

Other instructors shared cases where students reacted differently based on the topics and the language used when presenting. P5 found it interesting that the topics of the presentation could influence the emotional states of the ELLs. If the topics are familiar with the ELLs' domestic culture or prior knowledge/experiences, the students seemed comfortable to speak, whereas discussing the history of the U.S. was quite challenging for them to describe confidently in English. In addition to the choice of speaking topic, P2 shared her teaching strategy to help identify and address levels of anxiety. She asked the students to give a brief presentation about their weekend in their native language and then present it in English. She stated that many different personalities came out when the students spoke in their native language. Students who were not confident when delivering a presentation in class (in English) suddenly became more confident.

P9 also had an interesting strategy to alleviate anxiety, gamifying delivery of the presentation:

"They speak for a minute. The minute is over. Stop. Then you give them 30 seconds, condense it, make it quicker. Thirty seconds. Stop. You have fifteen seconds. Tell us everything you just told us, and they're done. So, it's this idea that they have to think on their feet. They have to think that it's like they don't have time to be anxious, especially for

15 seconds. The idea was just getting them talking and getting them more comfortable talking about something." (P9)

Interestingly, P7 and P12 provided an extra opportunity for extremely anxious ELLs to deliver the presentation in class. Furthermore, anxious students were given an opportunity to practice their talks one-on-one with the instructor outside of the class. Participants described sitting with the students and building up a rapport with them. This in turn would make them feel more comfortable and help them become more accustomed to speaking in front of others.

4.5.2 Findings from the Artifacts

The sketches developed by participants were analyzed to examine the ways in which biosensing technologies could be integrated within an educational prototype to support anxious ELLs when delivering presentations. The examples of all hand-drawn sketches are attached in Appendix F.

4.5.2.1 Appearance Dimension

As shown in Table 9, the most frequently selected platform/device that the instructors selected to illustrate their ideas upon was a tablet screen (P1, P4, P5, P9). This was followed by a smartwatch screen (P3, P5, P6), a desktop screen (P2, P8) and a mobile screen (P6, P8). Two instructors (P2, P7) designed a prototype, which consisted of a physical headband that collects sweat on the forehead and an arm patch that exudes an aroma scent to calm down the students based on the level and type of emotions that each student has. Although most instructors chose one main platform to design their ideas, four instructors (P2, P5, P6, P8) selected a secondary platform or device (e.g., headband, watch, mobile phone) that should be used with the main platform.

The investigator and two coders classified the platforms depending on whether it was intended for instructors or students. If it was for instructors to evaluate and monitor students, larger screens (desktop or tablet) were selected. It was because the dashboard delivers a larger visualization of information, which would signal changes among the students. The participants who selected the mobile phone and smartwatch screens as platforms were planning to design an interface for students to use, enabling them to practice delivering presentations, where they could check their physiological state at a glance. As mobile/smartwatch platforms have the advantage of being portable, participants thought these platforms would offer value when directly tracking speaking performance or autonomic arousal from learners when delivering a presentation in real time.

Table 9 Summary of selective artifact coding themes.

Dimension	**Sub-category**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**TO-TAL**
Appearance	Desktop		X						X		2
	Tablet	X			X	X				X	4
	Mobile						X		X		2
	Watch			X		X	X				3
	Others (band, arm patch)		X					X			2
Data	media (video)	X	X		X	X				X	5
	prosody (filles, pause, pitch, pace)	X			X	X					3
	behaviors (eye contact, eye gaze, body language)	X	X								2
	grammar, vocabulary	X	X								2

(continuous)											
	Sub-category	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**TO-TAL**
Data	EDA (physiological arousal, anxiety)	X	X	X		X	X	X		X	7
	parts of presentation		X		X						2
	rubric/evaluation feedback		X	X	X					X	4
	size of audience	X				X					2
	set time (presentation time, response time)					X			X		2
	set student seat arrangement								X		1
	course #, title, logo	X	X		X						3
Functionality (UI components)	click buttons (review information)	X							X		2
	click buttons (play/stop/record) buttons	X	X		X				X		4
	color coding		X	X	X			X	X		5
	hyper-text link		X								1
	graph visualization (arousals)		X	X	X						3
	Slider (sync. video and physio. arousal)		X		X	X				X	4
	set different mode (T-S)				X					X	2
	pancake menu				X						1
	profile				X						1
	logo	X			X						2
	text-field box				X					X	2
	pin comments on timeline									X	1

Interactivity	**Sub-category**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**TO-TAL**
With media	upload	X	X								2
	record	X				X				X	3
	view		X			X			X	X	4
With information											
	track	X	X	X	X	X			X		6
	click to see	X									1
	check/select presentation					X					1
	control (size of audience, delete talk)	X				X			X		3
	evaluation (attach rubric, grade, progress)		X	X		X				X	4
	link teachers to students			X	X						2
	feedback (post, real-time, notes, comments)		X	X	X					X	4
Spatial Structure	horizontal		X		X	X			X		4
	vertical	X							X	X	3
	single one unified frame			X		X					2

4.5.2.2 Data Dimension

Although participants were not required to output the physiological changes in their diagrams, this information was presented visually across some artifacts. In cases where it was difficult to draw, participants were able to describe how it would be presented (P1 to P3, P5 to P7, P9). Three participants sketched how the level of arousal could be shown changing over time. Those sketches presenting physiological arousal were drawn with the intention to provide personalized instructions based upon the emotional states of the learners (P2, P3, P7). The remainder of the instructors (P1, P5, P6, P9) mentioned that physiological cues should be detected by a secondary device such as a wristband (P5), a smartwatch (P6), or an accessory such as a headband (P1). P9 stated that classmates can detect anxiety by monitoring the physiological data and send encouraging messages or signs (e.g., ‘thumbs-up’ or ‘like’) to the presenter. This could be achieved by a system detecting at which point arousal is most prominent, indicating high levels of anxiety or frustration.

In addition to the physiological data, behavioral data (e.g., eye contact with the audience) (P1, P2) and speech prosody data (e.g., pause, filler words) (P1, P4, P5) were illustrated by participants. Having these components would provide other indicators to help instructors assess the emotional state of students when delivering presentations. P2’s sketch contained details of how the prosodic data should be presented. Her illustration contained a progress bar highlighting pauses which anxious students may exhibit. These bars would be presented using different colors to help the instructor differentiate between students.

Since the majority of participants in this study were English language instructors (n=6) and tutors (n=2), aspects of language such as fluency (e.g., speech rate, continuity), accuracy (e.g., grammar, pronunciation) (P1, P2, P4) and evaluation (P2-P4, P9) were considered as important aspects to be stored in the system. The instructors and tutors wanted to use their own rubric chart for the performance evaluation. They described wanting to track the number and type of grammatical errors made over time to see if improvements were made, and check whether the students use new expressions or vocabulary learned within the lesson when delivering a presentation.

4.5.2.3 Functionality Dimension

One of the more common controls illustrated by participants was a slider (P2, P4, P5, P9). More specifically, the user would be able to view the synchronized audio-video stream of the presenter and compare it with the corresponding level of autonomic arousal. The slider would then be able to be used to identify instances within the presentation of interest along with the corresponding level of arousal (and vice-versa). In order to do this, behavioral and physiological data could be recorded in seconds to catch the detailed changes of behaviors and autonomic arousal among the presenters.

It was obvious that the functionality illustrated by participants either reduced in number or became simpler as the screen size of the platform reduced in size. For example, P4 who selected the tablet screen platform, drew a hamburger menu at the top left corner of the interface to save space on the screen. The menu contained links to the grading rubrics and discussion board. The participant also presented a profile icon at the top right corner of the interface, which if selected would open up to enter in user details. The aim of the clickable icon was to provide a way to reduce content on the main page. P1 also drew clickable square buttons next to each evaluation item (e.g., Fillers, Pauses, Eye contact, Grammar, etc.) and hid detailed information of the evaluation to make the space simpler.

The clickable square buttons functioned differently in P5's sketch. The buttons were presented as action buttons to record the presentation again or indicate the satisfaction of the performance outcomes. For P8's mobile phone screen artifact, the participant drew clickable square buttons labeled with an initial of each student's name. The layout of buttons resembled the seating plan of the students, enabling the instructor to quickly identify students experiencing high levels of anxiety, especially towards the beginning of the semester when they may have difficulties recognizing students by name.

4.5.2.4 Interactivity Dimension

The interaction with the user interface prototype was classified into two types based on the main user of the system (e.g., instructors vs. students). The frequent action verbs voiced by participants included ‘upload’, ‘record’, and ‘track’. The instructors wanted their students to ‘upload’ their recorded presentation performance on the system or ‘record’ their voice, or audio/video on the system in real time autonomously. They wanted to empower students to ‘track’ their emotional changes and behavioral mannerisms when delivering a presentation. They also wanted students to be able to see grammatical errors made while watching their presentation performance. This implies that the instructors wanted to encourage the students to be active learners. One unique interaction found among the artifacts related to the ability for students to communicate with each other (P9), rather than the ability for just instructors to communicate with students. According to P9’s artifact, the audience were able to view the recorded video of the presenter at a later point in time and pin their comments in the time sequence. This could be used by the presenter and the instructor to help work towards strengthening performance in subsequent presentations.

For the interface designed for the instructors, 43 of the action verbs voiced by participants related to characteristics of their instructional role (e.g., ‘assess/evaluate/grade’, ‘give feedback, encourage/reinforce’ and ‘track’). The participants who illustrated interfaces for instructional purposes expected an efficient grading environment where instructors could simultaneously use evaluation rubrics and automatically track the errors in grading. As participants wanted to limit conversations regarding low grades with students, they hoped that the system could provide objective information to the students to help justify their grades. In addition, they wanted to provide positive or negative feedback/comments on the students' performance next to the grade in the system.

4.5.2.5 Spatial Structure Dimension

Icons related to the recorded audio/video media were presented at the top (P2, P4, P9) or on the left side of the paper prototype (P1, P5). The reason behind presenting this component at this particular location was due to an understanding that individuals often start viewing a web page or interface from the top left side, which was consistent with Weinreich's study of web use (Weinreich et al., 2008).

4.6 Discussion

To seamlessly integrate biosensing technology with teaching practices in an ESL/EFL educational setting, a participatory-based design approach was conducted with the adoption of the DTP with ELI stakeholders. This study design encouraged the stakeholders to be actively involved in the development process of the biosensor-based system that can support teaching anxious ELLs and allowed the investigator to achieve the first three stages of DTP: (1) Empathize – understanding end-users' needs and challenges; (2) Define – pinpointing the end-user challenges that need to be solved; (3) Ideate – selecting solutions to prototype with end-users. The focus group interview was conducted to empathize and identify the challenges that the participants have experienced with anxious ELLs in class and the ELI. To help participants generate potential solutions to support ELLs, the investigator introduced a biosensing technology. By conducting an ideation session, participants could brainstorm ideas to tackle ways to support anxious ELLs in a speaking class. They developed a variety of designs where biosensing technology could be used. The findings from this chapter laid the groundwork for the studies described in the remaining chapters.

4.6.1 Potential areas for biosensing technology to intervene and assist English language instructors

In the focus group interview, some of the in-service instructors expressed surprise and concern that they could not detect the level of anxiety or frustration among ELLs, as not all students exhibited visible signs or behaviors. They described being surprised by the fact that the ELLs who they thought to be relaxed during class, were in fact experiencing high levels of anxiety (i.e., "polar opposite state"). The instructors found it difficult to determine the level of struggle with anxiety until expressed by the students or contacted by the Division of Student Affairs or Student Disability Services department for accommodations to be put into place.

Prior studies indicated that individuals can deliberately manipulate the external channels of expression (e.g., facial expressions, voice intonation), which can be the artifact of social masking (Wagner et al., 2005). This point is echoed by Beck (2008) that ELLs mask their incomprehension and control their body language and facial expressions to protect themselves from undue embarrassment. Masking may have been contributing to instructors not being able to recognize levels of anxiety.

Cultural differences between the instructors and ELLs may sometimes cause misunderstanding (Kamiya, 2019). The non-native English-speaking teachers (NNEST) can empathize with the learning difficulties of ELLs because they have the same linguistic or cultural background as the students (Ma, 2012). However, there is still a barrier to fully recognize the inward emotional states of ELLs when students' cultures were different from the cultures of NNEST.

Conducting the focus group enabled the investigator to closely listen to the ESL/EFL stakeholders' personal teaching experiences with anxious ELLs. Participants were able to discuss and empathize with each other's experiences. It ultimately helped to find the potential area where an interface using biosensing technology needs to identify instances of anxiety and notify instructors. The biosensor-based system would offer greater assistance to the instructors to determine emotional states of the ELLs, addressing the misinterpretation of social masking that the students create, and cultural gaps with the students. The biosensing data, especially EDA data can offset biased and subjective interpretation of behaviors. If the researcher does not share the same cultural background as the participants, the behavioral data can be misinterpreted. Although an indicator can be developed based on the frequency of behaviors that are manifested. The data still needs to be carefully interpreted based on the cultural traits of the students. On the other hand, EDA is a common physiological attribute of human beings, regardless of cultural differences. It allows the researcher to obtain the normalized degree of emotional states across the participants. This may provide benefits to the researchers and educators enabling them to objectively compare the intensity of emotional states across ELLs and to identify ELLs who need more attention and assistance than other learners in the classroom. As sociocultural factors and subjective interpretations come into play when analyzing the behaviors of ELLs, EDA data can offer a more objective method of determining the degree of anxiety among ELLs.

4.6.2 Specific requirements when developing a biosensor-based system for ESL/EFL teaching

In the design thinking session conducted during the PD workshop, participants presented ideas for developing a prototype system. These included sketches, as well as verbal descriptions of how a system could work. A noticeable finding was that the sketches were divided in two categories: one for instructors' usage and the other for both instructors' and students' usage.

For instructors' usage, the participants expected that the EDA data could automatically detect or evaluate the aspects of linguistic performance including fluency (pause, stutters), correct usage of grammar and parts of speech, and vocabulary. Participants wanted to support students by helping them better understand the reasons for why they earned the grades awarded for presentations. An interesting aspect commonly identified in participants' sketches was that the design components were related to a specific teaching environment, namely ESL/EFL teaching. For the participants, what content to evaluate (rubric content) and how to evaluate the delivery of presentation performance (operation) were considered as the important elements of the system. They expected that biosensing technology could automatically detect the learners' mistakes to use for purposes of grading. There was a tendency observed in the sketches that the participants wanted to avoid or minimize tedious and laborious work detecting the learners' mistakes by using the technology. Also, some of the participants were skeptical about the automated assessment that the technology may replace the instructors' role as an educator in providing actual grades on the learners' performance. This implies the extent to which technology should not invade the boundary of the educators' role (Selwyn, 2019). An interesting finding from instructors' sketches related to the focus on evaluating the linguistic aspects of the learners' performance with the integration of biosensing technology (e.g., accuracy of grammar and pronunciation, the usage of new vocabulary taught in class). This suggested that customization of the biosensing technology that could fit into the specific context of ESL/EFL teaching where new educational technology tools for ELL and their teachers is missed in contrast to mainstream classes and higher educational institutes (Andrei, 2017).

In contrast to the prototype design for instructors' usage, some participants designed systems for students to practice speaking skills. This would allow them to feel comfortable using English, by monitoring changes in EDA when delivering a presentation. The instructors wanted to glance at the interface of students' usage to understand how they use the system and improve their performance.

4.7 Limitations and Future Work

This chapter has described an approach to identify ways to design for an educational prototype. While the focus group enabled participants to discuss points and build upon each other's suggestions, it is important to be aware of the phenomenon of group-think where participants get locked into one way of thinking. Even though the investigator tried to prevent participants from singlehandedly dominating the discussion, some participants may have been hesitant to express their honest and personal opinions against the views of other participants. Thus, an individual interview should be conducted in addition to the focus group to draw further insights from participants. A further limitation was that the participants were not professional designers, so the sketches they produced were not as clear as anticipated. Although the participants verbally explained the design components of the interface presented on their sketches, some of design ideas including presenting hovering effects, feedback of action buttons or functional buttons (e.g., pulling out rubrics, commentary notes) must have been missed or hidden in the sketches, as it may have been difficult for them to draw what they had envisioned in their heads. Spending more time reviewing sketches with each participant and encouraging them to elaborate on their designs, would offer promise to this scenario.

While the approach adopted led to the development of designs and rationale for creating them, it is acknowledged that the sample size of nine instructors was small. As a result, findings may not be representative of all instructors in the ESL/EFL environment and outside of UMBC. Further focus group sessions with a wider set of instructors would be useful to conduct, until saturation is gained.

4.8 Summary

In this chapter, a study has been described to develop an educational system using EDA technology. To achieve these goals, the following research questions were addressed in this chapter.

- RQ 3-1: Which traditional methods have EL instructors used to identify the emotional states of learners in class?

- RQ 3-2: How did/would the awareness of learners' emotional states impact instructors' teaching?
- RQ 3-3: How would the experienced EL instructors integrate biosensor-based feedback of learners' speaking anxiety into a prototype system?

Findings from the focus group interview addressed the first two research questions (RQ 3-1 and RQ 3-2). It uncovered the needs of ESL/EFL instructors in terms of perceiving emotional states of ELLs in class. This guided the investigator to identify participants to ideate an innovative education system using biosensing technology (RQ 3-3), which could address the needs of the ESL/EFL instructors. The findings helped the investigator scope the design prior to developing an actual educational prototype system using biosensing technology (described in Chapter 5). The main contribution from this current chapter relates to the development of a design approach that can be used by instructors and researchers as a first step to developing and evaluating technologies to determine levels of anxiety among students when teaching, with a view to tailoring instruction based upon levels of anxiety detected. The approach itself is also continued in Chapter 5 and 6.

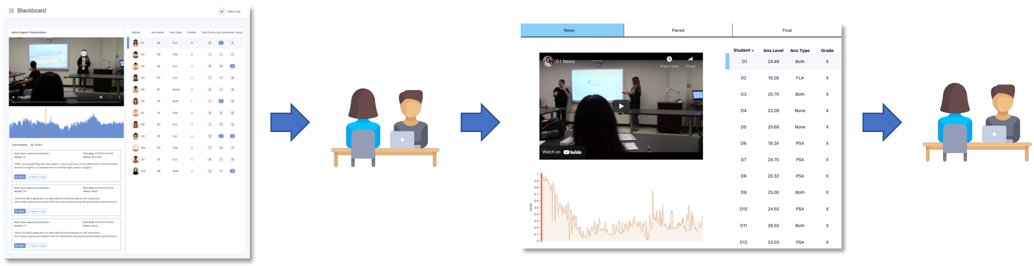
Chapter 5. Developing a Prototype System using Biosensor-based Information

5.1 Introduction

In this chapter, a prototyping phase is described. It was conducted with nine instructors from the ELI at UMBC. The prototype was intended to be used as part of an intervention (described in Chapter 6), with novice and experienced English language (EL) instructors to investigate the feasibility of using biosensing technology to support instructors when teaching ESL/EFL students. The prototype system was then evaluated in a pilot study with six EL instructors and three English language learners (ELLs). The contribution from this chapter (along with Chapters 4 and 6) is the development of an approach to design and evaluate educational prototypes using biofeedback. This can then be used by researchers and educators alike.

5.2 Process of Developing a Prototype System

As shown in Figure 11-a, the investigator began with designing a low fidelity prototype based on the instructors' design ideas (described in Chapter 4). Adobe XD was used to build the low fidelity prototype, as the tool offered a quick way of designing interfaces. The prototype could be easily refined based upon feedback from users.



(a) Low fidelity prototype　(b) Interview I (n = 9)　(c) High fidelity prototype　(d) Interview II (n = 6)

Figure 11 Process of developing a prototype system

In the initial low fidelity prototype (Figure 11-a), the audio-video recording of the student's presentation performance was located on the top left of the prototype interface. The EDA data relating to the presenter is presented beneath the audio-video recording. Under the graph relating to EDA arousal, the comments from the presenter, instructor, and classmates are presented. Each comment indicates the posting date and time, author, reading status of the post, and feedback on the presenter's presentation performance and emotional arousal. The most recent post is located at the top of the screen. However, the user can select the filter icon to sort the list by author and posting status whether the post was read/unread. On the right-hand side of the interface, the students of the class are listed in order of last name (ID). If the users select the header of the list including anxiety level ('Anx. Level'), anxiety type ('Anx. Type), and 'Grade', they can sort the order of the students based on the level of anxiety and grade. They can also classify the anxiety type into predominant public speaking anxiety (PSA), predominant foreign language anxiety (FLA), both PSA and FLA (Both), and no anxiety (None).

This low fidelity prototype was evaluated by six English language instructors (P10 to P15) and three ELLs (P16 to P18) in a pilot study (Figure 11-b). The aim was to gain feedback on its feasibility. As students were stakeholders of the system – one of the ELLs (P17) was the actual owner of the physiological data used in the prototype system, it was thought to be valuable gaining their viewpoint on instructors using physiological cues of students to support instruction. During the pilot study, both the instructors and students shared their perspectives regarding the idea of using the EDA data to support instruction and provided feedback on prototype designs. This feedback was applied, resulting in an interactive high-fidelity prototype (Figure 11-c).

The investigator and three external developers constructed the high-fidelity prototype system using Netlify. The platform offers hosting and server-less back-end services for web applications. Participants could access the system through an URL link (https://edu-project-8aa26e.netlify.app/) (Chapter 6), which enabled the investigator to deploy the prototype widely without the need for a complex set-up. To elicit more meaningful insights from the participants, the high-fidelity prototype system employed real EDA and audio-video data of the learners from one of the speaking classes in the ELI. The footage was gained from a class in the summer semester of 2019.

Using the high-fidelity prototype system (Figure 12), users could: 1) view footage of each student's presentation alongside viewing levels of physiological arousal (synchronized with the timeline of the audio-video recording); 2) click any point on the arousal graph to see the corresponding video footage; and 3) sort the list of students based on their name (ID), degree of anxiety, and type of anxiety. The degree of anxiety displayed the average EDA value throughout the presentation performance and the type of anxiety presented (PSA, FLA, Both, and None) based on each student's interview result conducted in Study 2. The investigator explained to each participant (P10-P15) how the pre-processed raw data (termed "raw data") and the information displayed on the system were processed, and how they could interact with each UI design components (e.g., tabs, a slider, sort list by column headers).

An interview was then conducted examining the experience with the redesigned high fidelity prototype system (Figure 11-d).

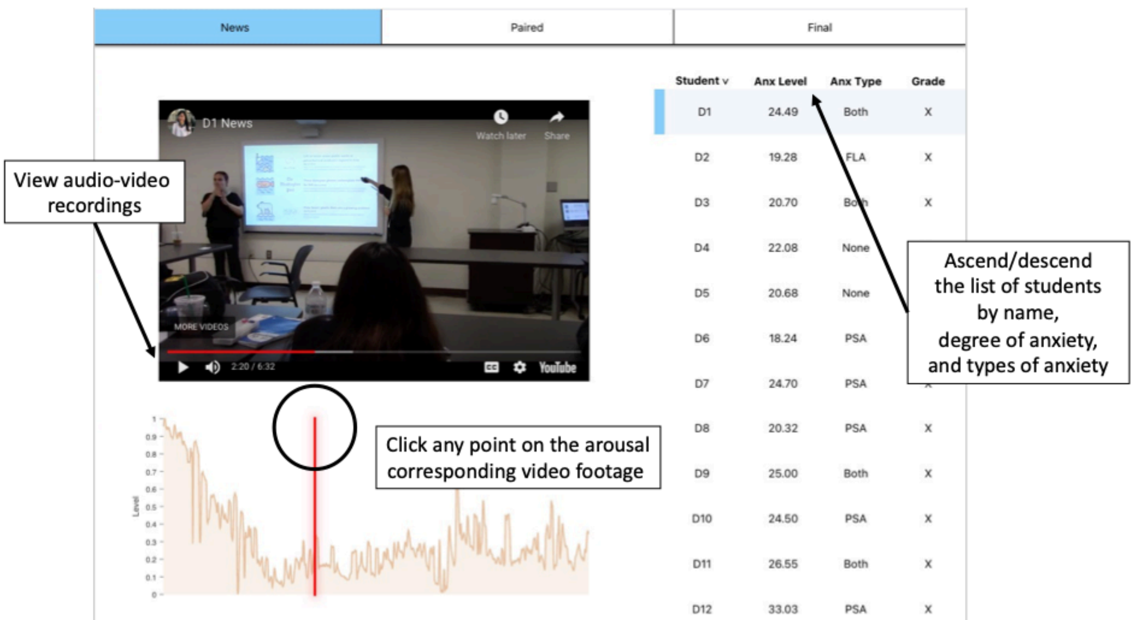


Figure 12 Interface of the high-fidelity prototype system

5.3 Pilot Study

5.3.1 Participants

Six instructors (1 male, 5 female) and three ELLs (1 male, 2 female) were recruited through the UMBC ELI liaison (Table 10). Most of the instructors had 10 years of teaching experience in either the U.S. or Asian countries. P16 was a graduate student in the Human-Centered Computing program and P17 and P18 were enrolled in the ELI at UMBC. The study was conducted remotely due to the social distancing mandates resulting from the COVID-19 pandemic.

Table 10 Demographic information of participants in pilot study

Instructors

ID	Age	Gender	Teaching Year	Teaching Country (Teaching year)	Language other than English
P10	34	M	9.5	US (2.5), Japan (7)	Japanese
P11	35	F	9	US (7), China (2)	Korean, Chinese
P12	30	F	8	US (4), Korea (2), Japan (2)	Spanish, Korean, Japanese
P13	34	F	10	US (8), China (1), Guatemala (1)	Chinese, Spanish, French
P14	32	F	7	US (4), Japan (3)	Japanese
P15	42	F	1.5	US (1.5)	Korean

English Language Learners

ID	Age	Gender	Year living in the U.S.	Language other than English
P16	33	F	9	Arabic
P17	22	F	1	Japanese
P18	42	M	5	Korean

5.3.2 Interviews

Interviews were conducted via an online meeting application. Interviews lasted 60 minutes in duration and were recorded. The investigator presented the prototypes to each participant. While the participants were looking at the prototype, the investigator asked an open-ended question such as “What do you like or dislike about the prototype design?” To explore the general perspectives of using the physiological data of students in an ESL/EFL class, the following questions were asked to both EL instructors and ELLs during the interview: “What do you think of using EDA data to understand students’ emotional states during the speech? Do you see any advantages or disadvantages of using it?”. With the high-fidelity prototype, which was developed by the feedback on the low fidelity prototype, the EL instructors were asked further questions: “In what ways, if any, would you use the levels of physiological arousal relating to students in the class to support classroom instruction? Can you provide some examples?”. Before ending the interview, the investigator debriefed the participants about this study.

5.4 Data Analysis

All transcriptions were categorized based on each interview question and repeated themes were highlighted.

5.5 Findings

Findings offered an insight into ways that the visual design and interaction of the prototype system could be further improved, and the factors that should be considered regarding the use of physiological data to support classroom instruction for foreign language teaching.

5.5.1 Feedback on the prototypes regarding design components

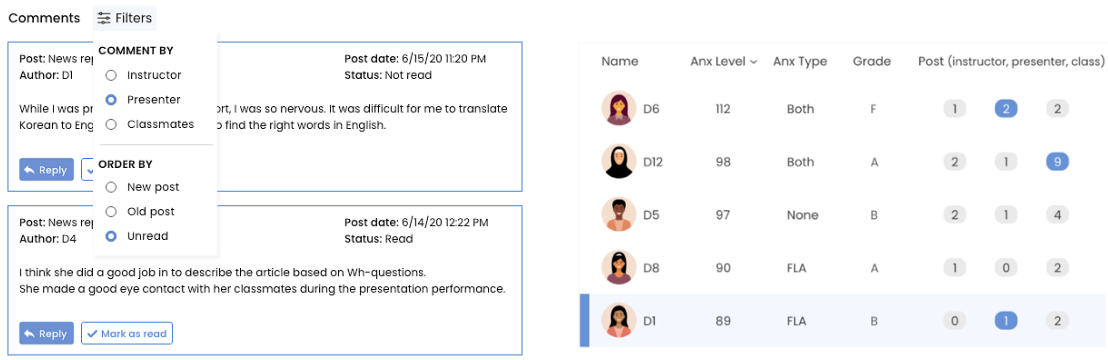
5.5.1.1 Visual feedback on low fidelity prototype

All participants were positive towards the design of the graphical interface itself, describing it as “simple and easy to read”. They highlighted that it was easy to understand the information present. As an example, P15 liked the layout of the interface. She felt that the content elements were ordered appropriately (e.g., video -> raw physiological data -> processed data), which would help a user accomplish their task. She stated that instructors may watch the video recordings first to grade students, and later on, data would be entered (e.g., grades, comments). In addition, P13-P15 appreciated the ability to view all content on a single page, without needing to scroll or keep multiple tabs open. The user could simply glance to gain an overview. For example, they examined the recording of presentations, changes in level of physiological arousal in real-time, and other information relating to the student. P14 stated that *"I can compare, contrast how I thought a student did with another student at a glance".*

In contrast, two instructors who have been teaching (P10) and had taught (P14) ELLs in Japan mentioned that most Japanese students they had taught did not have their own PC or laptop at home. However, they had smartphones. The two participants were concerned that some ELLs with limited access to technology may not be familiar with the system if instructors use it in class. The participants recommended keeping the system design as simple as the version presented in the pilot study, enabling access from both a desktop and mobile device.

Another concern raised by the participants in terms of interface design related to notifications and shading of the posts on the screen (Figure 13). P13 stated

"I am like a little confused…(what does the meaning of) instructor / presenter / class. I wasn't sure what that was referring to just at the first glance. I am sure like…if you hover over it, there would be probably an explanation or something, but I am not sure. Also, the highlights. Why does it have a different shade (on the notification)? (The investigator explained that the shade differentiates the un/viewed posts, and the numbering shows the total amount of posts created by each author – presenter, peers, instructor)." (P13)



(a) post list card view (b) post notification

Figure 13 Example interfaces of low fidelity prototype

In summary, the investigator kept the design and structure simple as the initial wireframed prototype was designed. However, there was some confusion faced by participants relating to an indicator on the notification of new posts. In addition to the misunderstanding of the design components on the system, some questions were proposed regarding the post components including when users can create posts (e.g., during or post presentation performance) and how they can share the posts with the instructor and peers in the class (e.g., privately or in public) while using the system. As a result, the investigator removed the posting section and the list view of the posting result section.

5.5.1.2 Visual feedback on high fidelity prototype

In general, all participants were interested in examining the clickable high-fidelity prototype. P15 who was also a participant in the professional development workshop (described in Chapter 4) responded with the excitement that, “we made (sketched) a number of ideas, but I really didn’t think you can literally develop it from there. This is a really good thing.” Suggestions were made by participants for visualizing data. These included: 1) displaying the anxiety type on the EDA graphs in real time along with the presentation performance and 2) integrating all EDA data to show how students have progressed to manage their anxiety over time and where they stood over the course of the semester.

P11, P12 and P15 responded that being aware of the specific types of anxiety faced (i.e., PSA vs FLA) would be helpful to develop personalized lesson plans for students with specific concerns. They suggest updating the display showing the types of anxiety faced at constant intervals (i.e., every second), so that instructors could immediately detect what event triggered a particular emotion of the ELLs throughout their presentation performance.

Another suggestion proposed by the participants (P12, P15) related to presenting a summary view. This could show how students have progressed over time to manage levels of anxiety. The user should be able to glance at the screen to compare progress with other students. P12 stated that, *"Looking at this (EDA data from one presentation performance) in isolation is not meaningful. It should inform how a student progresses over time as opposed to one presentation in isolation and see where they are in the course of the semester. I am not sure like where that progression happened and if she had gotten better over time…in a holistic way. (P12)"*

P15 suggested that a thumbnail icon that presents EDA arousal of all presentation topics could be valuable. Once a user selected the thumbnail icon, it would expand. The user can then compare the level of physiological arousal of the presenter by all presentation topics. She stated that instructors cannot always remember how effectively prior presentations were delivered by students, so she recommended a summary overview to monitor progress. Similar to the suggestions voiced by P12 and P15, P14 also suggested creating a panel on the interface to switch between content relating to different presentations delivered throughout the semester. *"...see if a student progressed or didn't progress throughout either the semester or like half a semester. That would be interesting to see if… during all of these…Give me an intervention to try and help them to see that. Like get a snapshot of one student for all different presentation topics on one page. (P14)"*

5.5.2 Perspectives of using physiological data in class

The following section describes the potential and concerns that the participants foresee on using physiological data in class.

5.5.2.1 Potential of using physiological data in teaching

Innovative way to empathize the emotions of ELLs

Presenting physiological data to instructors was thought to offer a means to promote empathy. P13 stated that *"I think I will be able to empathize with my students and their emotions. This is a very under explored area of education, especially second/foreign language education. I don't think that a lot of people notice that emotions are being portrayed when they are speaking a language that is not their own or that is not their language of nurture because they were not nurtured in this language. It is kind of [that] they only learn it in certain academic domains. And these academic domains can be very much trying to think of an analogy (through grading their performance). I don't know if you notice that you are able to use artificial intelligence on Blackboard to determine if the sentences the student writes are like what grade level it is, but it has nothing on speaking." (P13)*

Empower to visually see invisible physiological arousal in real-time

The instructors were interested in using biosensing technology, as it offers a method of detecting physiological changes and prominent arousal of the presenter over time, which would otherwise be difficult to gauge. P10 stated, *"It is very interesting because it shows to me… there is a lot you don't know based on what you see when it is too nervous. I like that."* (P10). P11 also highly valued the bioinformation provided by the sensing technology with the statement that, "*This is what their body is telling us, so it does help me."* (P11)

Helpful to link context with emotional state of ELLs

P11 to P14 stated that using autonomic arousal information would be helpful to understand how emotional states vary over time. P11 predicted that if instructors see this data for the same student throughout the semester, it would help instructors provide constructive personalized feedback. P12 showed high interest in using the physiological data as a chance to better understand their emotional state. If high levels of anxiety were detected for specific students over a period of time, instructors could have a conversation with students prior to their next presentation. P13 and P14 also predicted that they would go through and identify the specific instance where a high level of anxiety manifested, with a view to determining what the triggers were - *"I like to get an idea of how stressed the students are and would like to breakdown that you have where the student identifies."* (P13); *"If I see a lot of my students have the same type of anxiety or they all have similar anxiety levels from the EDA then that means, I think as an instructor I need to incorporate."* (P14).

5.5.2.2 Concerns of using physiological data in teaching

Face challenges to grade presentation performance

Using the physiological data in assessment became a controversial topic to discuss during the interview. Some instructors expressed that EDA data may bias assessment of a student's presentation. P10 provided an example case that *"So if, you know, students are nervous based on this data and you're like, well, should I mark them down on this part where my instinct was to mark them down because I feel like they weren't prepared, but oh, they were nervous. So, it could be used as like… a reason not to give them a lower mark or even the opposite. It could be a reason not to give them a better score, if you think."* (P10). P13 was concerned that the EDA data could be misconstrued, and may influence grading. She said *"Watching videos and the physiological data, it could be very much influencing the graded component. It could be very much misconstrued for a very performative aspect of language versus a form of self-improvement or process based in inquiry. Make sure to give a caveat, just like you are not being graded on your anxiety level, this is to help you. You will want the students to think that you know what their objectives are and what they are being graded are interconnected."* (P13). In contrast, P12 and P15 took the position that they would not modify their grading based on the physiological information, but they would refer to it and make comments about it.

Time and energy consuming for large class

Despite the benefit of using physiological data to support classroom instruction, most of the instructors (P10, P11, P14, P15) were concerned about the extra time and workload required for larger classes, and the amount of training needed for using the system on a wider scale. For example, P11 and P15 mentioned that interpreting emotional data in addition to grading is a lot of work for instructors. *"Looking in time, minutes and seconds, and then you have to decide whether that person is feeling this way or that way. It can be a lot of work and I am also listening to their (presentation) content to evaluate it at the same time."* (P11); *"It will be so much work for the instructors to manually input the types of anxiety for each student. As long as teachers do not need to evaluate about the peaks and the sensors grade them automatically, I am okay."* (P15).

P14 assumed that there might be some hesitation among instructors and students to try new technology and accept it. This aspect was observed and further discussed in Chapter 6.

Ambiguous interpretation and misuse of EDA data

Some participants highlighted concerns about misinterpreting the EDA data. P11 stated that *"The data can be accurate, and I still trust this data, but I could be wrong about the students. I can make mistakes when using it. I need more expert's ideas and suggestions on that."* (P11). Moreover, P12 wanted to avoid interfering in a situation and making it worse by using the EDA data, *"No need to tell students to calm down. My motivation of telling the students might affect their presentation in a good or bad way, so I don't know."* (P12).

5.5.3 Application of using physiological data in teaching and learning

5.5.3.1 Teaching

Prepare customized practice when high levels of anxiety are detected

A variety of use cases concerning the integration of physiological data with teaching were suggested. For example, P10 wanted to use the anxiety type data as an individual coaching tool to train students for public speaking contests. He stated that he would check the graphs relating to arousal present on an interface, to determine high levels of anxiety, and then offer personalized advice to students. He suggested grouping students with differing levels of anxiety together in pairs, and having students classified with 'lower levels of anxiety' coach those with 'higher levels of anxiety'. P11 stated *"I will develop different scaffolding activities based on the person who is more concerned about making the speech perfect versus being in front of the audience. Having different approaches to help students will probably inform me how to modify my lesson plans accordingly."*. P11 and P14 described wanting to adjust classes by having a 10–15-minute warm-up activity to help students practice dealing with instances of high anxiety such as right before a Q&A section, toward the beginning or end of the presentation, and when transitioning between slideshows.

Track progress of improvement in learning and anxiety

P13 and P14 stated that using the prototype system with the EDA data at the beginning of the semester might be helpful to see how a student's anxiety has progressed or gotten better throughout the semester.

Apply biofeedback to other scenarios

The participants were interested in exploring how EDA data could benefit a variety of scenarios other than when delivering presentations in class. These included when speaking in large discussion groups, small group discussions, and one-on-one conferences. They wanted to see how students' levels of anxiety vary based on the situation. P13 stated that *"How you would speak in performance speech events would be different from how you speak to your friend and see if they had the same anxieties regarding a certain language use?"* P13 also proposed the idea of using physiological data in classes relating to sociology and psychology to determine how levels of anxiety differ in other classes. She stated that *"It would be really cool if teachers and students could use EDA data in sociology or psychology topics. They could look at their anxiety levels and figure out themselves if there are trends."* (P13).

P14 described broader use cases of the prototype system using the EDA data*, "You could apply it outside of ESL. Some colleges require all students to take a speech class. They could use it. You can give it a broader scope."* (P14).

Assess course evaluation oneself

P13 wanted to use a system with EDA data to check at which point in the class higher levels of anxiety were experienced. P13 could then identify whether the particular part of the class where anxiety was provoked could be taught in a less anxiety-provoking way. *"It gives me data to be like maybe I need to reteach this or maybe I need to redo this because the students are having anxieties about it. From the teacher's standpoint, I could think of how I should construct my assignments to lower the affective filter (based on the biofeedback information), so the students don't feel so nervous."* (P13).

5.5.3.2 Learning

Long-term use of a biosensor for personal and educational purposes throughout the semester

P16-P18 expressed that they wanted to extend their use of the biosensor to track their physiological data throughout the period of a day. They wanted to review their own physiological data collected, and use it as a tool to identify contributing factors to their emotional states. P17 pointed out that wearing a biosensor for a short period of time probably made her feel more anxious than usual, so she wanted to wear the biosensor for a whole day, so that her instructors could view her daily levels of anxiety and provide more targeted help to her in class. Surprisingly, the students interviewed in the study described in this chapter and the students who allowed the investigator to collect their physiological data, were not concerned about having their anxiety displayed to the instructor. Rather, they expected that they could receive more guidance and personalized practice from their instructors to improve language learning.

P17 and P18 experienced wearing a biosensor while delivering presentations in class as a part of this study (Chapter 3). They knew what data the biosensor was collecting and how the data would be presented to users. However, only four biosensors were available to use in the study, in which they had to take turns wearing the biosensor with other students, while delivering a presentation.

5.6 Discussion

5.6.1 Using different degree of prototype fidelities in the Design Thinking Process

The overall opinion on the visual attributes of both low and high-fidelity prototypes were very positive. A noticeable result was that the feedback on the high-fidelity prototype had more suggestions for improvement, compared to the feedback on the low-fidelity prototype, in terms of integrating physiological data with the system. This may be attributed to the fact that the selectable elements on the screen helped participants to promote interaction with the system compared to the static low-fidelity prototype. However, the low-fidelity prototype helped the investigator weed out the design components which could distract or confuse the participants.

Moving through the design thinking process, followed by a period of iteration, resulted in the development of the final prototype to use as an intervention in Chapter 6.

5.6.2 Threat or Opportunity

5.6.2.1 Fairness in assessment

Although biosensor-based feedback can be used to help an instructor better understand the emotional states of students in a class, the ethics of using these cues remain unestablished. One example of a practical dilemma discussed by participants was determining whether the students' emotional states would be reflected in their grades. If high levels of anxiety are detected, it could be attributed to the lack of practice undertaken when preparing the presentation, and instructors may mark down the grade as a result. In contrast, a high level of anxiety could be a reason for students to ask instructors to offer another opportunity to them in order to gain an improved grade. In this context, the awareness of the learners' emotional state was a double-edged sword for instructors to evaluate the students' performance. Even though it was challenging to set the thresholds for certain metrics that inform educators when to intervene and help anxious students by using physiological arousal, it led to discussion among HCI researchers and TESOL educators.

5.6.2.2 Misuse, abuse vs. efficiency of technology

Automated assessment tools with artificial intelligence are already pervasive within the educational environment. A learning management system (LMS), such as Blackboard, employs artificial intelligence to automatically grade students' essays and catch instances of suspected plagiarism. The participants in this study took the position that if the technology is not biased, as accurate as human graders, and able to save time in the process of grading, it will be favored above hiring human graders. However, if it demands a lot of work and time to learn, they responded that they would not use it in their teaching. This led the investigator to consider whether there could be issues of misuse and/or abuse of using the sensing technology within an educational setting. While advances in technology may save time to complete tasks, it can lead to laziness or neglect of job duties (Tarafdar et al., 2015). It could cause instructors to become passive educators by relying on the outcomes provided by the system. This may trigger educators to blame students for their feelings of anxiety, simply due to a lack of preparation and practice for their performance.

As a result, the educators may give negative feedback on students' performance if they blindly trust and rely on the information provided by the system (Atabek, 2020; Glendinning, 2018). Moreover, the technology can be counter-productive to students if educators abuse the technology to share private biosensing data. This could be an invasion of personal privacy. To protect each student's information, data security measures including encryption or cryptography should be considered. Further discussion with users and stakeholders of biosensing solutions would be beneficial to explore how data has the potential for misuse/abuse, and ways in which this can be minimized. Educators, researchers, and developers also need to be transparent with students about how their bioinformation can be collected, analyzed, and used in a system, and for purposes of teaching. It is also necessary to inform students of the potential risks of sharing their bioinformation with educators and schools.

5.6.4 Giving autonomy to users in the adoption of new technology and its information

In this study, perspectives of using EDA data to support teaching were elicited when reviewing prototype designs and functions. Attitudes from participants could be classified under the following themes: technology acceptance (P10, P11, P13, P15, P17), doubtfulness (P12, P14, P16), and an in-between attitude (P10 and P11). These aspects motivated the investigator to deeply delve into instructors' perspectives on the use of biosensing technology in their teaching, through conducting a mixed method approach (e.g., survey to rank anxious students, interview). This has been described in Chapter 6.

The participants who placed trust in the physiological data were willing to use it to support classroom instruction. P11 highly trusted the physiological arousal data over her observations of students' facial expressions. This is because she thought that the cues presented on the screen would give her an insight into how the students' body is feeling. She could then compare the data with other students. P10 also noted the accuracy of physiological data in terms of detecting moments of anxiety among students when delivering a presentation, which could be missed by solely relying on the observation of facial gestures.

When traditional methods previously used to determine emotional state were not found to be satisfactory, the participants may expect that this new biosensing technology can bring more efficiency to support classroom instruction. In contrast to participants who favored the use of new technology, participants P12, P14, and P16 casted doubt on the accuracy of the data collected using the sensing technology, and how this would be interpreted. The first reason to be skeptical of this technology was that hand gestures made during the presentation may impact EDA levels. As a wristband can be used to detect emotional state, moving arms around may affect the way that EDA is collected They also mentioned that environmental factors other than PSA or FLA can affect the physiological arousal. Thus, they wanted to examine the physiological data with a video of presentation performance, and compare it with their traditional strategies of determining emotional state (e.g., observation of facial gestures), to make the physiological data feel more reliable.

5.6.5 Privacy of access physiological data

It was interesting that none of the instructors in this study raised concerns about privacy of learners' physiological data. Instead, they were highly interested in viewing recorded performance videos with the autonomic arousal changes among the students over time. In the same vein, when the investigator asked ELLs to wear a biosensor when presenting, participants expressed interest in seeing their own physiological data. They wanted to become more 'mindful' by observing when the emotional changes occurred and what triggered them while delivering a presentation in class. Additionally, they hoped that they can come up with strategies to reduce anxiety, and receive support and personalized practice by instructors to overcome emotional vulnerability in the learning process. In general, the instructors and students were optimistic regarding the potential that biosensing technology has to offer, and valued its innovativeness to support classroom interactions. Despite no concerns being raised by the participants, the investigator wanted to highlight that researchers and educators need to be transparent with learners about what they do with their physiological information, and make clear that their physiological data is stored safely and cannot be accessed for other purposes. In addition, it is necessary to create a comfortable environment where students can freely allow researchers and educators to gain access to their biomarkers, or withdraw if the students do not want to disclose their physiological data.

5.7 Limitations and Future Work

A pilot study was described in this chapter. Interacting with both low and higher fidelity prototypes allowed the participants to provide feedback on effectiveness and utility. However, the feedback was not attributed to the outcome of task-based users tests. It was rather an inspection of aesthetics and the functionality presented on the interface. To deeply investigate the efficacy of the system, whether it suits the practical needs of learners and educators in an ESL/EFL environment, participants should be given similar tasks to the actual tasks that learners normally conduct in class.

5.8 Summary

This chapter describes the process of iteration, moving from a low-fidelity to a higher-fidelity prototype. Although creating an interactive, clickable prototype is more time-consuming than developing a set of wireframes, it was found to be helpful to prompt participants to participate and share their perspectives. Discussion of the prototypes enabled the investigator to identify specific areas of the prototypes to refine. Participants suggested that the prototype should be simple and intuitively understandable, and should not require an extra training session to learn about the system. Moreover, the pilot study with target users including language teachers and ELLs allowed this study to explore their practical needs and expectations when integrating a new biosensing technology in a second/foreign language teaching and learning context. It provided concrete use cases (e.g., speech training, counseling with students, reference in assessment, students' self-monitoring, etc.). Concerns were also highlighted as this technology has been relatively unexplored in the field of education.

In Chapter 6, the investigator demonstrates the efficacy of using this prototype system in perceiving the emotional states of the ELLs and teaching.

Chapter 6. Evaluation of the Efficacy of a Prototype System using EDA Data

6.1 Introduction

Biosensing technologies are increasingly being used in a number of settings (Howell et al., 2018). These include educational environments (Lane & D'Mello, 2019; Sharma et al., 2019) and in the field of medicine (Merrill et al., 2019; Pataranutaporn et al., 2019), where biosensors have been used to collect physiological data to better understand individuals' emotions. However, research suggests that difficulties are faced interpreting biosensing data (Bahl et al., 2021; Kirsch, et al., 2013). Biosensing technologies have yet to be explored within the field of language education. The study described in this chapter was conducted to examine the potential of biosensing data integrated into an interface to enable instructors to identify students experiencing anxiety. In-service experienced and novice English language teachers/instructors were recruited with the aim of testing the efficacy of a prototype system (developed in Chapter 5). A mixed methods approach was adopted. The findings helped identify the potential problems that instructors may face while assessing the emotional states of learners including inconsistent and subjective emotional assessment within and between raters (discussed in Chapter 2). Furthermore, analysis from the individual interviews revealed participants' perspectives towards new biosensing technology. These included concerns regarding technology reliance, resistance, and acceptance. Findings also revealed more about how instructors would apply bioinformation to provide tailored emotional and instructional scaffolding for students. A range of use cases are described. The difficulties using this type of technology, along with ethical concerns, are also discussed.

In this chapter, the investigator addressed the following research questions:

- RQ 4-1: Assessment between raters

  Is there agreement between instructors on ranking learners' degree of speaking anxiety

  - **without** using the biosensing data presented on the prototype system and solely relying on observation (Evaluation 1)?
  - **with** - referring to the biosensing data on the prototype (Evaluation 2)?

- RQ 4-2: Assessment by rater

  When it comes to ranking the learners' degree of speaking anxiety, are there any similarities or differences between Evaluation 1, Evaluation 2, and biosensing based ranking data on the prototype (ground truth)?

  - Compare Evaluation 1 (observation) and Evaluation 2 (observation + referring to the biosensing data on the prototype system)
  - Compare Evaluation 1 and ground truth (biosensing data on the prototype system)
  - Compare Evaluation 2 and ground truth

- RQ 4-3: How does the instructors' awareness of emotional states of ELLs vary over the experiments?
- RQ 4-4: How would instructors use information presented via a system relating to the emotional states of students, in terms of teaching strategy or class activities?

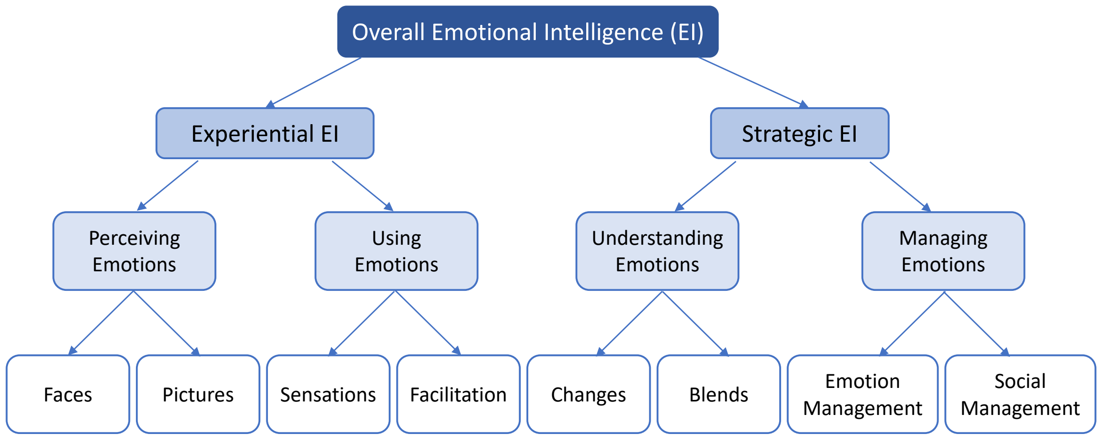
In terms of contributions, the evaluation described in this chapter could offer value to other researchers aiming to determine the value of biosensor-based feedback to support educators using this technology. Furthermore, findings from the evaluation have led to the development of design guidance that interface designers can use to assist ESL/EFL instructors when teaching, by aiding them to identify the emotional states of the ELLs and provide emotional and instructional support for them.

6.2 Method

6.2.1 Measures

6.2.1.1 MSCEIT Questionnaire

The Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT V2.0) questionnaire (Mayer et al., 2002) consists of 141 items using a five-point rating scale with images of facial expressions and given scenarios (Appendix G) to estimate a person's emotional intelligence (EI). The questionnaire was designed to measure the ability to *perceive* emotions in oneself and others, to *use* emotions in cognitive processes, to *understand* emotional information, and to *manage* them in oneself and others (Mayer et al., 2002), as shown in Figure 14. Item scores are compiled to generate a total EI score. A person with a score above 110 is considered to have a high level of EI. Someone scoring below 90 is counted to have perceptions of emotions so unique, which may cause interpersonal problems (Leddy et al., 2011). The MSCEIT was selected for this study for two reasons. Firstly, to see if the number of years of teaching experience influences the EI scores resulting from the questionnaire, and secondly, to investigate whether the participants who achieved high EI scores in the questionnaire can evaluate the degree of students' anxiety similar to the degree to which a biosensor measures.



Standard Score Range	Interpretive Guideline
69 or less	Improve
70-89	Consider Developing
90-109	Competent
110-129	Skilled
130+	Expert

Figure 14 Illustration of MSCEIT v2.0 scores structure

6.2.1.2 The Prototype System

The high-fidelity prototype developed in Chapter 5 (Section 5.2) was presented to participants to elicit further feedback on the efficacy of the prototype system to detect the emotional state of ELLs. The prototype system was evaluated with 17 participants who had no prior experience of interacting with biosensor-based feedback.

6.2.2 Participants

The participants in the study described in this chapter were selected through a simple random sampling approach. The investigator sent an email with a recruitment flyer (Appendix H) to M.A. TESOL programs at universities in the United States, including UMBC, American University, The University of Alabama, New York University, and Northeastern Illinois University. The study lasted 8 weeks, and participants received compensation of $180. This study was approved by the Institutional Review Board (IRB) at UMBC.

Twenty-five instructors were initially recruited for the study. Eight instructors withdrew due to either difficulties committing the time needed for the study, or withdrew for personal reasons. In total, 17 instructors (2 males, 14 females, 1 preferred not to answer) aged between 21-76 (mean = 34.53 ± 14.76) participated in the study (Table 10). Participants included seven *experienced* instructors (P1, P2, P20, P21, P23, P24, P25), who had experience of teaching ELLs for periods of longer than five years in either a formal school setting or an English language institute. Ten *novice* instructors (P3-P9, P12, P16, P22) were also recruited for the study. The novice instructors had less than five-years teaching experience as an assistant/co-instructor while attending the M.A. TESOL program. Nine out of 17 participants identified as Native English-Speaking Teachers (NESTs) whose first language was English, and eight participants identified as non-NESTs (NNESTs), hailing from different cultural backgrounds (Korean, Bolivian, Spanish, Chinese, Serbian, Polish, and Romanian).

Table 11 Summary table of demographic characteristics of the participants

ID	Gender	Age	Ethnicity	Occupation	NEST	Teaching experience
1	F	35	Asian	teacher – elementary school	NNEST	Experienced
2	F	32	White	other education	NEST	Experienced
3	-	-	Other	-	NNEST	Novice
4	F	23	White	teacher – high school	NEST	Novice
5	F	22	Hispanic	other education	NNEST	Novice
6	M	35	White	teacher – elementary school	NEST	Novice
7	F	23	Asian	other education	NNEST	Novice
8	F	-	Asian	other education	NNEST	Novice
9	F	21	White	social government services	NNEST	Novice
12	F	49	White	other education	NEST	Novice
16	F	22	White	college instructor	NEST	Novice
20	M	29	White	other education	NEST	Experienced
21	F	44	Black	social government services	NEST	Experienced
22	F	24	White	other education	NEST	Novice
23	F	76	White	college instructor	NEST	Experienced
24	F	45	White	other education	NNEST	Experienced
25	F	38	White	college instructor	NNEST	Experienced

Although 25 participants are in line with the numbers gathered for similar studies (Mestanlk et al., 2014; Pijeira-Díaz et al., 2016; Umair, 2021), it has been acknowledged that a larger sample would have been preferable. Challenges were faced recruiting instructors during the lockdown period. Difficulties obtaining sizable samples have been encountered in related studies. For example, Umair (2021) conducted his pilot study with six participants to explore understanding of thermochromic displays. Physiological arousal data was presented. In his later study, the researcher recruited 12 participants to examine how prototypes may shape people's understanding of their emotions in everyday life. In the study described in this chapter, working with groups of novice and experienced instructors offered the potential to further investigate whether teaching experience influences the ability to perceive the emotional states of learners in class, and technology acceptance in an ESL/EFL environment.

Due to lockdown mandates resulting from the COVID-19 pandemic, the study was conducted remotely. Tools such as email and web-based conferencing software (e.g., WebEx and Zoom) were used to communicate with participants. All participants who agreed to sign the consent form were required to have access to high-speed Internet and their own personal computer or laptop to participate in the study.

6.2.3 Procedure

As shown in Figure 15, 17 instructors (7 experienced and 10 novice) were asked to fill out the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT V2.0) questionnaire (Mayer et al., 2002) in Week 1. It took approximately 30 to 45 minutes to complete the assessment. The participants had to complete the questions in one sitting, or else the system would not save their responses. Participants were then asked to view presentation footage of students delivering presentations, and were asked to observe facial expressions and upper body gestures to evaluate the students' anxiety level on the ranking survey, so the ability-based scales in MSCEIT could serve as a baseline to assess instructors' emotional intelligence skills (Fiori et al., 2014).

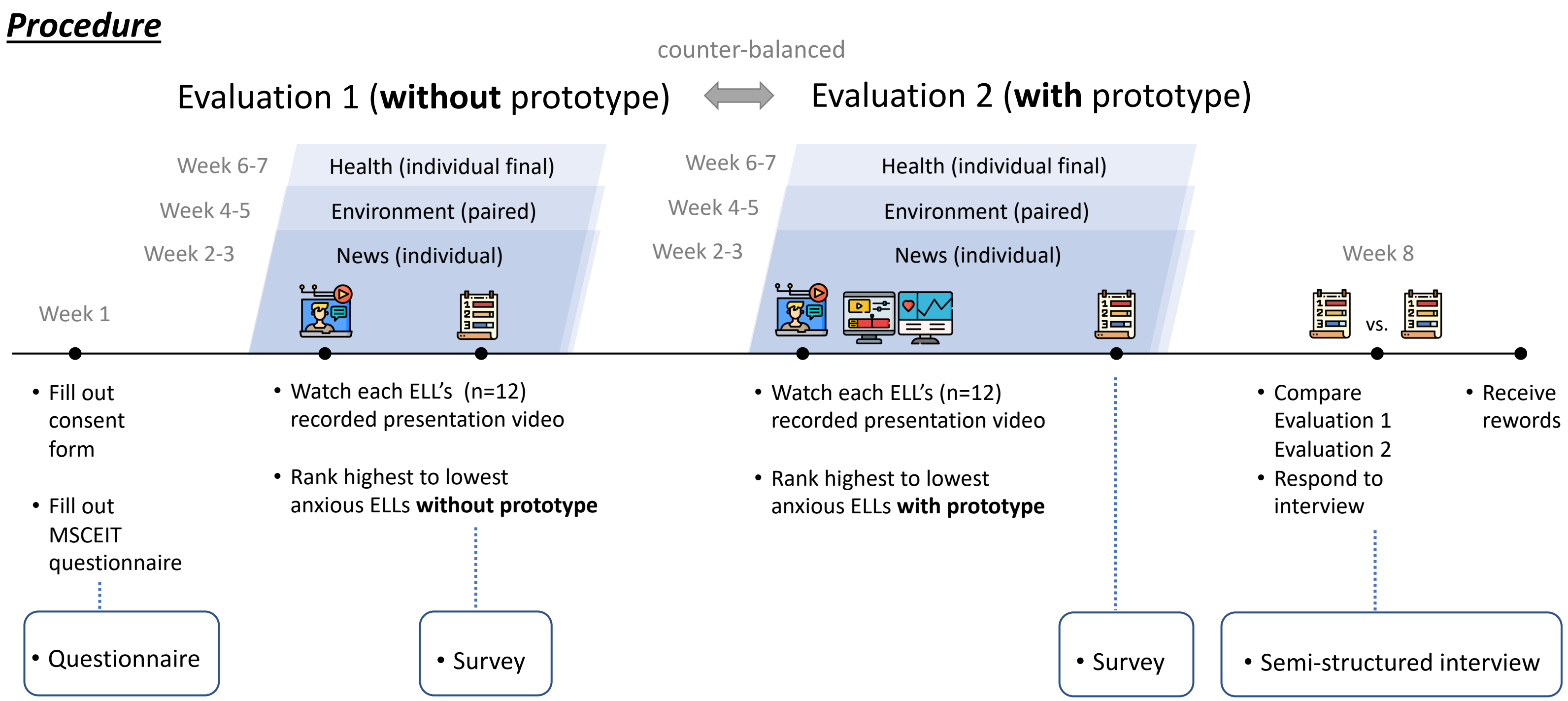


Figure 15 Procedure to compare human only vs. human with system evaluation

After completing the MSCEIT questionnaire, each participant was asked to watch videos that were recorded in one of the ELI Speaking classes in the Summer of 2019. The investigator positioned an audio/video Canon VIXIA HF R800 digital camcorder at the back of the classroom, and a Sony ECMAW4 microphone and receiver close to the presenter would be standing. Each presentation could be recorded, and any non-verbal behaviors could also be captured during this time. Table 12 shows the demographic characteristics of the ELLs. The class was an inclusive class where ten hearing ELLs and two Deaf/Hard of Hearing (DHH) ELLs, from a wide range of cultural and linguistic backgrounds, were learning English. The students ranged in age from 19 to 29 (mean age 21 ± 2.92 years). S11 described experiencing moderate to severe hearing loss (HL) (65dB HL), while S12 described having a greater level of loss (85dB HL). Utilizing cochlear implants led to S11 experiencing milder levels of hearing loss (40dB HL), with S12 experiencing moderate levels of hearing loss (41dB HL).

Table 12 Summary table of demographic characteristics of ELLs in recorded videos

ID	**Gender**	**Age**	**Nationality**
S1	F	29	South Korea
S2	F	20	Japan
S3	F	24	Saudi Arabia
S4	F	19	Japan
S5	M	20	Gabon
S6	F	20	Japan
S7	F	21	Japan
S8	F	20	Japan
S9	F	19	Saudi Arabia
S10	F	19	Japan
S11	M	19	Saudi Arabia
S12	F	22	Saudi Arabia

The ELI instructor assigned four presentation topics to students over the course of the semester. ELLs presented on topics relating to family trees, news reports (*News*), environmental issues (*Environment*), and health and cleanliness (*Health*). At the time of the presentations, the ELLs donned the biosensing technology on their wrists, which enabled the investigator to capture physiological information relating to their emotional state.

The first presentation (Family tree) served as an opportunity to acclimate the ELLs to the experimental protocol and to wearing the biosensor wristband. As a result, the other three presentation topics were used for purposes of the study.

Participants were required to watch videos of students presenting on these topics (*News, Environment, and Health*). Each video showcased two individual presentations relating to the topics of *News* and *Health*, and one paired presentation relating to the topic of the *Environment*. For the paired presentations, each member of the pair delivered part of the talk. The health-related presentation was the final presentation of the semester. News reports, environmental issues, and health and cleanliness were henceforth referred to as *News (Individual)*, Environment (*Paired)*, and Health (Individual *Final)* in this study.

From Week 2 to Week 7, the participants watched a total of 36 videos (12 students' presentations - 3 topics (*News (individual)*, *Environment (paired)*, and *Health (individual)*)) in the condition termed "Evaluation 1" (evaluation solely relying on observation **without** referring to prototype). Participants also watched the same 36 video recordings in the condition termed "Evaluation 2" (evaluation **with** observation and referring to prototype). In the Evaluation 2 condition, the participants were informed that the prototype was a tool that could be referred to, which showed the physiological information of each student alongside the video footage. The aim was not to force participants to use the biosensor-based feedback as the primary means of determining levels of anxiety, but to offer additional data points which they could use, should they wish to. Along with a growing critiqued notions of understanding "non-use" technology in HCI research (Knowles & Hanson, 2018; Waycott et al., 2016), this study aimed not to force participants to use the biosensor-based feedback as the primary means of determining levels of anxiety, but to offer additional data points which they could use or ignore.

The investigator randomized the order of Evaluation 1 and Evaluation 2 to minimize the likelihood of confounding variables being introduced. To minimize carryover effects, the participants watched 12 videos in a randomized order for each week.

The participants spent approximately 60-90 minutes on each task. This is similar to the duration of an actual language class where all the students in a class deliver a presentation one after the other. The tasks from Week 1 to Week 8 are presented in Table 13.

Table 13 The whole procedure of the tasks from Week 1 to Week 8

Week	Participants who started with Evaluation 1 (task without prototype) to Evaluation 2 (task with prototype)
1	▪ Fill out a questionnaire Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT)
2	▪ Evaluation 1 – Watch 12 videos (News - individual) and rank students
3	▪ Evaluation 2 – Watch 12 videos (News - individual) and rank students
4	▪ Evaluation 1 – Watch 12 videos (Envir. - paired) and rank students
5	▪ Evaluation 2 – Watch 12 videos (Envir. - paired) and rank students
6	▪ Evaluation 2 – Watch 12 videos (Health - individual final) and rank students
7	▪ Evaluation 1 – Watch 12 videos (Health - individual final) and rank students
8	▪ Interview

(a)

Week	Participants who started with Evaluation 2 (task with prototype) to Evaluation 1 (task without prototype)
1	▪ Fill out a questionnaire Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT)
2	▪ Evaluation 2 – Watch 12 videos (News - individual) and rank students
3	▪ Evaluation 1 – Watch 12 videos (News - individual) and rank students
4	▪ Evaluation 2 – Watch 12 videos (Envir. - paired) and rank students
5	▪ Evaluation 1 – Watch 12 videos (Envir. - paired) and rank students
6	▪ Evaluation 1 – Watch 12 videos (Health - individual final) and rank students
7	▪ Evaluation 2 – Watch 12 videos (Health - individual final) and rank students
8	▪ Interview

(b)

Evaluation 1 (observation **without** referring to prototype) served as a control. Each instructor was asked to rank ELLs based on their levels of anxiety manifested in the presentation recordings, using criteria which they would ordinarily use to determine anxiety in a class setting (e.g., observation of facial expressions, drawing upon previous experience with anxious students, etc.). They were then asked to score the quality of the presentation out of 100 points, using their own rubrics.

In Evaluation 2, each instructor was asked to perform the same tasks as in Evaluation 1. However, in Evaluation 2, participants could refer to a prototype system which provided information relating to learners' physiological responses. In total, participants would be able to view the audio-video recorded performance of 12 ELLs, their physiological arousal synchronized with the audio-video timelines, and information relating to the students' degree of anxiety and types of anxiety identified during the presentation (e.g., PSA, FLA).

There was a period of one week (at least 5 days) between conditions to reduce the likelihood of recall bias and an assimilation effect. As people tend to judge performance based on their pre-existing experiences or assumptions (Steiner, 1989), a certain period between Evaluation 1 and Evaluation 2 was necessary so that participants' memories of watching the video recordings would fade away.

On the ranking survey form (Appendix I), each instructor ranked student performance by level of anxiety. For example, students who appeared to experience the highest level of anxiety would be ranked closer to one. Participants were also asked to write the reasoning behind why they ranked students in the particular order selected.

For Evaluation 2, guidance (Appendix J) of how to use a prototype system (Figure 16) was shared with each participant by email. Three URL links containing each different presentation topic were sent to participants one at a time based on the weekly task. This was conducted to reduce the likelihood of hindsight bias, exaggerating the outcome after watching the entire series of videos in advance. (For the concerns with protecting privacy of the students in the URL links, the following link that made students' faces blur in the video is shared in this chapter: https://edu-project-8aa26e.netlify.app/).

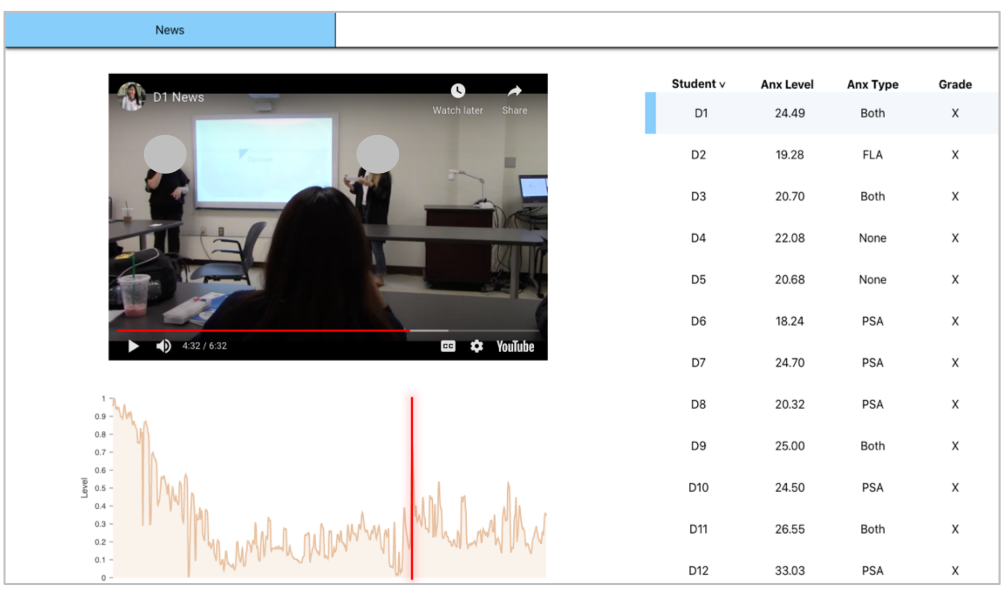


Figure 16 A prototype system displaying audio-video recorded performance, their physiological arousal synchronized with the audio-video timelines, and information of the students' degree of anxiety and types of anxiety during the presentation.

After Evaluation 2 was conducted, individual interviews were undertaken with participants (Appendix K) to elicit in-depth feedback on how the system could be used to support their teaching. During the interview, the investigator presented the pre- and post-ranking survey forms (Evaluation 1 and Evaluation 2) completed by each instructor, and asked them to compare how students' levels of anxiety were ranked. Participants were also asked to describe the likes/dislikes and advantages/concerns of using the prototype system. Each interview took around 40 minutes to an hour, and an online e-gift card ($180) was sent to each participant.

Participants received an email from the investigator with the tasks to be performed and links to each video. Videos were uploaded to YouTube so they could be readily streamed by participants. Links to ranking forms were also included into each email.

6.3 Data Analysis

6.3.1 Inter-rater Agreement among Instructors in Perceiving Emotions of ELLs (Between subjects)

Kendall's Coefficient of Concordance W

Evaluating one's emotional state provokes a controversial discussion in terms of its reliability and agreement among evaluators. As emotional states are not static and vary between individuals, it is challenging to reach consensus among instructors to evaluate the emotional states of learners (Garner, 2010). In addition, an instructor's own personal beliefs or ability to perceive the learners' emotions may be different and trigger differing interpretations among raters. For example, Korean teachers are more likely to be accepting of young children's emotional outbursts than U.S. teachers (Hyson & Lee, 1996). These concerns were examined by employing an inter-rater agreement test called Kendall's coefficient of concordance w (Kendall, 1948) with the ranked data evaluated by English language instructors. The Kendall's w test is a non-parametric statistic using either an interval scale (i.e., temperature, time on the clock) or an ordinal scale (i.e., ranks) to assess agreement among raters. The test ranges from 0 to 1. Zero relates to no agreement between raters whereas one relates to perfect agreement.

In the study described in this chapter, *Kendall's w test* was used to examine whether all instructors ranked lists of students similarly or differently. In addition, the investigator more closely examined if there was any agreement in the evaluation between two groups: novice instructors vs. experienced instructors. The investigator hypothesized that the results of both statistical tests might be closer to zero, which means that instructors are highly likely to disagree when ranking students based on the manifestation of anxiety among learners, no matter how many years they have taught ELLs. This could be attributed to the fact that instructors are not able to make consistent judgments due to the absence of objective dimensions/rubrics to evaluate the emotional states of learners.

6.3.2 Intra-rater Agreement Between Pre-Post Use of the Prototype Intervention (Within Subjects)

Kendall's Rank Correlation Coefficient (Kendall's Tau Coefficient)

To identify whether referring to the system influences the differences between the ranked data from Evaluation 1 (human only evaluation) to Evaluation 2 (human with prototype system evaluation), the *Kendall's tau coefficient test* was conducted using intra-rated ranked data.

6.3.3 Semi-structured Interviews

The investigator used an analysis method similar to that of the focus group interview. All transcriptions were categorized based on each interview question (Appendix H), and inductive coding for repeated themes was highlighted by the investigator using constant comparison (Boeije, 2002). The following themes were investigated in the transcripts: perspectives on the proposed prototype system including the feasibility of identifying the emotional states of ELLs, and opportunities to provide emotional scaffolding in teaching. In addition, concerns regarding ethical issues such as misuse and abuse of the data presented, complexity of using the system, training requirements of the biosensing technology, and potential for malfunction of the system, were investigated.

6.4 Data Results

6.4.1 MSCEIT questionnaire

Table 14 shows the descriptive statistics for the total standard EI score and each of the four branch standard scores of perceiving emotions, using emotions, understanding emotions, and managing emotions among instructors.

Table 14 Summary of MSCEIT questionnaire results

ID	Overall EI	Branch 1 Perceiving	Branch 2 Using	Branch 3 Understanding	Branch 4 Managing
1	98.58	106.08	97.75	99.02	89.24
2	126.01	117.85	123.24	110.69	109.62
3	98.02	94.88	105.23	90.24	100.60
4	96.12	106.75	91.27	96.41	90.10
5	108.66	112.05	104.36	98.53	101.53
6	83.19	80.39	77.51	90.55	99.94
7	100.37	99.29	113.49	91.19	95.66
8	84.92	85.41	95.45	87.52	89.41
9	93.41	83.60	87.89	109.96	101.56
12	85.19	84.44	78.36	94.86	97.99
16	112.84	111.96	108.51	106.79	104.82
20	118.09	119.31	124.01	102.86	104.87
21	105.39	106.27	109.26	106.51	91.37
22	104.49	107.85	101.20	101.49	96.75
23	104.38	101.31	99.93	108.79	96.91
24	109.06	104.45	100.89	116.32	98.28
25	103.52	106.18	89.35	118.26	93.86
mean	101.90	101.65	100.45	101.77	97.79
SD	11.56	12.02	13.32	9.34	5.84

Standard Score Range	Interpretive Guideline
69 or less	Improve
70-89	Consider Developing
90-109	Competent
110-129	Skilled
130+	Expert

The average score for the 17 participants on MSCEIT was 101.90 (± 11.56). This is within the ‘competent’ range according to the interpretive guide. The boxplot in Figure 17 shows that the experienced instructors achieved much higher median scores on the overall EI and the domains of the skills in Perceiving and Understanding emotions. The participants (P6, P8, P12) who scored under 90 are considered to develop the ability of EI. They were all novice instructors. The results of MSCEIT scores of experienced and novice instructors are listed on Table 15.

Table 15 Comparing mean scores and standard deviation (SD) of Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) for each of the four EI skills between experienced and novice instructors.

		Perceive	Use	Understand	Manage
Experienced instructors	Mean	109.29	108.78	106.35	108.92
	SD	9.53	6.93	13.16	6.90
Novice instructors	Mean	96.72	96.66	96.33	96.75
	SD	10.24	12.56	12.40	7.47

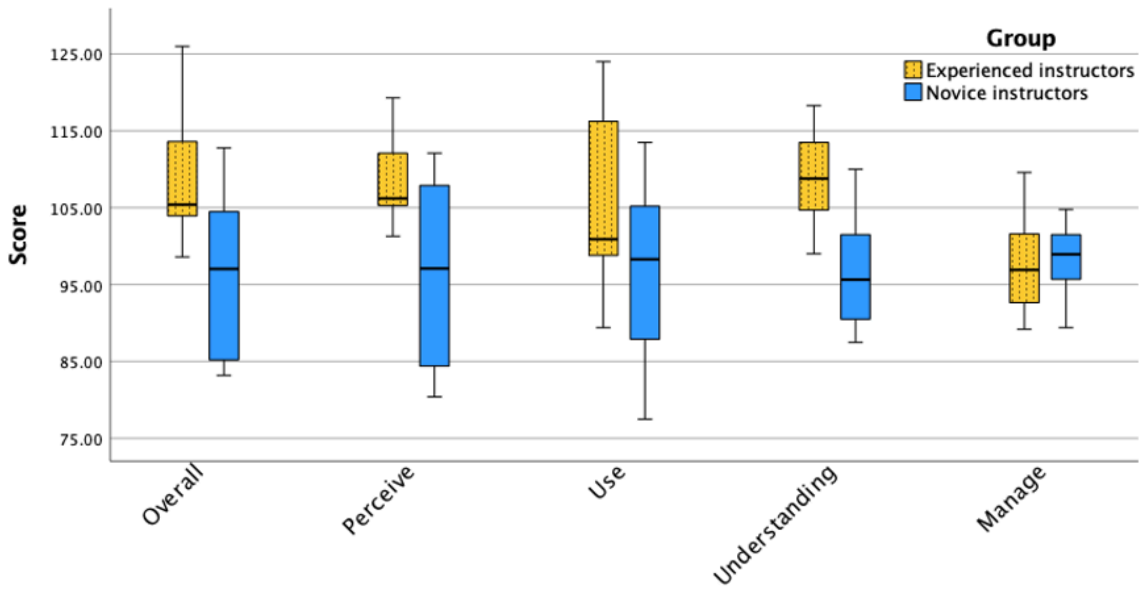


Figure 17 Boxplot of Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) standard score descriptive statistics among experienced instructors and novice instructors

6.4.2 Inter-rater Agreement among Instructors in Perceiving Emotions of ELLs (Between instructors)

Kendall's coefficient of concordance (W)

The ranking scores for agreement in Evaluation 1 (observation) and Evaluation 2 (observation with using the prototype system) were analyzed using Kendall's coefficient of concordance (W). The null hypothesis, that is, the 17 participants were not concordant with each other, was rejected. All p values were significant at 0.001.

When the participants ranked the level of anxiety among the ELLs by using a traditional method of observation (Evaluation 1), the concordance (W) values among the participants gradually decreased as the experiment progressed. 17 participants presented a moderate level of agreement (.539) in ranking the students' level of anxiety towards the beginning of the experiment from Week 2-3 (Table 16). However, the concordance (W) values decreased (.340) and showed a poorer level of agreement (.194) when the final presentation was evaluated.

Table 16 Kendall's coefficient of concordance (W) results with Evaluation 1 (observation)

	News	Paired	Final
Kendall's W[a]	.539	.340	.194
Chi-Square	100.86	63.52	36.25
Df	11	11	11
Asymp. Sig.	<.001	<.001	<.001

[a] Kendall's Coefficient of Concordance (n=17)

In Evaluation 2, when the participants ranked the anxiety level of the ELLs with the prototype system, a fair level of agreement was identified when ranking anxious students in order throughout all presentation topics. The concordance (W) values were .349, .331, and .367 on each presentation topic (Table 17). Although all instructors did not have high levels of agreement when assessing the level of anxiety among ELLs when using the prototype, a consistent degree of agreement throughout the presentations was identified.

Table 17 Kendall's coefficient of concordance (W) results with Evaluation 2 (observation with using the prototype system)

	NewsSys	PairedSys	FinalSys
Kendall's W[a]	.349	.331	.367
Chi-Square	65.31	61.84	63.64
Df	11	11	11
Asymp. Sig.	<.001	<.001	<.001

[a] Kendall's Coefficient of Concordance (n=17)

6.4.3 Intra-rater Agreement Between Pre-Post Use of the Prototype Intervention (Within instructors)

Kendall's Rank Correlation Coefficient (Kendall's Tau (τ) Coefficient)

According to Kendall's τ the value ranged from 1 (two rankings are identical marked as "="), 0 (two rankings are statistically independent), and -1 (two rankings perfectly disagree marked as "≠"), the test results showed four different cases as Table 18 below.

Table 18 Four cases of user experience with the prototype system

Eval. 1 = Eval. 2	= system	Case 1. Evaluate like a system
	≠ system	Case 2. Adhere to traditional observation (resist technology)
Eval. 1 ≠ Eval. 2	= system	Case 3. Defer to the system (accept technology)
	≠ system	Case 4. Evaluate inconsistently

[a] Evaluation 1 (without referring to prototype system) and Evaluation 2 (with referring to prototype system)

Case 1 indicated that the instructors' evaluation results were similar to the rank ordered by biosensor data (ground truth). Case 2 showed a high level of agreement between Evaluation 1 and Evaluation 2 in ranking the students as opposed to a disagreement between the Evaluation 2 and the system. It means the instructors were sticking to their own criteria to make the judgment, which may be attributed to the instructors' distrust of the computer system. This is a common concern when transitioning to new technology in education (Persico et al., 2014).

In contrast to Case 1 and Case 2, Evaluation 1 and Evaluation 2 showed low levels of agreement. The low level of agreement was attributed to factors affecting the participants' judgments between the two evaluations. If the ranking order in Evaluation 2 was identical or similar to the ranking order of the system as Case 3, it demonstrated that the participants may trust or refer to the ranking orders informed by the system. On the other hand, if there was a low level of agreement across all evaluations different from the prototype system results, it was classified under Case 4, which indicated that there is limited consistency ranking in the level of anxiety among ELLs. To gain further insights into the differences that the instructors experienced while optionally using the system to perceive the students' levels of anxiety, the quotes from individual interviews were added to each case (Section 6.4.3.1 - 6.4.3.4).

6.4.3.1 Case 1: Evaluate like a system

Three participants (P21, P22, P25) were consistent in ranking the anxiety level of the students when assessing the individual final presentation throughout Evaluation 1 and Evaluation 2, and their ranking order in Evaluation 1 and Evaluation 2 was similar to the ranking order presented on the system in the Final presentation evaluation (Eval 1 = Eval 2 = system), according to the result of the statistical test. These participants are all experienced instructors and achieved moderately high emotional intelligence scores (competent to skilled) on perceiving (Branch 1), and understanding (Branch 2) subcategories of the MSCEIT questionnaire, ranging from 106.18 to 107.85 and 101.49 to 118.26, respectively (Table 14). As the experiment progressed, a learning effect was observed. As the participants classified under Case 1 conducted Evaluation 1 (observation) prior to Evaluation 2 (observation with the system), they did not know what ranking order would be shown in the system. However, the results of their Evaluation 1 were moderately similar (.697, .626, .727) to the ranking order shown in the system (Table 19). This learning effect was supported by one of the interview responses as:

"I remember that the system said they (D11 and D12 – DHH ELLs) were anxious. And then even when I wasn't using the system, I rank them as anxious, which I am sorry if I shouldn't have done that…I just saw them and I remembered because it's hard to forget, especially because these students are unique." (P22)

Table 19 Kendall's correlation coefficient (τ) results in Case 1: perceiving anxiety level as the same as the system.

Topic	ID	Case 1	
		Evaluation 1 = Evaluation 2	= Prototype System
News		-	-
Paired		-	-
Final	P21	.002** (.697)	000** (.848)
	P22	.005** (.626)	.004** (.636)
	P25	.000** (.727)	.009** (.576)

[a] significant p-value at p<0.05* and p<0.01**, respectively
[b] τ value in parenthesis

6.4.3.2 Case 2: Adhere to traditional observation (resist technology)

For Case 2 shown in Table 20, the p-values were significant between the Evaluation 1 (observation) and Evaluation 2 (observation with the system), in contrast to the p-values between the Evaluation 2 and system. This indicated that findings from Evaluation 1 and Evaluation 2 were correlated, and a similar ranking order was present on both evaluation forms. As the Kendall's tau values indicated a positive and moderate high correlation between Evaluation 1 and Evaluation 2, as opposed to a weak correlation between Evaluation 2 and the system (ground truth), the participants classified under Case 2 were inclined to adhere to their traditional observation and resisted accepting the ranking order informed by the system.

Table 20 Kendall's correlation coefficient (τ) results in Case 2: adhere to traditional observation and resist accepting the ranking order information on the system.

Topic	ID	Case 2	
		Evaluation 1 = Evaluation 2	≠ Prototype System
News	P1	.009**(.576)	.337 (-.212)
	P2	.014* (.545)	.273 (.242)
	P4	.014* (.545)	.217 (.273)
	P8	.006**(.606)	.131 (-.333)
	P9	.020* (.515)	.493 (-.152)
	P12	.006** (.606)	.411 (-.182)
	P20	.000**(.788)	.681 (.081)
	P25	.040* (.455)	.583 (.121)
Paired	P4	.000** (.848)	.217 (-.273)
	P8	.000** (.758)	.100 (-.364)
	P9	.009** (.576)	.493 (-.152)
	P12	.000** (.727)	.411 (-.182)
	P20	.000** (.727)	.411 (-.182)
Final	P8	.014* (.545)	**.006**(-.604)**
	P9	.000** (.758)	.273 (-.242)
	P20	.000** (1.00)	**.040* (-.455)**

[a] significant at $p<0.05$* and $p<0.01$**, respectively
[b] τ value in parenthesis

Interestingly, P8 and P20 disagreed, and resisted accepting the ranking order information presented on the system. For example, the Kendall's tau values were negatively decreased from -.333, -.364, -.604** for P8 and .081, -.182, -.455* for P20 as the experiment progressed. During the individual interview, the participants classified under Case 2 explained why they could not defer to the ranking order shown in the system. The first reason was that they wanted to completely understand how the system worked, in order to justify that the system presented the correct information at the end. P9 stated that:

"I wasn't sure exactly how things (biosensor and biosensing data) are working, how things are being evaluated. (but I am a fan of the additional inputs like administering my feedback)." (P9).

P4 and P20 pointed out possible errors or missing values in the system or algorithm to inform the degree of anxiety, and contributing factors to the anxiety of learners. P4 stated that:

"I would need to know a lot more about the system to have full trust in it. There is so many things that students could be nervous about…and the system doesn't pick it up. (P4)"

P8 and P9 were very confident with their observation skills regarding detecting anxious students. They believed that their interpretation through observation of body language and gestures would be more accurate than the information shown in the system. They highlighted that they spend the bulk of time talking and interacting with students and felt more assured when they saw how they are reacting. Moreover, if there is a mismatch between levels of anxiety detected from their observations of videos of student performance vs content from the system, they stated that they would rather directly ask questions rather than relying on the system that they could not fully understand its mechanism.

6.4.3.3 Case 3: Defer to the system (accept technology)

No significance was identified between findings from the test examining Evaluation 1 and Evaluation 2 (Table 21). As indicated by the high Kendall's tau values, more than half of the participants deferred to the ranking list presented on the system. They were more dependent on the system when evaluating the *Paired* presentations compared to evaluating the individual presentations with the topic of *News* and *Health*. As the experiment progressed, the participants classified under Case 3 were more likely to defer to the system. As the Kendall's tau values of one, that is a 100% association between the Evaluation 2 and the system, the ratio of fully relying on the rank order of the system increased to 50% (4 out of 8 participants) for the first presentation (individual - News), 63.6% (7 out of 11 participants) with the paired presentation on the topic of Environment, and 88.9% (8 out of 9 participants) with the final individual presentation on the topic of Health.

Table 21 Kendall's correlation coefficient (τ) results in Case 3: defer to the system accepting the ranking order information on the system.

Topic	ID	Case 3	
		Evaluation 1 ≠ Evaluation 2 = Prototype System	
News	P3	.273 (.242)	.020* (.515)
	P5	.337 (-.212)	.000** (1.00)
	P6	.583 (-.121)	.000** (1.00)
	P7	.681 (.091)	.000** (.879)
	P16	.784 (-.061)	.000** (1.00)
	P21	.784 (-.061)	.002** (.697)
	P22	.170 (303)	.003** (.667)
	P23	.170 (-.303)	.000** (1.00)
Paired	P1	.273 (.242)	.000** (1.00)
	P3	.493 (.152)	.000** (1.00)
	P5	.273 (.242)	.000** (1.00)
	P6	.681 (-.091)	.000** (1.00)
	P7	.273 (.242)	.000** (.909)
	P16	.681 (.091)	.000** (1.00)
	P21	.583 (-.121)	.000** (.879)
	P22	.100 (.364)	.000** (.697)
	P23	.493 (-.152)	.000** (1.00)
	P24	.681 (.091)	.000** (1.00)
	P25	.100 (.364)	.009** (.576)
Final	P1	1.00 (0.00)	.000** (1.00)
	P3	.681 (.091)	.000** (1.00)
	P4	0.75 (-.394)	.000** (1.00)
	P5	.583 (.121)	.000** (1.00)
	P6	.337 (-.212)	.000** (1.00)
	P7	.784 (-.061)	.000** (.848)
	P16	.493 (-.152)	.000** (1.00)
	P23	.891 (-.030)	.000** (1.00)
	P24	1.00 (0.00)	.000** (1.00)

[a] significant at p<0.05* and p<0.01**, respectively
[b] τ value in parenthesis

The following interview responses supported the statistical results in Case 3. Some participants (P1, P6, P16, P23) claimed that the system could surpass the human capability to detect the specific instances where students feel anxious when delivering a presentation. P6 reported:

"I felt the system is smarter than me. If the tool whatever monitors they had on their bodies, I feel like it can denote stress and anxiety, and I am going to trust that. I have done a lie detector test myself for a previous job many years ago." (P6)

Additionally, when the participants did not have enough teaching experiences or enough time interacting with ELLs, they seemed uncertain about this ability to perceive the emotions of the students and tended to rely more on the bioinformation displayed on the system. P5 stated:

"I am not a certified teacher. I was a little worried that I was doing it wrong because it is the first time I evaluated them. I was a little thrown. I don't trust my own instincts at this point." (P5)

Moreover, they described that people may interpret learners' anxiety-related behaviors differently, whereas biosensing technology would be unbiased and objective by applying the same conditions to evaluate students' levels of anxiety. A further reason as to why participants were relying on the technology, was the challenge of identifying emotions through facial expressions or behaviors of learners from different cultural backgrounds to the observer (participants in this study), which was mentioned as challenging depending on how well emotions could be masked.

6.4.3.4 Case 4: Evaluate inconsistently

For Case 4, rankings from three participants (P2, P12, P24) showed no significance across Evaluation 1, Evaluation 2, and the system regarding ranking the anxiety level among ELLs (Table 22). Although in the beginning of the experiment, participants P2 and P12 showed similar ranking orders between Evaluation 1 and Evaluation 2 with News presentations by continuously using their traditional method of observation (Case 2), ranking orders across evaluations became inconsistent (Case 4) in the subsequent tasks with the Paired and Final presentations. On the other hand, participant P24 did not show consistent ranking orders (Case 4) between the evaluations in the beginning of the task and the News presentation, but she ended up relying on the ranking order informed by the system (Case 3) and showed consistent ranking orders for the rest of the tasks.

The participants classified under Case 4 indicated either very high or very low scores on the MSCEIT questionnaire. For example, participants P2 and P24 scored high in *Skilled* range in the overall Emotional Intelligence ability and Understanding Emotions section each on the MSCEIT questionnaire. On the other hand, P12 mostly scored high in *Consider Developing* range in all sections of the MSCEIT questionnaire.

Table 22 Kendall's correlation coefficient (τ) results in Case 4: inconsistent evaluation across Evaluation 1, Evaluation 2, and the system.

Topic	ID	Case 4	
		Evaluation 1 ≠ Evaluation 2	≠ Prototype System
News	P24	.681 (.091)	.100 (-364)
Paired	P2	.583 (.121)	.681 (.-091)
Final	P2	.170 (.303)	.891 (.030)
	P12	.217 (.273)	.273 (.-242)

[a] significant at p<0.05* and p<0.01**, respectively
[b] τ value in parenthesis

It was also interesting to identify the reasoning as to why participants classified under Case 4 lacked consistent evaluations across the tasks. From reviewing the interview responses, P2 described being confident with her own teaching. She described interpreting the raw physiological data using her own methods, instead of solely relying on her observation or following the filtered bioinformation (e.g., numerical degree of anxiety) by the system. She stated that:

"I would be able to tell who is engaged and motivated during class…So, I guessed the numerical values (degree of anxiety) from the raw physiological arousal graph." (P2)

On the other hand, P12 and P24 claimed the difficulty in comparing students' levels of anxiety in both Evaluation 1 and Evaluation 2.

"I assumed that as this study went on, my way of assessing whether with or without the prototype system would settle in and become more similar, and it didn't seem to happen that way. I thought each week, whether with or without the prototype system, it was hard to know how to compare the students' mannerisms." (P12)

"[As time went by] I kind of lost trust in my own ability to recognize emotions. Maybe I shouldn't have done it this way, but when I look deeper into the system and look for the signs in the body language, and the way they (ELLs) are presenting, I decided to rely on the system. I feel like relying on the physiological information. It's…it's science, right?." (P24)

6.4.3.5 Change of perspectives towards biosensing technology over time

At the beginning of this study, the same number of participants were found in Case 2 and Case 3 (n=8, n=8). However, as the experiment went on, more participants described changing their attitude regarding accepting the ranking order presented on the system. The participants who kept using the traditional observation method (Case 2) had shifted their views on accepting and relying on the ranked order informed by the system.

As Table 23 shows, the number of participants who were classified under Case 2 gradually decreased from 47% (News presentation) to 29% (Paired presentation), and 17.6% (Final presentation). In contrast, the number of participants who were classified under Case 3 did not change much, but it was interesting that there was a higher tendency to rely on the system in the evaluation with the Paired presentation as opposed to the evaluation with the individual presentations (e.g., News and Final presentations).

Table 23 Change of perspectives towards accepting the information provided by the system

ID	Individual News		Presentation Paired Environment		Individual Health
P1	Case 2	→	Case 3	→	Case 3
P2	Case 2	→	Case 4	→	Case 4
P3	Case 3	→	Case 3	→	Case 3
P4	Case 2	→	Case 2	→	Case 3
P5	Case 3	→	Case 3	→	Case 3
P6	Case 3	→	Case 3	→	Case 3
P7	Case 3	→	Case 3	→	Case 3
P8	Case 2	→	Case 2	→	Case 2
P9	Case 2	→	Case 2	→	Case 2
P12	Case 2	→	Case 2	→	Case 4
P16	Case 3	→	Case 3	→	Case 3
P20	Case 2	→	Case 2	→	Case 2
P21	Case 3	→	Case 3	→	Case 1
P22	Case 3	→	Case 3	→	Case 1
P23	Case 3	→	Case 3	→	Case 3
P24	Case 4	→	Case 3	→	Case 3
P25	Case 2	→	Case 3	→	Case 1
Case 1	0/17 (0%)		0/17 (0%)		3/17 (17.6%)
Case 2	8/17 (47%)		5/17 (29%)		3/17 (17.6%)
Case 3	8/17 (47%)		11/17 (65%)		9/17 (53%)
Case 4	1/17 (6%)		1/17 (6%)		2/17 (11.8%)

*Case 1: Evaluate like a system
Case 2: Stick with observation (resist tech.)
Case 3: Follow system (accept tech.)
Case 4: Random, Inconsistent

Although some participants who were classified under Case 2 transferred to Case 3 as the experiment progressed, three participants (P8, P9, and P20) kept resisting accepting the information provided by the system and adhered to their traditional observational methods throughout the whole experiment.

6.5 Discussion

This study examined the efficacy of using the prototype system to perceive the level of anxiety encountered by students. Findings from the statistical tests confirmed low agreement in ranking the levels of anxiety when ELLs were delivering presentations. The subsequent statistical tests and interview responses representing those four cases in Table 24 revealed where the low agreement stemmed from. The findings of each case demonstrated how participants go about assessing the anxiety level of learners. Four cases were identified: (ranking the anxiety level of ELLs similar to the ranking order presented on the system (Case 1), adhere to traditional observation (Case 2), defer to the system (Case 3), inconsistent evaluation (Case 4)) and how those perspectives had changed over time during the experiment.

6.5.1. Need for Emotional Rubrics

As noted in Table 14 and Table 15, the statistical test showed low agreement in ranking the anxiety level of ELLs among participants (novice and experienced instructors). Researchers (Munn, 1940; Parthasarathy et al., 2016; Shields, 1984) have stated that low levels of agreement when assessing the emotions of individuals can be attributed to the fact that emotions are very complex and subjective to evaluate. Low agreement among instructors was also found in the evaluation of the students' academic performance (Goldhaber et al., 2015). ESL/EFL instructors usually develop their own rubrics and checklists, and share it with other instructors/co-instructors to grade their students' academic performance in a consistent manner.

Similarly, the prototype system used in this study can play a role similar to a formal rubric, to support instructors when attempting to assess emotional states of ELLs. This in turn would lead to greater levels of consistency (i.e., if students are being taught by more than one instructor and exhibit early signs of anxiety in their classes, instructors can modify their teaching at those early points with help from the system). The numeric values of physiological data can be used as an objective index to refer to when the instructors cannot come to consensus when assessing the emotional states of the students. Moreover, at the beginning of the semester when the rapport between the instructor and the students has yet to be established, this prototype system can be employed to understand the learners' emotional states while observing their behaviors and having conversations with them. The interview statements described in this chapter (Section 6.4.3.3) also supported that the prototype system developed, can be useful as guidance for novice instructors, should they have limited experience teaching ELLs from diverse cultural backgrounds. Participants stated that the prototype system can be useful to uncover the inward emotional states of the ELLs including DHH ELLs and assess their anxiety level through a visual representation on the screen. Although this study shows the potential benefit of using the prototype system to help the instructors identify anxious ELLs and emotionally/instructionally scaffold those learners accordingly, a threshold needs to be established to indicate what point the system should inform the instructors to intervene (i.e., modify instruction to address learners' anxiety). Since 'facilitating anxiety' motivates the learner to 'fight' the new learning task, whereas 'debilitating anxiety' demotivates the learner to 'flee' the new learning task

(Scovel, 1978), further work is needed to investigate the threshold of anxiety where learners continue to persevere with the task..

6.5.2 Blind faith vs. skepticism on new biosensing technology

The findings showed four cases relating to participants' perspectives on whether to openly accept or resist the bioinformation presented via the system. Unlike prior work, this study invited ESL/EFL instructors for the first time, who would be actual end-users of the biosensing technology, as the participants in the experiment, and investigated how they embraced the new technology compared to their traditional method of observation.

When there was a sizable gap in ranking order between what the participants observed and the bioinformation presented by the system, the findings from participants could be classified into exhibiting blind faith or skepticism regarding the innovative biosensing technology.

The larger the gap, the more dependent the participants were on content presented via the system to make assessments of emotional state. Those who considered that the system measuring biometric information outperforms human observation, tended to easily shift their judgment, and blindly rely on the system. Some participants blamed themselves for their lack of ability to perceive the emotional states of ELLs. They attributed this to lower levels of experience teaching and lower levels of intercultural communicative competence (ICC). This aspect of blind faith in technology should be noted because it can lead users to passively accept the information presented via a system. By over-trusting a system, the instructors may neglect to actively observe students during teaching.

On the other hand, when the difference between the rankings in the observation and the system was large, some participants resisted to adopting the information presented by the system. They suggested that the biosensor may have malfunctioned while collecting the physiological information of the learner, or a data flaw may have been present, caused by the learners' frequent hand movements/hand gestures when presenting. Concern was also expressed that the system was only evaluated with a small number of students, which made them uncertain about relying solely on this data. In addition to the possible defect in technology, the participants said that the reason for taking such a skeptical point of view in accepting information presented via the system was because they could not trust it until they fully understood the principles and mechanism of the biosensing technology. They suggested that professional development training was needed from experts in the field of neuroscience, cognitive-behavioral psychology, and education. They wanted to understand how the nervous system of each student is connected to cognition and behavior when learning in class, how the biosensor collects physiological responses of users, and the steps taken to present processed bioinformation via the system.

6.5.3 Design implication - complementarity between raw and processed data for self-adaptive systems

When designing a system relating to personal bioinformatics, prior studies have focused on delivering big data and computational results via systems (Singh et al., 2021). However, there is growing consensus that ethical principles should be considered to support public good. These include data transparency and explainability (Weyns, 2020). In order to reduce bias and ensure fairness when presenting data related to emotional state, the prototype system developed in this study provided both raw and processed physiological data of learners to help instructors make informed decisions. It was interesting to see how the participants retrieved information from both raw and processed data on the system to perceive the emotional state of the learners throughout the experiment. With the raw data presenting prominent peaks of the physiological arousal graph, some participants tried to find environmental factors, such as an uncomfortable classroom climate rooted in different cultural values and learning/teaching styles that caused students' tension during the presentation performance. For example, students hailed from individualistic societies (e.g., U.S.) are known to speak up in class, whereas students from more collectivistic societies (e.g., China, Japan) tend to speak only when called upon (Phuong-Mai et al., 2005). By considering the learners' cultural backgrounds and the actions/events that happened prior to and after the peaks, the participants were comparing and attempting to attribute meaning to what they observed from the students' behaviors. However, not all participants were able to identify meaning. When the physiological data showed a steady level of arousal without any prominent peaks, the participants did not know how to best interpret the data.

In parallel with being able to access raw data, participants also appreciated being able to see the processed data in order to view the degree and types of anxiety among students in class. They liked the fact that they could learn more about the students' strengths and weaknesses by monitoring their progress across presentation topics over time. Moreover, the types of anxiety including PSA and FLA, were identified as useful data points for purposes of teaching. This would enable them to prepare personalized class activities for ELLs. They also described the benefit of using the processed information to quickly aid them in making a decision about modifying instruction and thought that this would be helpful for novice instructors who might not be familiar with perceiving learners' emotions. Another interesting finding was that the participants who appeared to resist technology were dismissing the processed information on the system, but they partially referred to the physiological data to compare it with judgements made through traditional methods of observation.

The findings in this study demonstrated the importance of presenting both raw and processed data to provide the information necessary to support decision making.

6.5.4 Validity of assessment

Self-report assessments are commonly used as measures of ground truth, as they provide an insight relating to emotional and mental state by the users themselves (Di Lascio et al., 2018; Gao et al., 2020; Wang et al., 2014; Zhang et al., 2018). In addition to self-report surveys, behavioral annotation from observations (including facial expressions and body gestures) have been used by researchers to determine an emotional state of a student (Petrovica et al., 2017). The latter approach has been widely adopted in the affective computing research community, where a variety of emotional expressions and interaction dynamics can be annotated (Metallinou & Narayanan, 2013).

However, self-report assessments and behavioral annotations can be subject to bias. Challenges to accuracy can be faced, as these methods are conducted post-interaction and take time to implement. Moreover, difficulties can be encountered establishing an agreement metric among multiple annotators (McKeown et al., 2011; Nicolaou et al., 2011). Gao et al. (2021) have described the risks associated with using annotated data as the ground truth for recognition of emotion. The researchers have proposed the use of physiological signals as an objective measure of student engagement in learning. Other researchers have also suggested that EDA can be a more useful index to monitor an individual's emotional state (Braithwaite et al., 2013; Choi et al., 2019; Picard et al., 2016) as EDA offers a continuous, objective, and unobtrusive method to detect time-varying nature of human emotion in real-time in (Egger et al., 2019; Kritikos et al., 2019; Poh et al., 2010).

The study described in this chapter, used the physiological dataset of learners as the ground truth dataset, to compare the degree of anxiety in presentations. However, future work could examine using multiple methods to validate the dataset. These could include self-report questionnaires, interviews, and behavioral annotations.

For the emotional assessment, the EDA data (e.g., state-based assessment) presenting objective numerical data will complement the cognitive bias and subjectivity resulting from learners' self-reported measures of emotions (e.g., student-based assessment) during the retrospection (Dewaele & Li, 2020). In addition, referring to the EDA data of learners will minimize educators' discordance of evaluating students' emotional states. Until recently, educators have evaluated students' emotional states by traditional methods of observation, so the emotional states have been recognized differently depending on the educators' familiarity with the learners' cultures and the interpretation of the learners' facial expressions and behaviors. To develop a consistent pedagogical content in a curriculum, it is necessary to have instructors' reach consensus (McCaughtry & Rovegno, 2003). Perceiving the degree of anxiety among students in the same manner may address the issues in assessing the emotional states of the ELLs differently.

6.6 Limitations and Future Work

The spread of the COVID-19 virus limited the ability to conduct in-person experiments. If the participants were able to use the prototype system and perceive the emotional states of their own students in class throughout the semester, it is possible more detailed insights could have been gleaned.

This dissertation has mainly explored the perspectives of the educators on the efficacy of the prototype system using learners' bioinformation to perceive their emotional states and provide personalized emotional and instructional scaffolding. However, a further series of experiments are needed to investigate if the instructors' customized scaffolding with the aid of the prototype system is helpful to learners. This can be done through interviewing the learners and monitoring their learning progress throughout the semester.

6.7 Summary

This chapter described a study investigating how ESL/EFL instructors applied the prototype system using biosensing technology to perceive and rank the anxiety level of the ELLs. Four perspectives were identified (termed: 'cases'), relating to how bioinformation was used. Findings led to the development of design guidance and discussion of ethical considerations when using biosensing technology in an ESL/EFL context.

6.7.1 Developing four types of personas in an ESL/EFL context

The mixed methods approach adopted in this chapter allowed the investigator to develop a framework of four types of potential end-users who use biosensing technology for teaching ELLs. Findings from the quantitative analysis show how similar (Case 1) or different (Case

4) between the results from the participants' observational evaluation of students' anxiety (Evaluation 1) and the bioinformation in the system. Moreover, the comparison between the correlation coefficient values of Evaluation 1 and Evaluation 2 illustrates two opposing personas of those who have the perspective of technology resistance (Case 2) and technology acceptance (Case 3). Conducting interviews with each participant provided more detailed perceptions of biosensing technology, along with more reasoning behind why bioinformation was used to its full capacity to make judgements on levels of anxiety. The interview responses also provided a deeper view of possible issues using biosensing technology in ESL/EFL class including ethical concerns (e.g., social stigma, degree of intervention).

6.7.2 Importance of designing a system with both raw data and processed information

Findings from the study confirmed the importance of designing a system presenting both raw data and processed information of the ELLs. By providing an option to the end-users in terms of referring to the prototype system, this study found that the participants appropriately examined the raw data of learners' physiological data and the processed data (e.g., degree of the anxiety, types of anxiety) presented via the system to make judgements about levels of anxiety. The participants who resist biosensing technology (Case 2) tended to meaningfully interpret the raw data instead of relying on the processed data, whereas the participants in favor of the system tried to learn about the emotional states of the ELLs by comparing their judgments made from observing students with the processed data. As the raw EDA data was obtained directly from the participants, some participants were less reluctant to accept the raw data even though they were doubtful about biosensing technology. On the other hand, the participants who relied on the processed information of learners favored the fact that they glance at the screen to glean information to make a decision on levels of anxiety. They suggested that the system presenting processed information would offer benefits to instructors handling larger sized classes. It could also aid novice instructors who may have low levels of data literacy skills.

Chapter 7. Contributions and Discussion

7.1 Chapter introduction

In this chapter, the four main contributions are described. Design guidance for an educational system using EDA data within an ESL/EFL environment, is presented. Guidance is based upon findings from the studies described in this dissertation, covering ways to link assessment with EDA data, and the ethical implications of using this data.

7.2 Summary of significant contributions

7.2.1 Demonstrating the feasibility of using electrodermal activity (EDA) to measure the emotional state of English language learners when delivering presentations

Wearable devices, such as wristbands, have become a popular tool among researchers in HCI. These offer a way to non-invasively measure the psychological states of users (Babaei & Tag, 2021). While studies have examined their feasibility for gaining emotional information, only a few studies (e.g., Gashi et al., 2018) have been conducted outside of a controlled environment (e.g., laboratory). EDA data collected from a wearable device is known for being very useful in a multitude of international settings and cross-cultural studies. The data can provide an objective means of determining someone's psychological state, in contrast to the alternative of gaining qualitative feedback where linguistic and cultural problems may occur in language translation (Charles et al. 2017). Moreover, users cannot control their levels of physiological arousal at their own will, whereas facial expressions and behavioral gestures can be masked and controlled. Despite the potential value of using EDA data in cross-cultural studies (Bartolomé-Tomás et al., 2020; Kaneko et al., 2021; Torrico et al., 2019), researchers have yet to demonstrate the feasibility of using EDA with ELLs in an ESL/EFL classroom. Thus, in this dissertation, biosensing technology was used to determine whether it could offer promise to instructors. Findings provided empirical evidence that EDA data presented via an interface can help educators visually recognize a student's emotional state while delivering a presentation in class. In addition, the system notifies educators regarding the intensity and types of anxiety faced through the processed EDA data. Perceiving the learners' emotional states and providing emotional support has been considered an essential and desirable skill for educators to have (Greenberg et al., 2017; Quezada et al., 2020; Vandenbroucke et al., 2018). The studies described in this dissertation have demonstrated the feasibility of using EDA data to interpret emotional state.

Reproducibility has been a key element in good research (Engel et al., 2017) as it accelerates scientific discovery that researchers should surely do better both within a laboratory and beyond, especially in an era of interdisciplinary science where people wish to apply methods that they are not experts in (Garijo et al., 2013). To achieve reproducibility in terms of using EDA technology in an educational field, the studies described in this dissertation reused the same biosensing instrument (e.g., E4 wristband) used in a prior study. The same method of analysis (Continuous Decomposition Analysis (CDA)) was used, which is yet to be explored by other researchers in the field of education. It allowed the investigator to examine the feasibility of using EDA data, and in a setting where the experimental tasks (e.g., presentation topics and guidelines) were created by the instructor as a part of the class (i.e., a high-stake activity). While prior studies using EDA technology have been conducted in a controlled environment, the findings of this study have shown that EDA technology offers promise to a classroom setting with learners hailing from diverse linguistic and cultural backgrounds. These findings will add to the body of knowledge relating to areas of affective computing, biosensing technology, and even the field of ESL/EFL education.

7.2.2 Development of an algorithm for classifying degree of anxiety and identifying predominant type of speaking anxiety among English language learners

Findings from Study 1 and Study 2 (described in Chapter 3) led to the identification of EDA features that enabled the investigator to classify levels of anxiety among ELLs. Students were classified into two groups: high levels of anxiety and low levels of anxiety. The learners' predominant type(s) of anxiety (e.g., PSA, FLA) was also identified. A machine learning algorithm was developed to conduct this classification. K-means clustering was used to group students into higher and lower levels of speaking anxiety. Findings showed the potential for the development of a reflective system where students and educators can monitor the emotional states of a presenter and can actively work towards addressing anxiety if it is impacting performance. In Study 2, ten features from the phasic and tonic components of EDA (e.g., mean, standard deviation, minimum and maximum values in a component, locations of minimum and maximum values, mean peak amplitudes, number of peaks, slope, and area under the curve) and the time-frequency features of EDA signals extracted from the Hilbert Huang Transformation method combined with the gradient boosting tree (GBT) classifier, were outperformed in predicting the predominant types of speaking anxiety among ELLs. Furthermore, a modified Locally Interpretable Model Explanation (LIME) method (cite) was adopted to identify the subset of features that are most important for modeling predictions of different types of anxiety among ELLs. The top three supporting input features of the GBT algorithm identified in this study contributed to the prediction of the predominant type of speaking anxiety among ELLs, which is the first attempt to disentangle the interchangeable anxiety through a machine learning approach. These findings may encourage both ELLs and language instructors to be aware of the origins of learners' anxiety subtypes, and it may eventually lead to the development of personalized practice for foreign language skills and emotional intelligence skills.

7.2.3 The development of an approach to investigate methods to measure levels of anxiety among ELLs through the use of a biosensor, and to design and evaluate an educational prototype

Researchers suggest that there are few design methods that have had educators participate from the beginning of the process when developing an educational system (Carroll & Rosson, 2013; Hansen & Iversen, 2013). Most researchers often invite educators to evaluate a system once it has been developed, to gain their input (Gupta & Pathania, 2021; Kakoulli Constantinou, 2021). Furthermore, the existing prototype systems introduced in prior studies are mainly designed for students who are directly wearing a biosensor and accessing their bioinformation (Di Lascio et al., 2017; Gashi et al., 2018). These systems aim to help the students monitor their own physiological changes through a system in real-time, with a view to using information presented by the system to better perform in learning and decision-making tasks. In addition to students' self-monitoring of emotional changes, this dissertation noted the importance of educators' emotional support for high quality instruction (Claessens et al., 2017; Pianta et al., 2008). However, few systems have been designed to help instructors support the learning experience, and provide the emotional support needed by referencing the students' levels of physiological arousal. Therefore, the latter part of this dissertation (Chapter 4 to Chapter 6) describes an adaptation of the Design Thinking Process (DTP) to encourage in-service ESL/EFL stakeholders to actively participate in developing a teaching aid. Their vision on using biosensing technology to support teaching and learning could be shared with potential designers and researchers. As the participants were either ESL students or worked in EFL instruction, features suggested by participants were found to be different from those seen in other biosensor-based feedback solutions. For example, functionality was desired to support language assessment. This included grammatical errors, usage of new vocabulary, and identifying accurate use of speech, which were highlighted as important aspects to detect on the system in relation to the emotional states of learners. This was because it provided some interpretable evidence for educators to identify students' most vulnerable moments in foreign language learning. Furthermore, the instructors can reflect on whether the content covered in class and their teaching technique can be modified to better support students who exhibit anxiety. They can encourage fluency over accuracy, until ELLs become more confident when using English. This can be achieved by practicing the same content when abruptly high levels of EDA arousal are detected. Unlike previous systems that have mainly focused on assessing public speaking skills (e.g., Chen et al., 2014; Schneider et al., 2015; Tanveer et al., 2015), or virtual reality therapy interface interventions to reduce public speaking anxiety (e.g., Bickmore et al., 2020; Chollet et al., 2018; Murali et al., 2021), the approach described in this dissertation addressed the relationship between learners' emotional states and foreign language performance, through the use of EDA technology.

7.2.4 Developing design guidance for developing an educational system using EDA data in an ESL/EFL environment

The following recommendations were developed based upon insights from the studies conducted, along with suggestions from participants.

7.2.4.1 Link foreign language assessment and EDA arousal

This was one of the first initiatives to apply EDA technology to an ESL/EFL environment, where students from different cultural and linguistic backgrounds were taking English language classes. Through the study described in Chapter 4, instructors revealed that their main interest was to find out the causal relationship between students' emotional states and their foreign language performance. The instructors wanted to use biosensing technology to detect any correlation between the levels/types of anxiety and foreign language performance. For example, they expected that the system informs the user when ELLs need emotional and instructional support. Suggestions included the system detecting incorrect parts of speech, frequent grammatical mistakes, and identifying a limited range of new vocabulary, which would signal that they were nervous. Systems using EDA technology in the past have been mainly designed for users wearing a biosensor to monitor their own bioinformation and interpret their emotional states (Howell et al., 2018). However, this dissertation has focused on the perspective of educators who can identify state from the physiological information collected, and scaffold the learners in terms of foreign language learning.

7.2.4.2 Present both raw EDA data and processed EDA information on the system

Findings from the study described in Chapter 6 led to the classification of educator profiles. Four profiles (termed: 'personas') were developed. The personas facilitate a common understanding of target users, mapping the users' archetypes with their goals and needs in digital products or services (Nieters et al., 2007), each of which represent the characteristics of ESL/EFL instructors using EDA technology in classroom settings. These four representations include: an educator who perceives emotions of learners same as the biosystem detects (Case 1); an educator against using EDA technology and would rather adhere to an individual's traditional observation skills to recognize the students' emotions (Case 2); an educator who highly trusts and relies upon the information presented by the system (Case 3); and an educator who does not take a consistent stance on ways to perceive learners' emotions (Case 4).

To support educators falling into the categories of Case 2 and Case 3, design guidance is needed to have a more balanced view of the value of EDA presented via a system. The participants classified under Case 2 had a higher tendency to distrust the processed EDA data presented on the system because they felt that the instructors' observational abilities are more accurate or meaningful, compared to using the processed EDA data for this purpose. For the persona who has this point of view, the system should be designed to present raw EDA data in addition to the processed EDA information. As these users are concerned about whether researchers might intentionally misinterpret the EDA data when preprocessed, and fabricate or falsify data (Miyakawa, 2020), it is critical to present them with the raw EDA data so that they can draw their own conclusions with a physiological state-based assessment. The physiological changes in every moment associated with the learner's presentation performance can help instructors understand contributing factors to the learners' emotional states by comparing it with observation. The instructors can connect high peak or rapid changes found in the EDA arousal to infer the environmental factors (e.g., interaction with classmates, presentation format), and cognitive processing required to formulate language (e.g., word choice, grammar usage, pronunciation, topic/content of the presentation). In this regard, instructors can balance out their traditional assessments on students' anxious behaviors and EDA state-based assessment to perceive and evaluate the emotional states of the ELLs to improve their teaching.

Moreover, the design should create an environment where educators can continuously evaluate students in a more convenient and efficient way. For example, the instructors had to open video files of each student's presentation, watch, and evaluate them one-by-one. To reduce the time burden, the system should be designed to present a list of all students in the class and once a student's name is selected, the system should offer the ability to view the corresponding video of the student delivering a presentation. As individuals who are more ambivalent about relying on new technology (i.e., Case 2) tend to adhere to traditional teaching methods (i.e., focusing on observing facial expressions of students to determine emotional state), the design should highlight the benefits of using the system.

In addition, raw EDA data should be presented for educators classified under Case 3, who are more likely to refer to the information presented by the system. To prevent instructors from undermining their own judgment, by relying too much on the information offered by the system, the raw and processed EDA information should be delivered in concert.

7.2.4.3 Provide ethical implications for using EDA data in an ESL/EFL class

The ethics underlying using video recordings to support decision making have been widely discussed by researchers (Mackay, 1995). The necessity of ethical guidance for neurotechnology has been described in a range of prior studies (Goering & Yuste, 2016). However, there has been a paucity of discussion relating to ethical guidance for foreign language educational systems using EDA data. In this dissertation, ethical guidance was developed for the context of an ESL/EFL class based on the findings from the interviews with ESL/EFL instructors who monitored students' emotions through EDA technology.

Suggestions from instructors included that information relating to learners' emotional states should ideally be used for formative assessment rather than for summative assessment. If a system showing the emotional state of students is used for purposes of grading the outcome of an instructional unit (e.g., summative assessment), it bears a high risk of being misused or abused. So, the instructors may quickly judge student performance based on the results provided by the system, rather than taking time and scrutinizing the progress of each student in a formative way. The capabilities of humans and technologies should be in a relationship that complements each other's weaknesses. Technology should not infringe on an individual's capabilities.

Secondly, the need for preserving privacy, confidentiality, and security of the EDA data should be prioritized. EDA technologies constantly generate data which is stored within a system. According to Fairclough (2014), physiological information gathered should be owned by that person, and by default, this information should be confidential in a similar way to medical records. To address fragility in educational data privacy, Amo et al. (2019) proposed Personal Data Broker, which is a cloud-based solution that students can take control of and manage their own data, and decide when and who can create, read, update, and delete actions. Thus, an educational system containing physiological information should ensure students privacy and authority to access and manage their data. Moreover, the guidance should be transparent, and easily inform students of how their data is used in class. In particular, the learners discussed in this dissertation are English language learners (ELLs) who may have a limited understanding of the concept of data privacy and protection written in English. To address this, the guidance should be developed in a range of languages, suitable for the different linguistic backgrounds that ELLs come from.

7.3 Implications

7.3.1 Need for a Paradigm Shift

A range of studies have focused on ways to address levels of anxiety through the development of systems using biosensing technology to detect emotional states (Kritikos et al., 2019; Pali et al., 2021; Zheng et al., 2016). Anxiety often negatively impacts academic performance. Thus, researchers have been focusing on ways to accurately detect and measure levels of anxiety, aimed at developing interventions to address these using biosensing technologies (Alfaras et al., 2020; Lakshmi et al., 2017). Interestingly, some of the participants in this study perceived low levels of anxiety as a necessity for students, as it facilitates a stronger performance in class. Scovel (1978) stated that 'facilitating anxiety' motivates students to 'fight' the new learning task, whereas 'debilitating anxiety' stimulates the learners to avoid the challenge. In the studies described in this dissertation, participants provided insights on ‘facilitating anxiety’, which encouraged the investigator to think outside the box when developing a prototype system using biosensing technology. The studies described in this dissertation have focused on the benefits of using a system to detect high levels of anxiety among students, and how this data can be used to support teaching. However, there may be value to highlighting periods when students are emotionally stable or ‘positively’ anxious, rather than pinpointing their most debilitating moments. As discussed in Chapter 6, some participants said they would use this system to better understand and learn about their students, particularly those who hail from cultures which encourage individuals to keep their emotions masked. In this context, the system can be used to both notify instructors about specific instances of anxiety, but also inform users as to when a certain amount of anxiety encourages the students to perform at their full capacity. This may allow the instructors to learn about different strengths and weaknesses of individual learners and evaluate the outcome of their teaching through monitoring levels of facilitating and debilitating anxiety among learners.

7.3.2 Freedom of Choice to Avoid Response Bias

When it comes to studies using an intervention to evaluate its effects before (pre-) and after (post) usage, the danger of response bias can be involved with participants when conducting usability studies (Dell et al., 2012). Participants may feel some level of pressure to only provide positive feedback on evaluations of technological artifacts to please the researcher (Vashistha et al. 2018). This makes it challenging for many researchers to receive constructive and negative feedback from participants, which would help them to improve their designs or interventions (Anokwa et al., 2009; Ho et al., 2009).

To elicit honest and critical feedback from participants in the studies described in this dissertation, participants were provided with some element of choice when conducting tasks within studies. For example, in the study described in Chapter 4, participants were able to select any type of platform (e.g., desktop, tablet, mobile, smart watch screen) which they either felt most comfortable with, or thought would be most practical in a classroom environment. They could then sketch their educational technologies to help instructors identify and support anxious students in class. As a result, this inspired some level of creativity in designs (e.g., P7 - an arm patch diffusing aroma to anxious students, integrating technologies including VR and automatic speech recognition with biosensing technology (P1, P4, P5)). Some participants combined two platforms together when sketching, to show how different devices could be used to interact with the system.

Another way to minimize response bias from participants related to making the intervention optional for participants to use (a prototype system informing the degree of anxiety among students in order) described in Chapter 6. Participants could take the information relating to emotional state provided by the intervention, partially refer to it, or not use it and rely on their observation skills instead based on their preference and attitude towards biosensing technology. For each task, the participants could decide how much they want to rely on the information presented via the system. As a result, the investigator was able to develop four types of personas including participants who exhibited a negative attitude towards biosensing technology. Those who did not want to use the intervention during the task had strong levels of confidence in their teaching skills. They also tended to believe that they have a very close relationship with students, so supplementary assistance from biosensing technology was unnecessary. Observation of facial gestures would be enough to determine emotional state.

7.4 Limitations and Future Work

The students who participated in this study described wanting to wear the biosensor over an extended period of time to monitor their physiological changes. This would allow them to be conscious of changes in emotional state, and learn to react appropriately. However, the cost associated with biosensors was a limiting factor. Conducting an extended field trial using biosensors would be a natural extension of the work, as it would help shed light on factors contributing to PSA and FLA faced by ELLs outside of delivering presentations. Findings would also help instructors use the information to scaffold ELLs.

Although the sample size of ELLs (n=19 in Study 1 and n=33 in Study 2) and educators (n=9 in Study 3 and n=19 in Study 4) was small, this dissertation demonstrated the feasibility of using EDA technology in an ESL/EFL environment. The findings from these studies can serve as a cornerstone for further research and it can be applied to a wide range of people who would feel anxious to overcome the barriers in language and culture. Another logical step to extend the work would be to examine the merits of the solution with a wider, more diverse sample. As an example, the classes surveyed as part of this research included students with hearing disabilities. Challenges can be faced when delivering a presentation if working in concert with an interpreter, as it may be unclear to the instructor as to whether anxiety is being shown during the signing process. EDA can offer a data point to help the instructor identify how best to support students in this process.

Chapter 8. Conclusion

8.1 Summary of Research Questions

In this dissertation, the feasibility of using EDA technology to support ESL/EFL instructors was investigated. The research has centered on the following question: "How does providing awareness of the emotional states of ELLs, through the use of a biosensor, impact teaching?". To address the question, a structured set of studies have been undertaken. In particular, the feelings of anxiety which are frequently experienced by English language learners while delivering a presentation in class were investigated (sub-research questions in Chapter 3).

Study 1:

- RQ 1-1: Which features extracted from electrodermal activity (EDA) data collected when presenting in class can be used to classify levels of anxiety?
- RQ 1-2: Which non-verbal behaviors made when presenting correspond to EDA arousal (peaks), indicating anxiety?

Study 2:

- RQ 2: Can PSA and FLA be classified using EDA data among ELLs when delivering a presentation in English?

The degree and predominant types of students' anxiety characterized by the EDA technology (described in Chapter 3), led to the development of questions relating to how instructors would use this information in connection with their existing teaching methods. Using an adapted version of the Design Thinking Process (DTP), the investigator examined how instructors identify emotional states among students in their classes, and modify their teaching to support anxious learners. In Chapter 4, a focus group interview was conducted. The sub-research questions addressed in the focus group interview included:

Study 3

- RQ 3-1: Which traditional methods have been used by experienced EL instructors to identify the emotional states of learners in class?
- RQ 3-2: How did/would the awareness of learners' emotional states impact instructors' teaching?

The findings from the focus group interview led to the design of Study 3. The aim was to introduce biosensing technology to this educational environment. In the following stage, the investigator and instructors investigated ways to integrate classroom activities with the new biosensing technology that could complement instructors' existing teaching methods. The following research question was developed:

- RQ 3-3: How would the experienced EL instructors integrate biosensor-based feedback of learners' speaking anxiety into a prototype system?

Sketches of systems integrating biosensor-based feedback developed by participants provided a blueprint for developing an educational prototype (described in Chapter 5). To evaluate the efficacy of the prototype system, a comparison study was undertaken. Participants were asked to determine emotional states using their own methods (e.g., observation of facial gestures) and using the prototype (Chapter 6). The following research questions were developed:

Study 4

- RQ 4-1: Assessment between raters

  Is there agreement between instructors on ranking learners' degree of speaking anxiety

  - **without** using the biosensing data on the prototype and solely relying on observation (Evaluation 1)
  - **with** referring to the biosensing data on the prototype (Evaluation 2)?

- RQ 4-2: Assessment by raters

  When it comes to ranking the learners' degree of speaking anxiety, are there any similarities or differences between Evaluation 1, Evaluation 2, and biosensing based ranking data on the prototype (ground truth)?

  - Compare Evaluation 1 and Evaluation 2
  - Compare Evaluation 1 and ground truth
  - Compare Evaluation 2 and ground truth

- RQ 4-3: How does the instructors' awareness of emotional states of ELLs vary over the experiments?

Interviews were conducted to explore in more depth how novice and experienced ESL/EFL instructors would apply the EDA technology to their teaching.

- RQ 4-4: How would instructors use information presented via a system relating to the emotional states of students in terms of teaching strategy or class activities?

8.2 Summary

The main research question of this dissertation related to "how does providing awareness of the emotional states of ELLs, through the use of a biosensor, impact teaching?". The contributions of this research include: (1) the demonstration of the feasibility of using electrodermal activity to measure emotional states of ELLs during a classroom presentation; (2) the development of an algorithm for classifying degree of anxiety and predominant type of speaking anxiety among the learners; (3) the development of an approach to investigate methods to measure levels of anxiety among ELLs through the use of a biosensor and to design and evaluate solutions to support educators using this technology; and (4) design guidance for an educational system using EDA data within an ESL/EFL environment.

The series of studies conducted in this dissertation allowed the investigator to develop an educational system for instructors using biosensing technology to perceive the emotional states of ELLs. The investigator aimed to present the instructors with objective and unbiased information relating to ELLs, which could provide them flexibility to interpret students' physiological data, in conjunction with information gleaned from observing students within classroom activities. In order to do this, both quantitative and qualitative data were presented within the system ensuring that the weakness of one type of data can be offset by the strengths of the other. The system provided all time stamped data points relating to physiological changes identified. These aided instructors to associate them with the emotional states of students while delivering a presentation. Overall, the combination of more than one data source, and information provided in the system, enabled the instructors to interpret emotional states of ELLs. They could then make a decision on how to modify their instructional style to better support anxious students, helping to address the central research question.

Appendices

Appendix A-1. Public Speaking Anxiety (PRPSA) Questionnaire

ID: ____________________

Thank you very much for participating in this study. The following is a questionnaire concerning your level of anxiety experienced in the language class, particularly when delivering in-class presentations. This questionnaire is the Personal Report of Public Speaking Anxiety (PRPSA) which has been adapted from the work conducted by McCroskey (1970). Please read each statement carefully and indicate the extent to which you agree or disagree by circling your choice on the five-point scale. The results of this survey will be used only for research purposes. Therefore, please be honest with your responses.

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
1. While preparing for giving a speech, I feel tense and nervous.	1	2	3	4	5
2. I feel tense when I see the words “speech” and “public speech” on a course outline when studying.	1	2	3	4	5
3. My thoughts become confused and jumbled when I am giving a speech.	1	2	3	4	5
4. Right after delivering a speech, I feel that I have had a pleasant experience.	1	2	3	4	5
5. I get anxious when I think about a speech coming up.	1	2	3	4	5
6. I have no fear (= afraid) of giving a speech.	1	2	3	4	5
7. Although I am nervous just before starting to deliver a speech, I soon settle down after starting and feel calm and comfortable.	1	2	3	4	5
8. I look forward to delivering a speech. (= I expect to give a speech.)	1	2	3	4	5
9. When the instructor announces a speaking	1	2	3	4	5

assignment in class, I can feel myself getting tense.					
10. My hands tremble (= are shaking) when I am giving a speech.	1	2	3	4	5
11. I feel relaxed while delivering a speech.	1	2	3	4	5
12. I enjoy preparing for a speech.	1	2	3	4	5
13. I am in constant fear of forgetting what I prepared to say. (=I keep worrying about forgetting what I prepared to say.)	1	2	3	4	5
14. I get anxious (= worried) if someone asks me something about my topic.	1	2	3	4	5
15. I face the prospect (=likelihood, possibility) of delivering a speech with confidence.	1	2	3	4	5
16. I feel that I am in complete possession of myself (= I can control my emotion/reaction) while delivering a speech.	1	2	3	4	5
17. My mind is clear when delivering a speech.	1	2	3	4	5
18. I do **NOT** dread (=feel very unhappy/worried) delivering a speech.	1	2	3	4	5
19. I perspire (= sweat) just before starting to deliver a speech.	1	2	3	4	5
20. My heart beats very fast just as I start delivering a speech.	1	2	3	4	5
21. I experience considerable anxiety (= feel very worried) while sitting in the room just before my speech starts.	1	2	3	4	5
22. Certain parts of my body feel very tense and rigid (=tight, not flexible) while delivering a speech.	1	2	3	4	5

23. Realizing that only a little time remains in a speech makes me very tense (=nervous) and anxious (=worried).	1	2	3	4	5
24. While **delivering** a speech, I know I can control my feelings of tension and stress.	1	2	3	4	5
25. I breathe faster just before starting to deliver a speech.	1	2	3	4	5
26. I feel comfortable and relaxed in the hour or so just before delivering a speech.	1	2	3	4	5
27. I do poorer (=not very good) on speeches because I am anxious.	1	2	3	4	5
28. I feel anxious when the teacher announces the date of a speaking assignment.	1	2	3	4	5
29. When I make a mistake while delivering a speech, I find it hard to concentrate on (=focus on) the parts that follow (next part).	1	2	3	4	5
30. During an important speech, I experience a feeling of helplessness building up inside me. (=I feel that I do not have power to do anything useful by myself.)	1	2	3	4	5
31. I have trouble falling asleep the night before a speech. (=I can't sleep well the night before a speech.)	1	2	3	4	5
32. My heart beats very fast while I deliver (=speak) a speech.	1	2	3	4	5
33. I feel anxious (=worried) while waiting to deliver my speech.	1	2	3	4	5
34. While delivering a speech, I get so nervous and I forget the facts I really know.	1	2	3	4	5

Appendix A-2: Foreign Language Classroom Anxiety Scales (FLCAS)

ID: ______________________

Thank you very much for participating in this study. The following questionnaire relates to levels of anxiety experienced in foreign language classes. It is adapted from work by Horwitz et al. (1986). Please read each statement carefully and indicate the extent to which you agree or disagree by circling your choice on the five-point scale. The results of this survey will be used only for research purposes. Therefore, please be honest with your responses.

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
1. I never feel pretty sure of myself when I am speaking in my English class.	1	2	3	4	5
2. I don't worry about making mistakes in English class.	1	2	3	4	5
3. I am trembling (=am shaking) when I know that I'm going to be called on in English class.	1	2	3	4	5
4. It frightens (=scares) me when I don't understand what the instructor is saying in English.	1	2	3	4	5
5. It wouldn't bother me at all to take more English classes	1	2	3	4	5
6. During English class, I find myself thinking about things that have nothing to do with the course.	1	2	3	4	5
7. I keep thinking that the other students are better at English than I am.	1	2	3	4	5
8. I am usually at ease (=comfortable) during tests in my English class.	1	2	3	4	5
9. I start to panic when I have to speak without preparation in English class.	1	2	3	4	5

10. I worry about the consequences (=results) of failing my English class.	1	2	3	4	5
11. My stomach becomes upset before important tests.	1	2	3	4	5
12. I don't understand why some people get so upset over English classes.	1	2	3	4	5
13. In English class, I can get so nervous that I forget things I know.	1	2	3	4	5
14. It embarrasses me to volunteer answers in my English class.	1	2	3	4	5
15. I would not be nervous speaking English with native speakers.	1	2	3	4	5
16. I get upset (= unhappy) when I don't understand what the teacher is correcting.	1	2	3	4	5
17. Even if I am well prepared for English class, I feel anxious (= nervous, worried) about it.	1	2	3	4	5
18. I often feel like **not** going to my English class.	1	2	3	4	5
19. I feel confident when I speak English in English class.	1	2	3	4	5
20. During tests, I sometimes get so nervous that I forget facts I really know.	1	2	3	4	5
21. I am afraid that my English instructor is ready to correct every mistake I make.	1	2	3	4	5
22. I can feel my heart pounding when I'm going to be called on in an English class.	1	2	3	4	5
23. The more I study for an English test, the more confused (= don't understand) I get.	1	2	3	4	5
24. I don't feel pressure to prepare very well for English class.	1	2	3	4	5
25. I always feel that the other students speak English better than I do.	1	2	3	4	5

26. I feel very self-conscious about speaking English in front of the other students.	1	2	3	4	5
27. English class moves so quickly, so I worry about getting left behind.	1	2	3	4	5
28. I feel more tense and nervous in my English class than in my other classes.	1	2	3	4	5
29. I get nervous and confused (=don't understand) when I am speaking English in my English class.	1	2	3	4	5
30. When I'm on my way to English class, I feel very sure and relaxed.	1	2	3	4	5
31. I get nervous when I don't understand every word the English teacher.	1	2	3	4	5
32. I feel overwhelmed (= I feel too much of things) by the number of rules you have to learn in order to speak English. (e.g. He go**es**, five flower**s)**	1	2	3	4	5
33. I am afraid that the other students will laugh at me when I speak English.	1	2	3	4	5
34. I would probably feel comfortable around native speakers of the foreign language	1	2	3	4	5
35. I get nervous when the English instructor asks questions which I haven't prepared in advance (=before).	1	2	3	4	5

Appendix B. Interview Questions for English Language Learners

ID: ___________________________

1. Why are you taking this speaking and listening class?
   What is your goal of taking this class? What do you expect to have achieved by the end of the course?

2. Have you lived in English-speaking country (countries) such as the United States, the United Kingdom, Australia, New Zealand, etc.? (If yes) How long have you lived in _______?

3. Do you speak English every day? / How often do you speak English? (If yes)

   a. Who do you usually speak English with?

   b. Have you tried to practice speaking English?

   c. While you were in your home country, have you had many chances to practice speaking English in front of a class?

4. Can you describe your experience of delivering a presentation in class using English? How was the experience of delivering a presentation in this speaking and listening class?

5. When you deliver a presentation in class in English, what is the biggest thing you think about?

(If a response includes experiences relating to anxiety, worries, negative emotion, complete the sub-parts below)

   a. What caused you to feel (*previous response*) and what happened?

   b. Can you describe what you feel (*previous response*) when speaking in English in front of class?

   c. Which one do you feel more anxious to talk about? Your opinions? Or facts/truths?

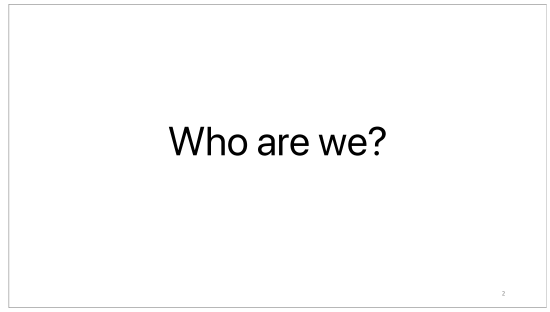
   d. Before delivering a presentation in class, is there anything you do to reduce your anxiety? When do you use this technique? How often?

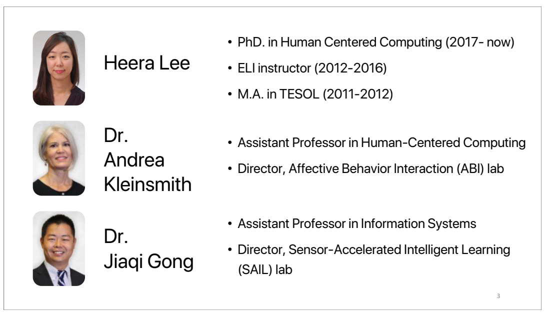
   e. How do you know you are feeling anxious? When you are delivering a presentation in class, are there any specific/ particular gestures or mannerisms you exhibit?

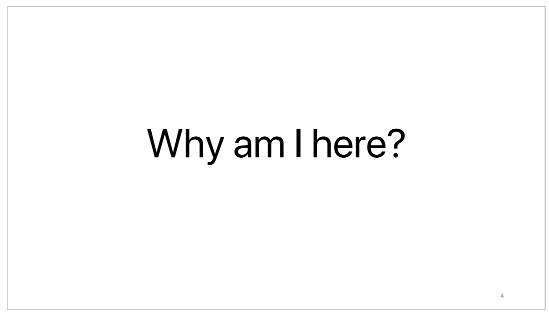
6. In general, how do estimate your performance when you speak in English in front of class? How effective is your delivery of English? (e.g., scale out of 10) Why do you think this?

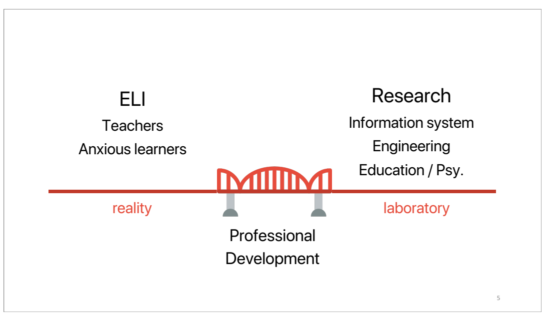
7. Do you have any additional comments or questions?

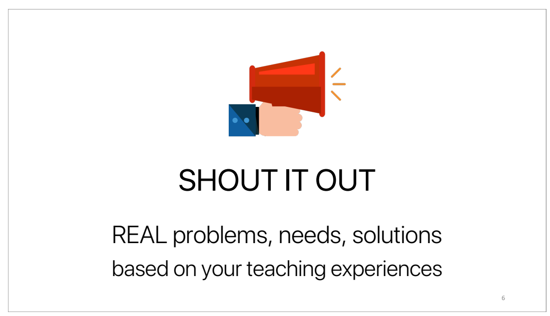
Appendix C. Professional Development Workshop PowerPoint Slides

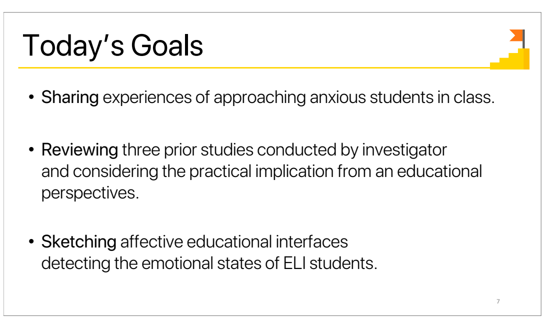


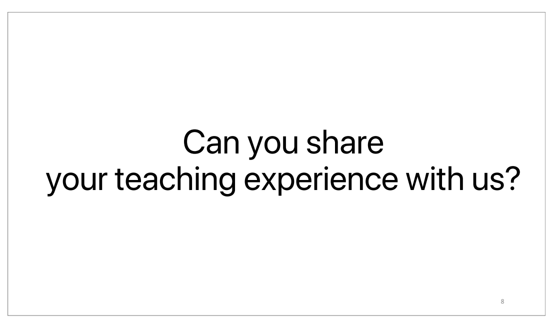


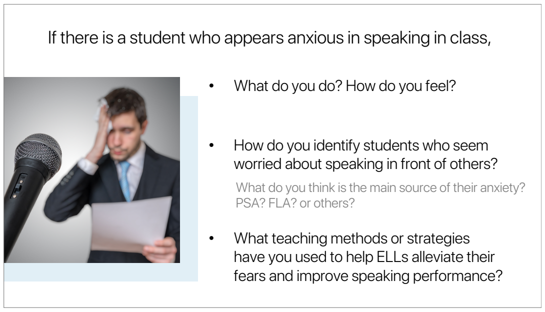


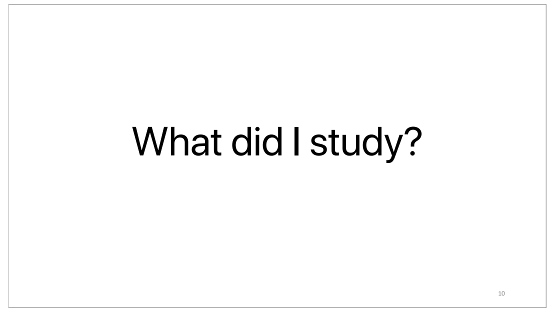


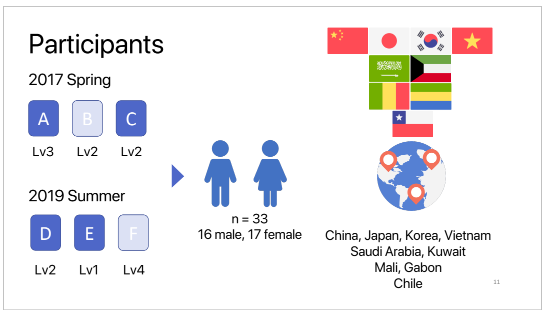


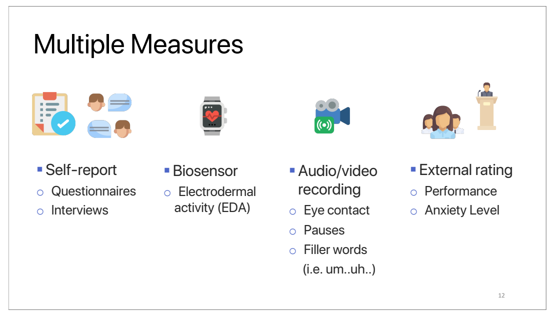




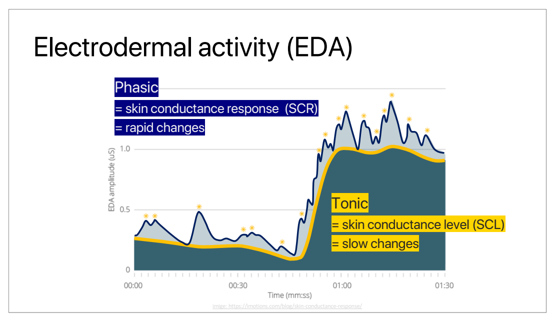

If there is a student who appears anxious in speaking in class,
What do you do? How do you feel?
How do you identify students who seem worried about speaking in front of others?
What do you think is the main source of their anxiety? PSA? FLA? or others?
What teaching methods or strategies have you used to help ELLs alleviate their fears and improve speaking performance?

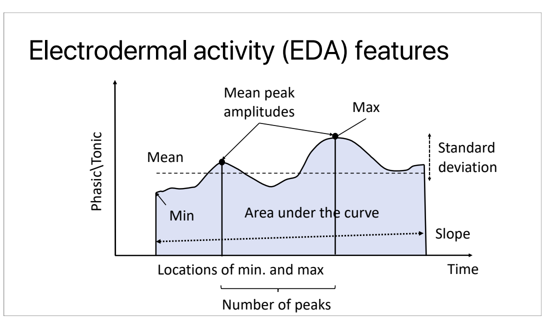

What did I study?
10

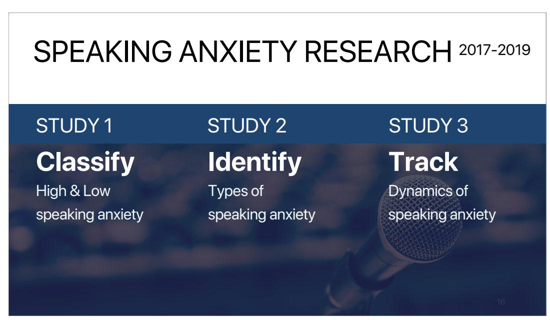

Participants
2017 Spring
A
B
C
Lv3
Lv2
Lv2
2019 Summer
D
E
F
Lv2
Lv1
Lv4
n = 33
16 male, 17 female
China, Japan, Korea, Vietnam
Saudi Arabia, Kuwait
Mali, Gabon
Chile

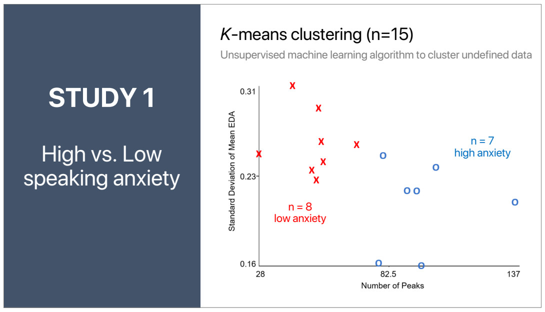

Multiple Measures
Self-report
Questionnaires
Interviews
Biosensor
Electrodermal activity (EDA)
Audio/video recording
Eye contact
Pauses
Filler words
(i.e. um..uh..)
External rating
Performance
Anxiety Level
12


E4 wristband
Heart rate
Skin temperature
3-axis accelerometer
Amount of sweat on skin
(electrodermal activity = EDA)
https://www.empatica.com/research/e4/

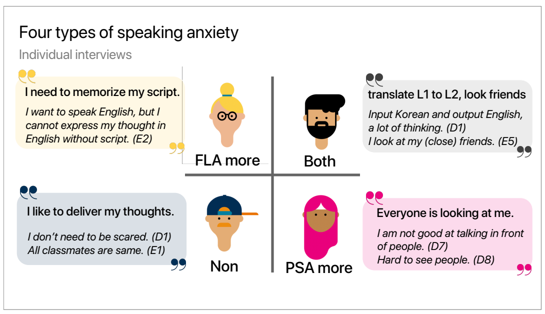

Electrodermal activity (EDA)
Phasic
= skin conductance response (SCR)
= rapid changes
Tonic
= skin conductance level (SCL)
= slow changes
EDA amplitude (uS)
1.0
0.5
0
00:00
00:30
01:00
01:30
Time (mm:ss)

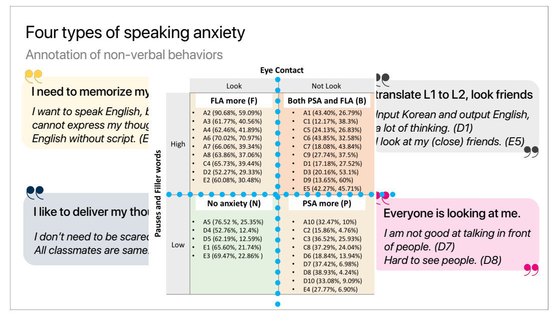

Electrodermal activity (EDA) features
Phasic\Tonic
Mean peak amplitudes
Max
Mean
Standard deviation
Min
Area under the curve
Slope
Locations of min. and max
Time
Number of peaks

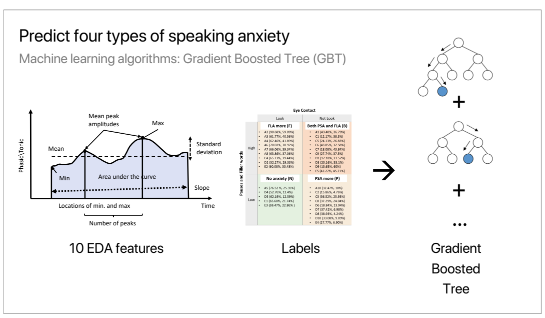

SPEAKING ANXIETY RESEARCH 2017-2019
STUDY 1
Classify
High & Low speaking anxiety
STUDY 2
Identify
Types of speaking anxiety
STUDY 3
Track
Dynamics of speaking anxiety

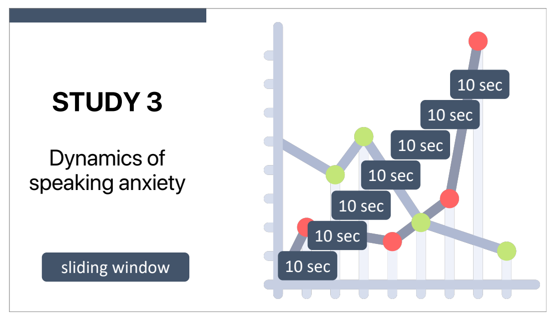

STUDY 1
High vs. Low speaking anxiety
K-means clustering (n=15)
Unsupervised machine learning algorithm to cluster undefined data
Standard Deviation of Mean EDA
0.31
0.23
0.16
28
82.5
137
Number of Peaks
n = 7
high anxiety
n = 8
low anxiety

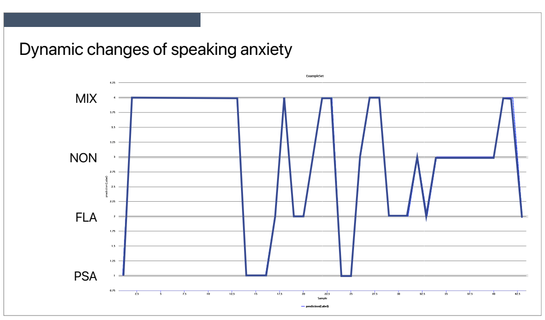

STUDY 2
Types of speaking anxiety

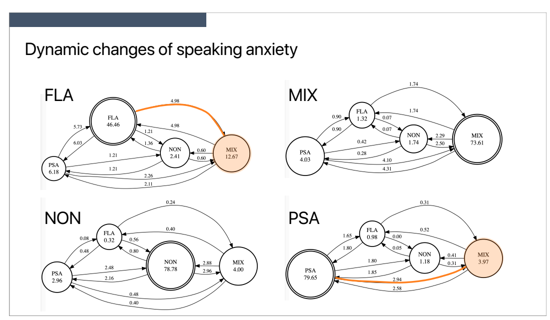

Four types of speaking anxiety
Individual interviews
I need to memorize my script.
I want to speak English, but I cannot express my thought in English without script. (E2)
translate L1 to L2, look friends
Input Korean and output English, a lot of thinking. (D1)
I look at my (close) friends. (E5)
FLA more
Both
I like to deliver my thoughts.
I don't need to be scared. (D1)
All classmates are same. (E1)
Non
PSA more
Everyone is looking at me.
I am not good at talking in front of people. (D7)
Hard to see people. (D8)


Four types of speaking anxiety
Annotation of non-verbal behaviors
Eye Contact
Look
Not Look
FLA more (F)
Both PSA and FLA (B)
No anxiety (N)
PSA more (P)
Pauses and Filler words
High
Low


Predict four types of speaking anxiety
Machine learning algorithms: Gradient Boosted Tree (GBT)
10 EDA features
Labels
Gradient Boosted Tree


STUDY 3
Dynamics of speaking anxiety
sliding window
10 sec


Dynamic changes of speaking anxiety
MIX
NON
FLA
PSA


Dynamic changes of speaking anxiety
FLA
MIX
NON
PSA

Educational implications

STUDY 1
Classify
High & Low speaking anxiety

STUDY 2
Identify
Types of speaking anxiety

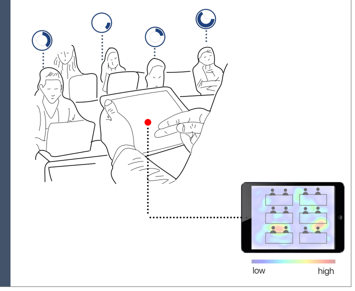
STUDY 3
Track
Dynamics of speaking anxiety

25

Educational implication

STUDY 1
High vs. Low speaking anxiety

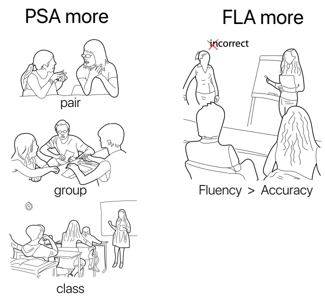
- Awareness frustrated students
- Pairing / Grouping



Educational implication

STUDY 2 - 3
Types and dynamics of speaking anxiety

- Customized drills
  - PSA
  - FLA



Let's take a break (10 min)

29

Today's Goals

- Sharing experiences of approaching anxious students in class.
- Reviewing three prior studies conducted by investigator and considering the practical implication from an educational perspectives.
- Sketching affective educational interfaces detecting the emotional states of ELI students.

28

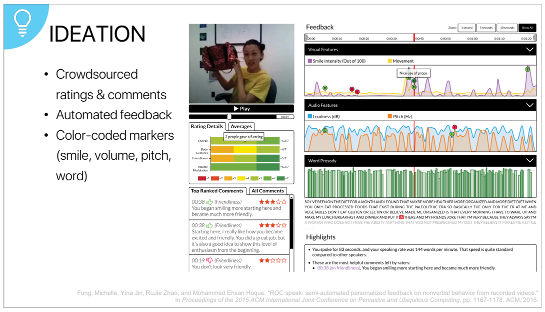
Let's make your idea real

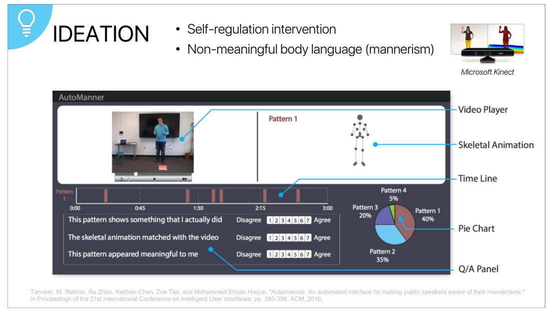
31

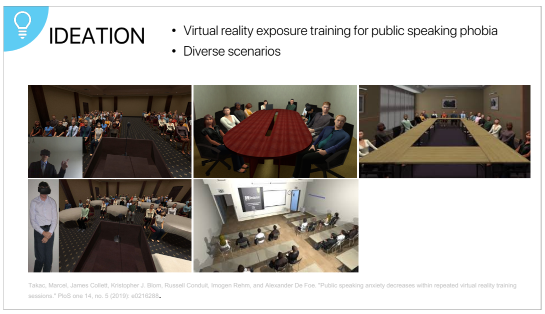
IDEATION → SKETCH PROTOTYPE → PRESENT (IMPLEMENT) → FEEDBACK (TEST)

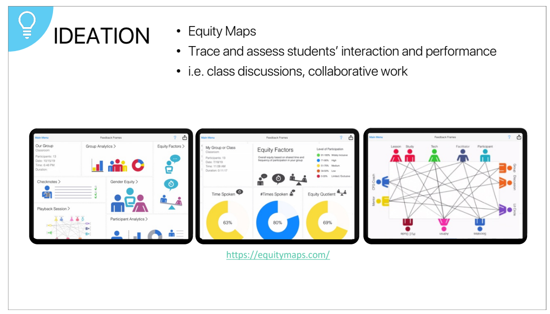
IDEATION

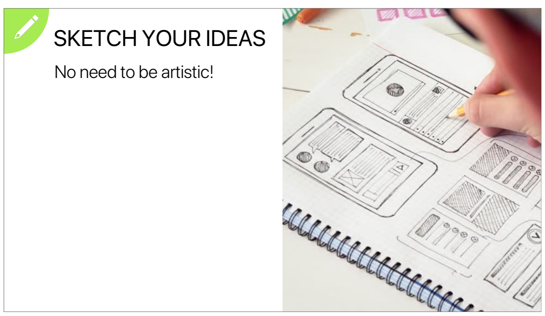
- What technology can be useful
- to understand learners' emotional states and
- eventually help their learning process?
- Write 1-2 ideas on post-it.

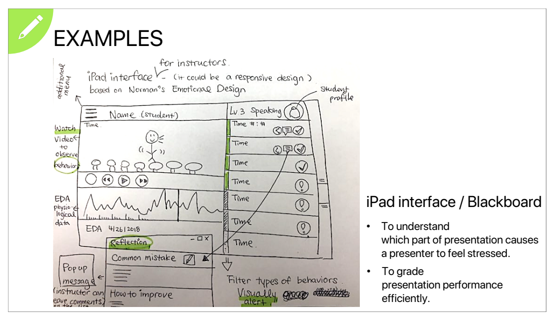

IDEATION
Crowdsourced ratings & comments
Automated feedback
Color-coded markers (smile, volume, pitch, word)
Feedback
Highlights
Fung, Michelle, Yina Jin, RuJie Zhao, and Mohammed Ehsan Hoque. "ROC speak: semi-automated personalized feedback on nonverbal behavior from recorded videos." In Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 1167-1178. ACM, 2015.

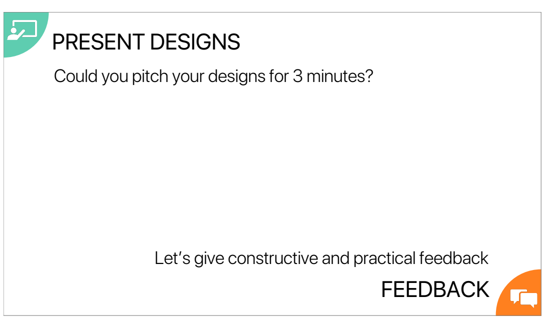

IDEATION
Self-regulation intervention
Non-meaningful body language (mannerism)
Microsoft Kinect
AutoManner
Video Player
Skeletal Animation
Time Line
Pie Chart
Q/A Panel
This pattern shows something that I actually did
The skeletal animation matched with the video
This pattern appeared meaningful to me
Pattern 1 40%
Pattern 2 35%
Pattern 3 20%
Pattern 4 5%
Tanveer, M. Iftekhar, Ru Zhao, Kezhen Chen, Zoe Tiet, and Mohammed Ehsan Hoque. "Automanner: An automated interface for making public speakers aware of their mannerisms." In Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 385-396. ACM, 2016.

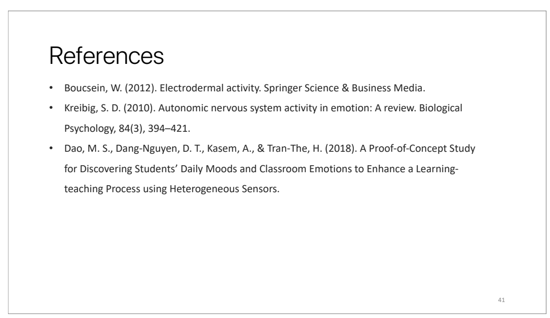

IDEATION
Virtual reality exposure training for public speaking phobia
Diverse scenarios
Takac, Marcel, James Collett, Kristopher J. Blom, Russell Conduit, Imogen Rehm, and Alexander De Foe. "Public speaking anxiety decreases within repeated virtual reality training sessions." PloS one 14, no. 5 (2019): e0216288.


IDEATION
Equity Maps
Trace and assess students' interaction and performance
i.e. class discussions, collaborative work
https://equitymaps.com/


SKETCH YOUR IDEAS
No need to be artistic!


EXAMPLES
for instructors
iPad interface (it could be a responsive design)
based on Norman's Emotional Design
Student profile
Name (student)
Speaking
EDA
Common mistake
How to improve
Filter types of behaviors
iPad interface / Blackboard
To understand which part of presentation causes a presenter to feel stressed.
To grade presentation performance efficiently.


PRESENT DESIGNS
Could you pitch your designs for 3 minutes?
Let's give constructive and practical feedback
FEEDBACK


References
Boucsein, W. (2012). Electrodermal activity. Springer Science & Business Media.
Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. Biological Psychology, 84(3), 394–421.
Dao, M. S., Dang-Nguyen, D. T., Kasem, A., & Tran-The, H. (2018). A Proof-of-Concept Study for Discovering Students' Daily Moods and Classroom Emotions to Enhance a Learning-teaching Process using Heterogeneous Sensors.
41

Appendix D: Evaluation Form for the Professional Development Workshop

PROFESSIONAL DEVELOPMENT EVALUATION CHECKLIST

Please rate how strongly you agree or disagree with each of the following statement by circling the appropriate number.	Strongly Agree	Agree	Neither Agree Nor Disagree	Disagree	Strongly Disagree
Content					
1. The objectives for today's session were clearly stated.	☐	☐	☐	☐	☐
2. Today's session was aligned to its stated objectives.	☐	☐	☐	☐	☐
3. Today's session was useful and practical.	☐	☐	☐	☐	☐
4. Today's session advanced the development of my teaching methods or class activities.	☐	☐	☐	☐	☐
Process					
5. Today's activities (presentations, focus group interview, group activity, etc.) increased my capacity to understand my students.	☐	☐	☐	☐	☐
6. The facilitators of today's session effectively demonstrated how to understand.	☐	☐	☐	☐	☐
7. The facilitators of today's session incorporated our experiences into today's activities.	☐	☐	☐	☐	☐
8. Time was allocated effectively today to deepen my understanding of the presented material.	☐	☐	☐	☐	☐
Context					
9. There were opportunities during today's session to collaborate on shared activities.	☐	☐	☐	☐	☐

10. Today's activities (presentations, scenarios, group exercises, etc.) were relevant for my job-related needs.	☐	☐	☐	☐	☐
11. Today's sessions advanced my understanding of how to engage in a continuous improvement cycle.	☐	☐	☐	☐	☐
12. The organization of the learning environment (tools, materials, participant groupings, etc.) met my learning needs.	☐	☐	☐	☐	☐

Comments

PROFESSIONAL DEVELOPMENT – EVALUATION QUESTIONS

Teaching subject and level: ____

Please respond to the following questions.

1. Of all the things learned in today's session, what was the most valuable learning experience and least helpful experience?

2. What things did you learn through today's professional development that were unexpected? Why?

3. As a result of today's session, what will you do differently in the future? Why?

4. What could be done to improve this seminar?

5. For future studies, what topics would be most helpful in teaching and learning process in the ELI?

6. Additional comments?

Appendix E. Coding form

Coder: _______________

Example variables of each filtering dimension adapted by (Lim et al., 2008)

Filtering	Example variables
Appearance	Platform/device, size, color, shape, margin, etc.
Data	Data type (number, string, media), data use, organization, hierarchy
Functionality	Functional components, system function, users' functionality need (buttons, menu bar, dropdown boxes, text-field)
Interactivity	Input behavior (click, scroll), output behavior, feedback behavior
Spatial structure	Arrangement of interface or information elements (frame division)

Coding Participant: __________

- Appearance (e.g., Desktop, Tablet, Mobile phone, Smart watch, Others)

- Data: (Find necessary data type)

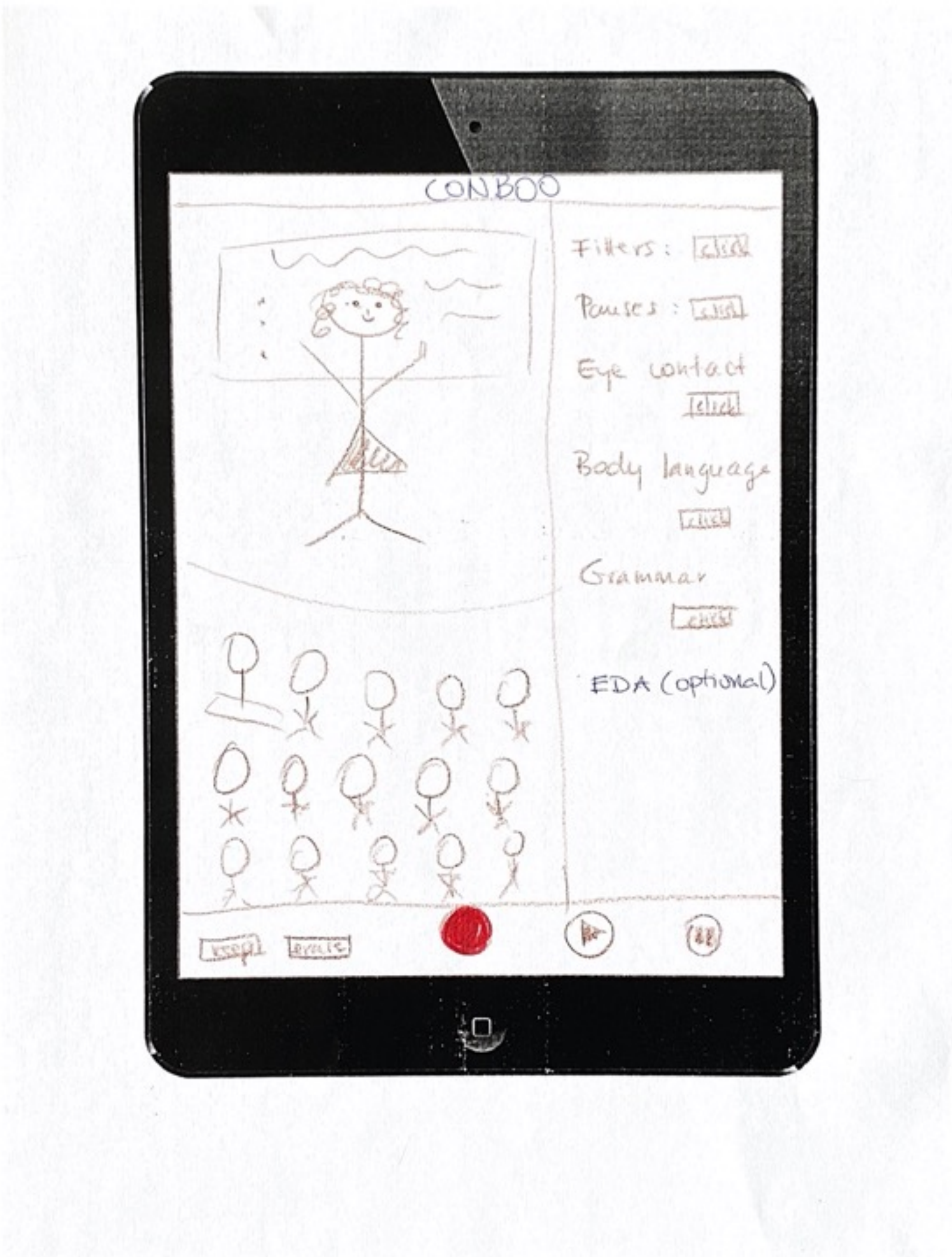
- Functionality: (Find UI components and attach image capture)

- Interactivity: (Find action verbs for input/output/processing behaviors (e.g., click, see, search, write etc. and add image capture)

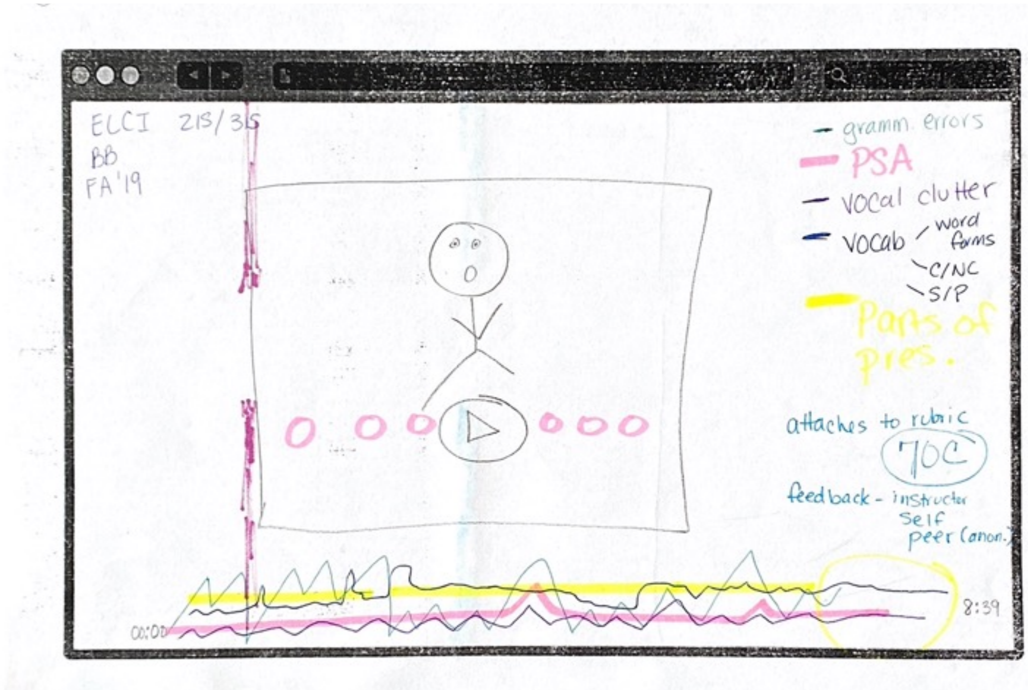
- Spatial structure: (Draw the frame of the design)

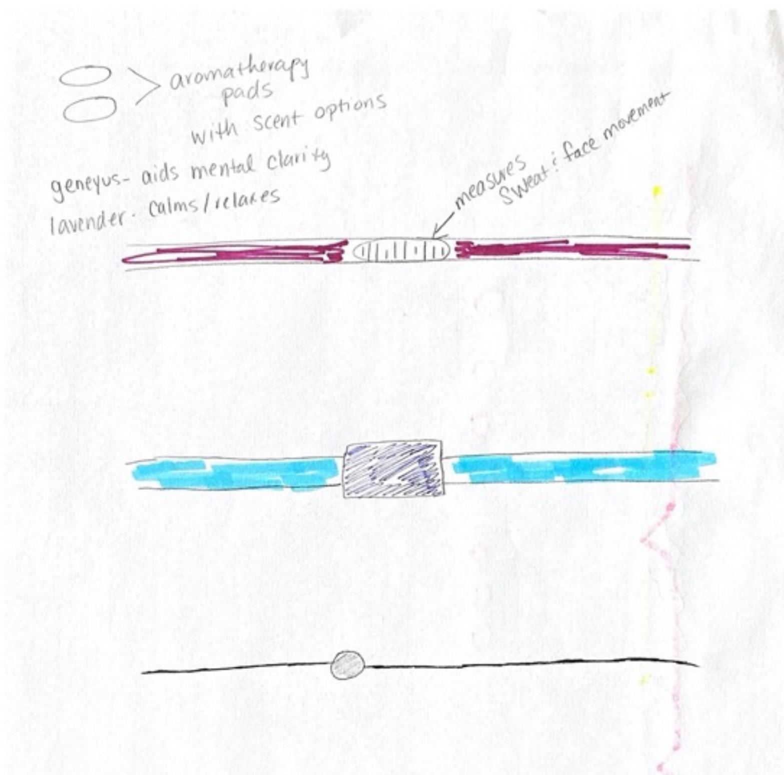
Appendix F. Prototype Hand-drawn Sketches from Participants in the Professional Development Workshop

- **P1**

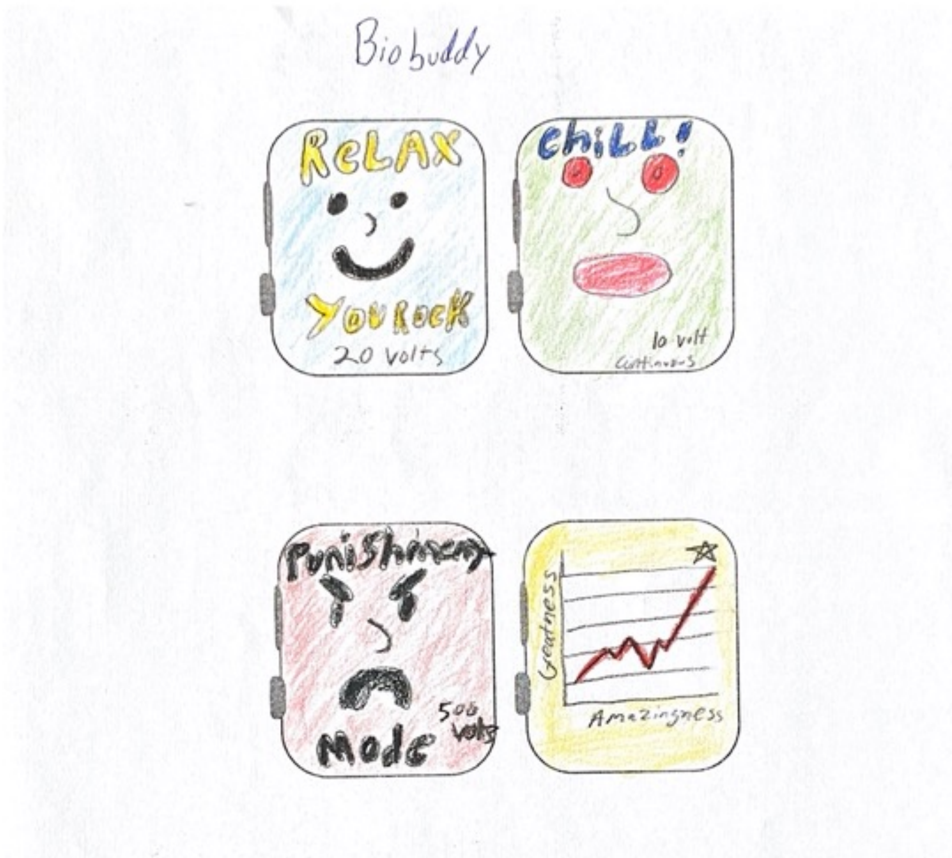


- **P2**

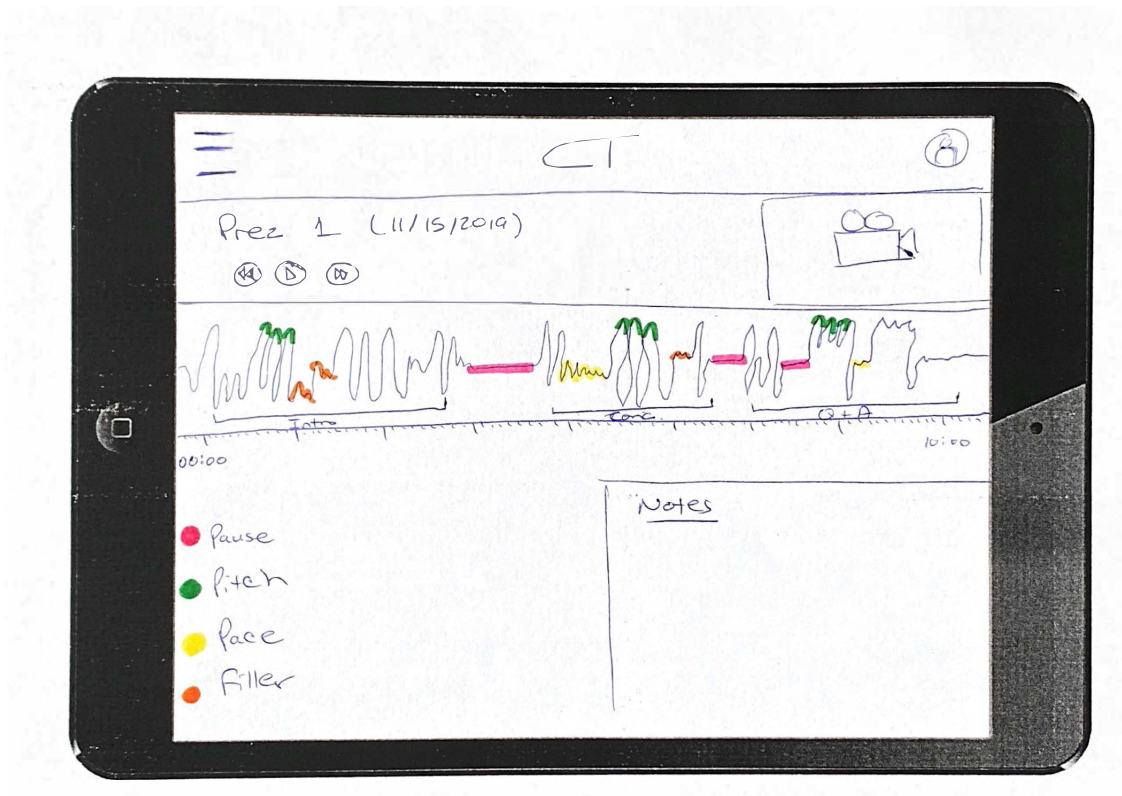

ELCI 215/35
BB
FA'19
gramm. errors
PSA
vocal clutter
vocab
word forms
C/NC
S/P
Parts of pres.
attaches to rubric
70C
feedback - instructor
self
peer (anon.)
00:00
8:39


aromatherapy pads
with scent options
geneyus - aids mental clarity
lavender - calms/relaxes
measures sweat & face movement

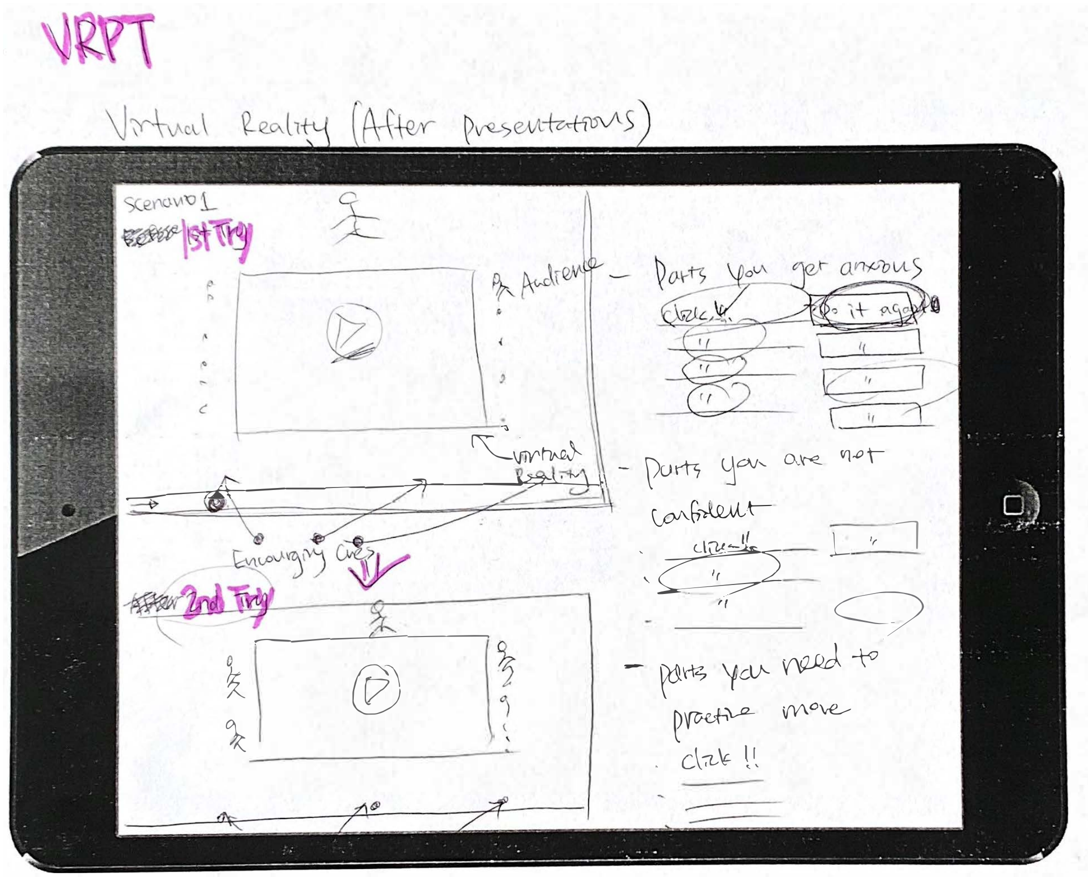
- **P3**

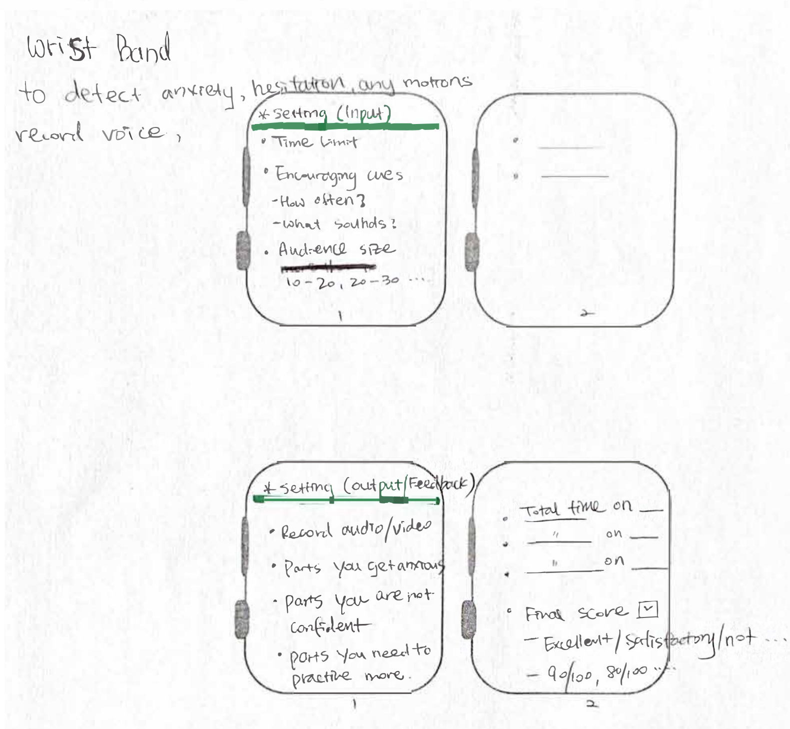

Biobuddy
RELAX
YOU ROCK
20 volts
CHILL!
10 volt
Continuous
Punishment
Mode
500 volts
Greatness
Amazingness

- **P4**

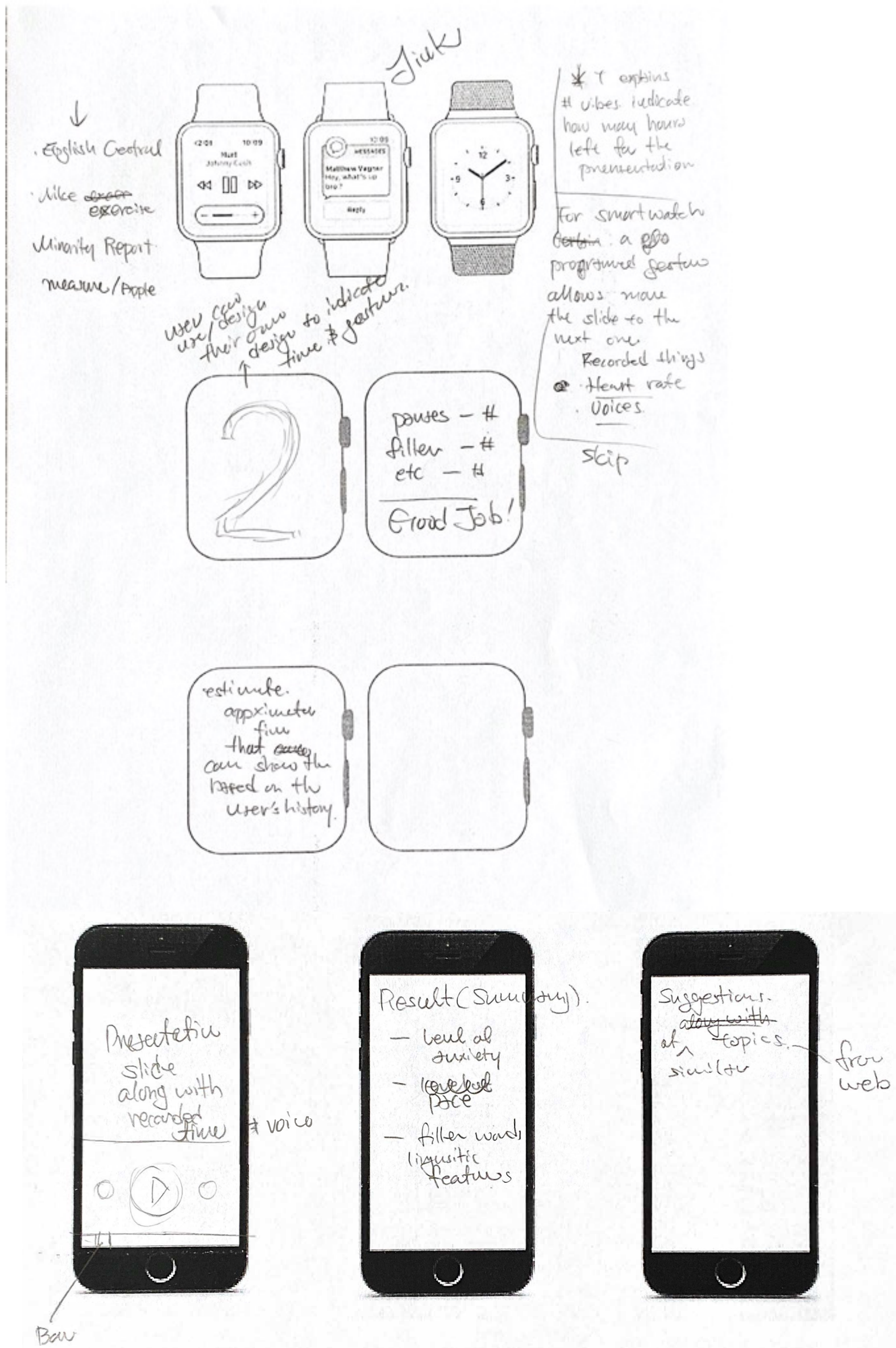

CT
Prez 1 (11/15/2019)
Intro
Conc.
Q+A
00:00
10:00
Notes
Pause
Pitch
Pace
Filler

- **P5**


VRPT
Virtual Reality (After presentations)
Scenario 1
1st Try
Audience
Parts you get anxious
click
Do it again
virtual Reality
Parts you are not confident
click!!
Encouraging Cues
2nd Try
parts you need to practice more
click !!

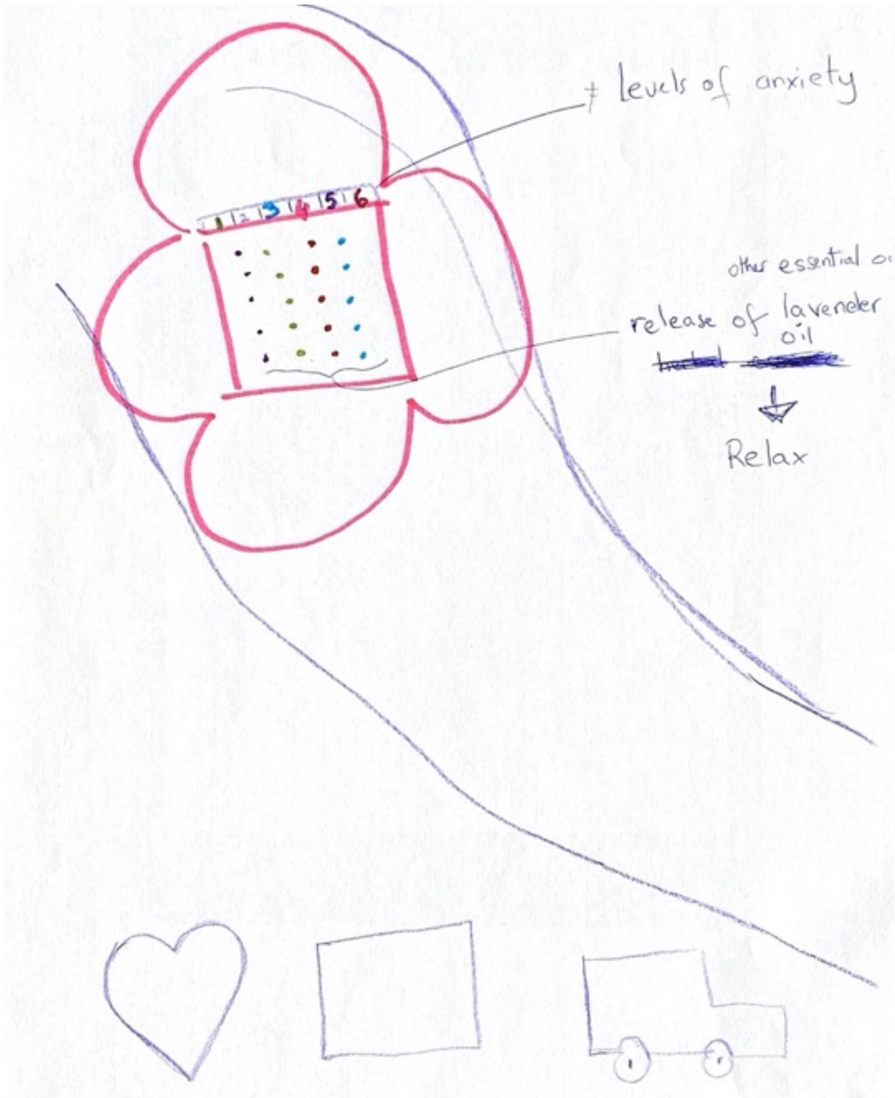

Wrist Band
to detect anxiety, hesitation, any motions
record voice,
* setting (Input)
Time Limit
Encouraging cues
-How often?
-What sounds?
Audience size
10-20, 20-30 ...
* setting (output/Feedback)
Record audio/video
Parts you get anxious
parts you are not confident
parts you need to practice more.
Total time on
on
on
Final score
- Excellent/satisfactory/not ...
- 90/100, 80/100 ...

- **P6**



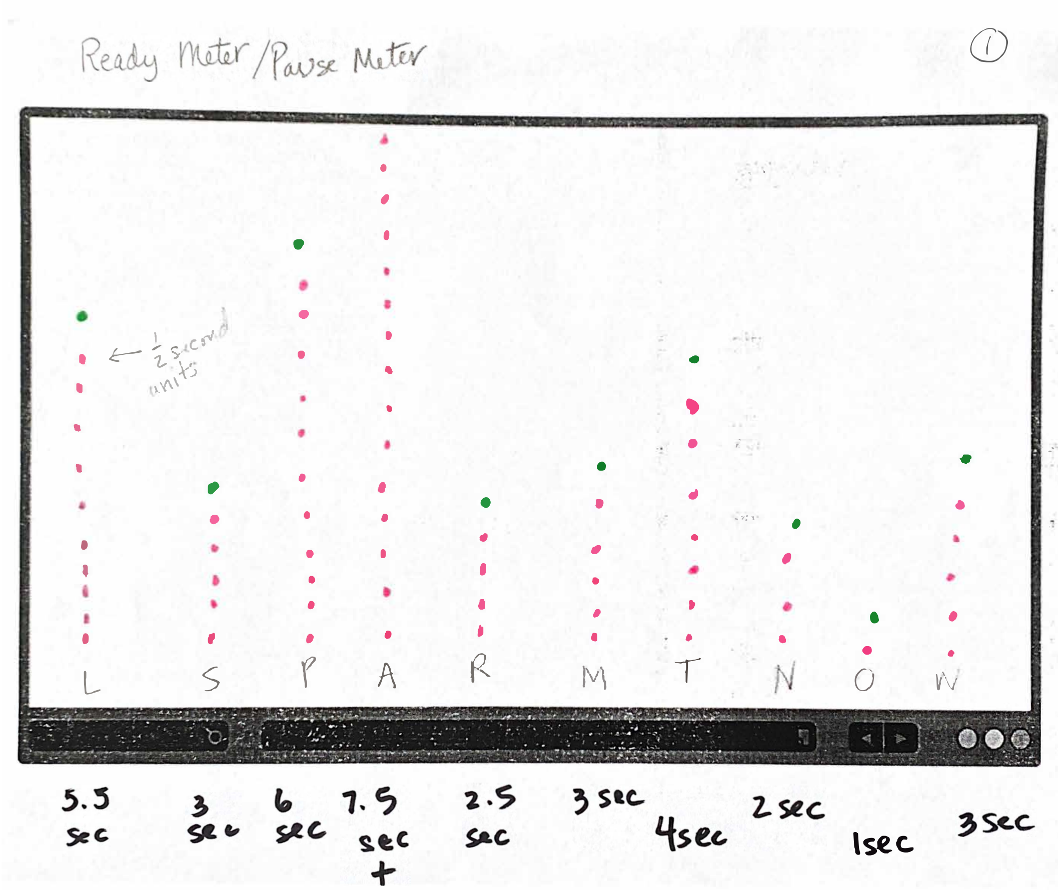
English Central
Nike exercise
Minority Report
measure/Apple
user can use/design their own design to indicate time & gesture.
* explains # vibes indicate how many hours left for the presentation
For smartwatch : a programmed gesture allows move the slide to the next one.
Recorded things
Heart rate
Voices.
skip
pauses — #
filler — #
etc — #
Good Job!
estimate. approximate time that can show the based on the user's history
Presentation slide along with recorded time
& voice
Bar
Result (Summary).
— level of anxiety
— pace
— filler words linguistic features
Suggestions. of similar topics.
from web

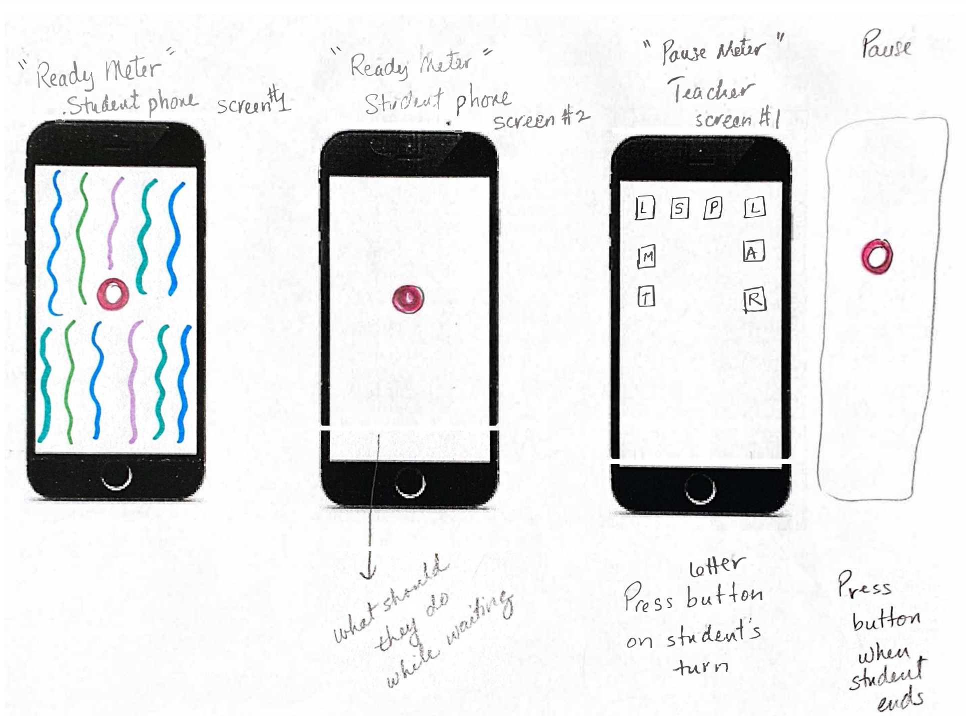
- **P7**



1 2 3 4 5 6
↑ levels of anxiety
other essential oils
release of lavender oil
Relax

- **P8**


Ready Meter/Pause Meter
1
← 1/2 second units
L
S
P
A
R
M
T
N
O
W
5.5 sec
3 sec
6 sec
7.5 sec +
2.5 sec
3 sec
4sec
2 sec
1sec
3sec


"Ready Meter" Student phone screen#1
"Ready Meter" Student phone screen #2
"Pause Meter" Teacher screen #1
Pause
L S P L
M A
T R
what should they do while waiting
letter Press button on student's turn
Press button when student ends

- **P9**

Appendix G. Examples of the Mayer Salovey-Caruso Emotional Intelligence Test

(MSCEIT V2.0)

(As per Terms & Conditions for use of MSCEIT for research purposes, MHS Inc. permitted researchers to cite no more than six items)

The MSCEIT test has a wide range of test question types or styles. In some you will be presented with faces, and you will need to select the emotion it corresponds to, and level of intensity of this emotion. In other questions, you will be presented with scenarios, and be asked to select the most appropriate response. You will also be shown photos of objects and scenery and be asked to identify the emotions that these photos are likely to elicit in people.

Example Item 1.

Joan felt stressed and became a bit anxious when she thought about all the work she needed to do. When her manager brought her an additional project, she felt ____. (Select the best choice.)

a) Overwhelmed
b) Depressed
c) Ashamed
d) Self-conscious
e) Jittery

Example Item 2.

How much is each feeling below expressed in this picture?

- Happy 1-5
- Sad 1-5
- Fear 1-5
- Anger 1-5
- Disgust 1-5

Example Item 3.

How much is each feeling in the list below expressed by this face?

- Happiness (1-5)
- Sadness (1-5)
- Fear (1-5)
- Anger (1-5)
- Disgust (1-5)

Appendix H. Participant Recruitment Flyer

Heera Lee (Alumni TESOL 12')
Human-Centered Computing | Information Systems

Are you interested in evaluating speaking skills
among English language learners?

We are conducting a study designed to develop an approach to support instructors to provide emotional and instructional scaffolding for English language learners (ELLs) through biosensor-based feedback. This study will take approximately 8-12 hours in total, over two months (eight weeks) to complete (1-1.5 hr/week) online.

You may qualify if you

- are taking TESOL course
- have less than a year of teaching experience
- have a desktop/laptop and Internet access at home

You will

- fill out a questionnaire
- watch recorded presentation videos of a class with 12 ELLs
- rank students from the most anxious to the less anxious in order
- have an individual interview with the investigator

Benefit

- compensation with **$180** amount of e-gift card.
- opportunity to develop own TESOL research studies

For more information, please contact heera1@umbc.edu or visit www.heeralee.com/papers

Study Protocol # Y20RK12246

Appendix I. Ranking Survey Form

Evaluation Form of Ranking Emotional States of English Language Learners

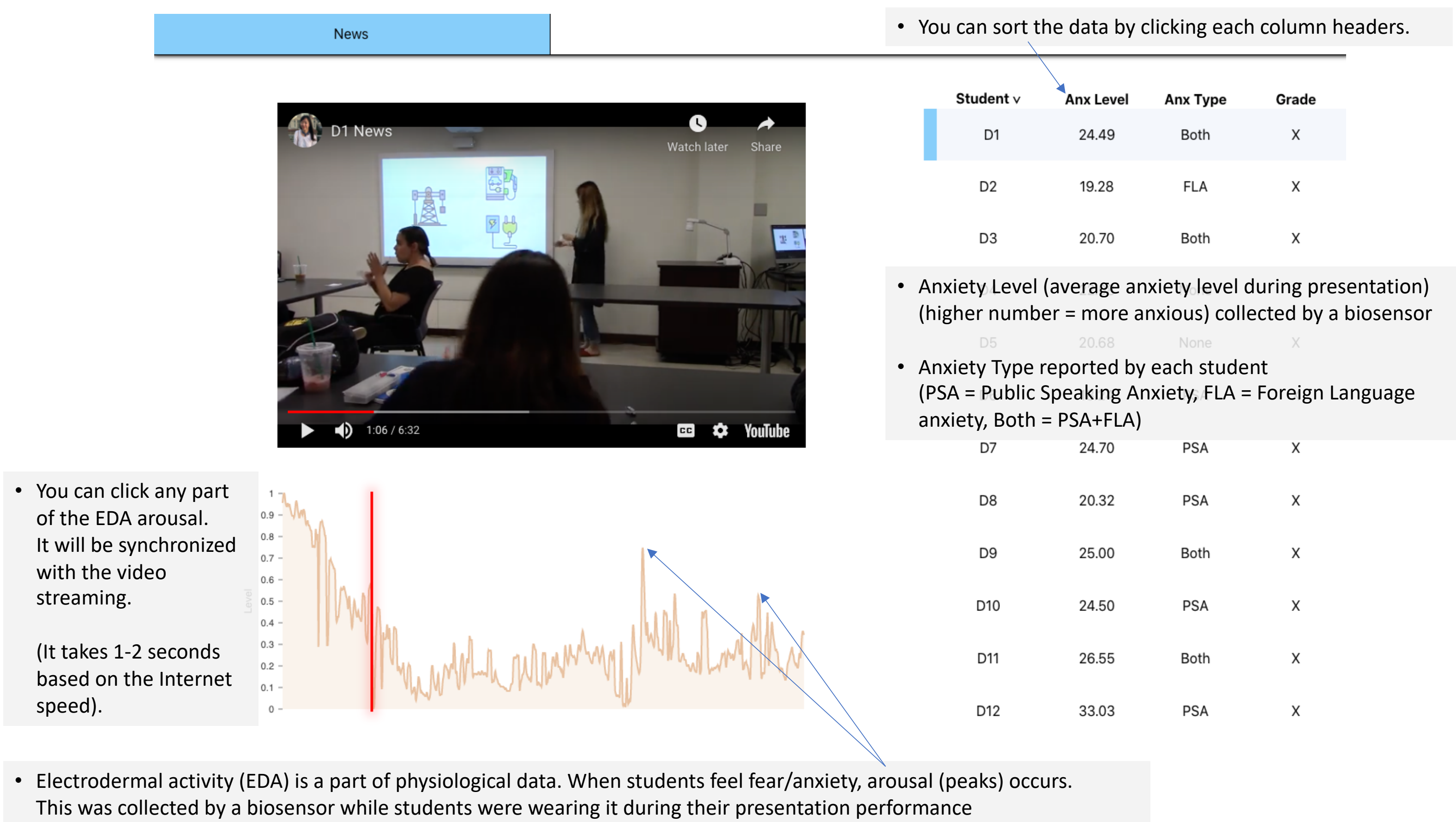
Rank Emotional States of Learners

As part of the task, you will be presented with a range of videos relating to student presentations. After watching each video, please rank each student's emotional state on the continuum by referring to their ID number. All students from D1 to D12 should be marked on the continuum. Add in the reasons for why you ranked the students the way you did. Please do **NOT** rank more than two students in the same position.

Example
Most Anxious
D4
D2
D3
D1
D9
D5
D8
D7
D12
D6
D11
D10
Least anxious

Rank students	Reason	Performance Grade
Most Anxious		(e.g., 75/100)
Least anxious		

Appendix J. Guidance of How to Use a Prototype System.



Appendix K. Interview Questions for English Language Instructors

General Teaching Experience

1. At what stage during a class, or outside of a class, do you provide feedback to students about their presentation performance?
2. How often do you provide feedback to students about their presentation performance?
3. How do you provide feedback to students about presentation performance (e.g. verbal, written, etc.)?
    a. Why do/did you select those types of feedback?
    b. What resources or apps have you used to provide your feedback?
4. How did your students respond to the feedback? (Did their public speaking/emotional state change when speaking?)
5. Among students who appear emotionally anxious, how do you decide which students to help first?

Comparison Between Evaluation 1 (Human rate) and Evaluation 2 (Human rate with system)

1. How similar or different are Evaluation 1 and Evaluation 2 in terms of ranking/grading students?
2. Why do you think Evaluation 1 and Evaluation 2 are similar/different?

Opinions About Using the Prototype System (Widdows, 1991; Mouza, 2015)

https://edu-project-visible-all.netlify.app/

1. What do you like most and least about the prototype system?

   (This open-ended question can be narrowed down with guidance/examples such as: What do you like most about the prototype system in terms of designs and features?)
2. What changes would you make to the prototype system to support instructors when detecting emotional states among students? Are there any functions in the prototype system you would like to add/modify/remove?
3. In what ways, if any, would you use this prototype system to support your teaching?
   a. Can you provide some examples?
4. Are you optimistic or worried about how this prototype system would be used in the future?

Glossary

Definitions of key terms used in the proposal are described below:

- *Instructors*

  The term 'instructors' has been selected for use in this work, as it specifically refers to the term used for English language educators working at a language institute. Both experienced and novice instructors were recruited for studies described in the dissertation.

  The term 'experienced instructors' has been applied to individuals who have been teaching English at a college or university language institute for more than three years. Experienced instructors recruited in this study have experience teaching four specific skills (reading, writing, listening and speaking skills) to international students who study abroad and hail from diverse cultural backgrounds. "Novice instructors", on the other hand, refer to students or pre-service instructors studying in the MA Teaching English to Speakers of Other Languages (TESOL) program at the University of Maryland, Baltimore County (UMBC). They have developed skills teaching English, along with gaining an insight into cross-cultural communication and instructional system design.

- *English language learners (ELLs)*

  In general, ELLs are defined as students with limited levels of English proficiency who speak a language other than English at home (Center, 2015). In the studies described in this dissertation, the term 'ELLs' has been applied to students who enrolled in English language courses in the U.S. who hail from a country where English is considered a foreign language.

- *Emotions*

  The term ‘emotion’ relates to a complex set of interactions, mediated by neural-hormonal systems, which can (a) give rise to affective experiences such as feelings of arousal, pleasure, and displeasure; (b) generate cognitive processes such as emotionally-relevant perceptual effects, appraisals, and labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive (Ekkekakis, 2013). The terms ‘affect’ and ‘emotion’ can be used interchangeably, but ‘affect’ is the most general (Batson, Shaw, & Oleson, 1992) and can be defined as “mental states involving evaluative feelings, psychological conditions when the person feels good or bad, and either likes or dislikes what is happening” (Parkinson, Totterdell, Briner, & Reynolds, 1996). The investigator selected to use the term ‘emotion’ in this work, as the studies described explore learners’ feelings of arousal through data captured using a biosensor (a and c) while delivering a presentation in class (b), and the emotional state was indicated through behaviors (d).

References

ACTFL, J. (2012). ACTFL proficiency guidelines 2012.

Adamson, L. B., Deckner, D. F., & Bakeman, R. (2010). Early interests and joint engagement in typical development, autism, and Down syndrome. *Journal of Autism and Developmental Disorders*, *40*(6), 665–676.

Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *The Modern Language Journal*, *78*(2), 155–168.

Al-Nouh, N. A., Abdul-Kareem, M. M., & Taqi, H. A. (2015). EFL College Students' Perceptions of the Difficulties in Oral Presentation as a Form of Assessment. *International Journal of Higher Education*, *4*(1). https://doi.org/10.5430/ijhe.v4n1p136

Al-Saraj, T. M. (2014). Revisiting the Foreign Language Classroom Anxiety Scale (FLCAS): The Anxiety of Female English Language Learners in Saudi Arabia. *L2 Journal*, *6*(1). https://doi.org/10.5070/l26121650

Albrecht, T. L., & Adelman, M. B. (1987). *Communicating social support.* Sage Publications, Inc.

Alemi, M., Daftarifard, P., & Pashmforoosh, R. (2011). The impact of language anxiety and language proficiency on WTC in EFL context. *Cross-Cultural Communication*, *7*(3), 150–166.

Alfaras, M., Primett, W., Umair, M., Windlin, C., Karpashevich, P., Chalabianloo, N., … Höök, K. (2020). Biosensing and actuation—Platforms coupling body input-output modalities for affective technologies. *Sensors*, *20*(21), 5968.

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review*, *42*(1), 76–98.

Altman, D. G. (1991). Inter-rater agreement. *Practical Statistics for Medical Research*, 403–409.

Amo, D., Fonseca, D., Alier, M., García-Peñalvo, F. J., Casañ, M. J., & Alsina, M. (2019). Personal data broker: A solution to assure data privacy in EdTech. In *International Conference on Human-Computer Interaction* (pp. 3–14). Springer.

Andrei, E. (2017). Technology in teaching English language learners: The case of three middle school teachers. *TESOL Journal*, *8*(2), 409–431.

Anokwa, Y., Smyth, T. N., Ramachandran, D., Sherwani, J., Schwartzman, Y., Luk, R., … DeRenzi, B. (2009). Stories from the field: Reflections on HCI4D experiences. *Information Technologies & International Development*, *5*(4), pp-101.

Atabek, O. (2020). Associations between Emotional States, Self-Efficacy for and Attitude towards Using Educational Technology. *International Journal of Progressive Education*, *16*(2), 175–194.

Awan, R.-U.-N., & Anwar, M. N. (2010). *An Investigation Of Foreign Language Classroom Anxiety And Its Relationship With Students" Achievement*. *Journal of College Teaching & Learning-November* (Vol. 7). Retrieved from https://clutejournals.com/index.php/TLC/article/view/249

Ayres, J., & Heuett, B. L. (1997). The relationship between visual imagery and public speaking apprehension. *Communication Reports*, *10*(1), 87–94.

Babaei, E., & Tag, B. (2021). A critique of electrodermal activity practices at chi. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3411764.3445370

Bagchi, B. K., & Greenwald, D. U. (1937). Electrodermal changes during certain types of attention. *The Journal of General Psychology*, *17*(2), 235–248.

Bahl, S., Bagha, A. K., Rab, S., Javaid, M., Haleem, A., & Singh, R. P. (2021). Advancements in biosensor technologies for medical field and COVID-19 pandemic. *Journal of Industrial Integration and Management*.

Bailey, K. M. (1983). Competitiveness and anxiety in adult second language learning: Looking at and through the diary studies. *Classroom Oriented Research in Second Language Acquisition*, 67–102.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, *20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Bartolomé-Tomás, A., Sánchez-Reolid, R., Fernández-Sotos, A., Latorre, J. M., & Fernández-Caballero, A. (2020). Arousal Detection in Elderly People from Electrodermal Activity Using Musical Stimuli. *Sensors*, *20*(17), 4788.

Baskinger, M. (2008). COVER STORY Pencils before pixels: a primer in hand-generated sketching. *Interactions*, *15*(2), 28–36.

Batson, C. D., Shaw, L. L., & Oleson, K. C. (1992). Differentiating affect, mood, and emotion: toward functionally based conceptual distinctions.

Beamer, L. (1992). Learning intercultural communication competence. *The Journal of*

Business Communication (1973), *29*(3), 285–303.

Beck, J. S., & Nunnaley, D. (2020). A continuum of data literacy for teaching. *Studies in Educational Evaluation*, 100871.

Beck, T. A. (2008). Behind the Mask: Social Studies Concepts and. *Social Education*, *72*(4), 181–184.

Beidel, D. C., Turner, S. M., & Dancu, C. V. (1985). Physiological, cognitive and behavioral aspects of social anxiety. *Behaviour Research and Therapy*, *23*(2), 109–117.

Bell, A. M. (1853). *Observation on Speech; the causes and the cure of stammering, mal-articulations, and defects, etc*. WP Kennedy.

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, *47*(4), 647–658.

Benesch, S. (2013). *Considering emotions in critical English language teaching: Theories and praxis*. Routledge.

Berger, T., Lettner, D., Rubin, J., Grünbacher, P., Silva, A., Becker, M., … Czarnecki, K. (2015). What is a feature? a qualitative study of features in industrial software product lines. In *Proceedings of the 19th International Conference on Software Product Line* (pp. 16–25).

Bickmore, T., Kimani, E., Shamekhi, A., Murali, P., Parmar, D., & Trinh, H. (2020). Virtual agents as supporting media for scientific presentations. *Journal on Multimodal User Interfaces*, 1–16.

Blöte, A. W., Kint, M. J. W., Miers, A. C., & Westenberg, P. M. (2009). The relation between public speaking anxiety and social anxiety: A review. *Journal of Anxiety*

Disorders, *23*(3), 305–313.

Bodie, G. D. (2010). A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety. *Communication Education*, *59*(1), 70–105.

Boeije, H. (2002). A purposeful approach to the constant comparative method in the analysis of qualitative interviews. *Quality and Quantity*, *36*(4), 391–409.

Bosch, N., D'Mello, S., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., … Zhao, W. (2015). Automatic detection of learning-centered affective states in the wild. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (Vol. 2015-January, pp. 379–388). Association for Computing Machinery. https://doi.org/10.1145/2678025.2701397

Boucsein, W. (2012). *Electrodermal activity*. Springer Science & Business Media.

Brackett, M. A., & Katulak, N. A. (2006). Emotional intelligence in the classroom: Skill-based training for teachers and students. *Applying Emotional Intelligence: A Practitioner's Guide*, 1–27.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, *49*(1), 1017–1034.

Bransford, J. D., Brown, A. L., & C. (2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Retrieved from https://books.google.com/books?hl=en&lr=&id=QZb7PnTgSCgC&oi=fnd&pg=PR1&dq=How+People+Learn.+Brain,+Mind,+Experience,+and+School&ots=FtPWnK9z-F&sig=0m5LqTnDMS3voyvqID9uThVKAHo

Brantmeier, C. (2005). *ANXIETY ABOUT L2 READING OR L2 READING TASKS? A STUDY WITH ADVANCED LANGUAGE LEARNERS*. *The Reading Matrix* (Vol. 5).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Brookes, H. F., & Ross, H. (1967). *English as a foreign language for science students*. Heinemann.

Brooks, G., & Wilson, J. (2014). *Using Oral Presentations to Improve Students' English Language Skills*. *Kwansei Gakuin University Humanities Review* (Vol. 19). Retrieved from https://www.researchgate.net/publication/324217578

Bubel, M., Jiang, R., Lee, C. H., Shi, W., & Tse, A. (2016). AwareMe: addressing fear of public speech through awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 68–73).

Butler, R. M., Boden, M. T., Olino, T. M., Morrison, A. S., Goldin, P. R., Gross, J. J., & Heimberg, R. G. (2018). Emotional clarity and attention to emotions in cognitive behavioral group therapy and mindfulness-based stress reduction for social anxiety disorder. *Journal of Anxiety Disorders*, *55*, 31–38.

Buxton, W. (2007). *Sketching user experience : tting the design right and the right design / Bill Buxton*.

Çağatay, S. (2015). Examining EFL Students' Foreign Language Speaking Anxiety: The Case at a Turkish State University. *Procedia - Social and Behavioral Sciences*, *199*, 648–656. https://doi.org/10.1016/j.sbspro.2015.07.594

Campbell, C. M., & Shaw, V. M. (1994). Language anxiety and gender differences in adult second language learners: Exploring the relationship. *Faces in a Crowd: The*

Individual Learner in Multisection Courses, 47–80.

Carroll, J. M., & Rosson, M. B. (2013). Wild at home: The neighborhood as a living laboratory for HCI. *ACM Transactions on Computer-Human Interaction*, *20*(3). https://doi.org/10.1145/2491500.2491504

Center, K. C. D. (2015). Children who speak a language other than English at home. *Retrieved from Annie E. Casey Foundation Website: Http://Datacenter. Kidscount. Org/Data/Tables/81-Children-Who-Spe# Detailed/1/Any/False/36,868,867,133*, *38*.

CHARLES E., B., JON B., M., LYNNA J., A., ROBERT, D., & INA, A. (2017). Electro-Dermal Activity (Eda) and Sensors: an Emerging Technology for Educational Research. *I-Manager's Journal of Educational Technology*, *14*(2), 20. https://doi.org/10.26634/jet.14.2.13716

Chaturvedi, S., & Shweta, R. C. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: an overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, *41*(3), 20–27.

Cheemalapati, S., Gubanov, M., Del Vale, M., & Pyayt, A. (2013). A real-time classification algorithm for emotion detection using portable EEG. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)* (pp. 720–723). IEEE.

Chen, Lei, Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. M. (2014). Towards automated assessment of public speaking skills using multimodal cues. In *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction* (pp. 200–203). Association for Computing Machinery, Inc. https://doi.org/10.1145/2663204.2663265

Chen, Lujie, Li, X., Xia, Z., Song, Z., Morency, L.-P., & Dubrawski, A. (2016). Riding an Emotional Roller-Coaster: A Multimodal Study of Young Child's Math Problem Solving Activities. *International Educational Data Mining Society*.

Cheng, S., Keyes, K. M., Bitfoi, A., Carta, M. G., Koç, C., Goelitz, D., … Pez, O. (2018). Understanding parent–teacher agreement of the Strengths and Difficulties Questionnaire (SDQ): Comparison across seven European countries. *International Journal of Methods in Psychiatric Research*, *27*(1), e1589.

Choi, B., Jebelli, H., & Lee, S. (2019). Feasibility analysis of electrodermal activity (EDA) acquired from wearable sensors to assess construction workers' perceived risk. *Safety Science*, *115*, 110–120.

Chollet, M., Ghate, P., Neubauer, C., & Scherer, S. (2018). Influence of Individual Differences when Training Public Speaking with Virtual Audiences. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 1–7).

Chollet, M., & Scherer, S. (2017). Assessing public speaking ability from thin slices of behavior. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 310–316).

Chollet, M., Stefanov, K., Prendinger, H., & Scherer, S. (2015). Public speaking training with a multimodal interactive virtual audience framework - Demonstration. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 367–368). Association for Computing Machinery, Inc. https://doi.org/10.1145/2818346.2823294

Chollet, M., Wörtwein, T., Morency, L.-P., & Scherer, S. (2016). A multimodal corpus for the assessment of public speaking ability and anxiety. In *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC 2016) (pp. 488–495).

Chollet, M., Wörtwein, T., Morency, L. P., Shapiro, A., & Scherer, S. (2015). Exploring feedback strategies to improve public speaking: An interactive virtual audience framework. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 1143–1154). Association for Computing Machinery, Inc. https://doi.org/10.1145/2750858.2806060

Claessens, L. C. A., van Tartwijk, J., van der Want, A. C., Pennings, H. J. M., Verloop, N., den Brok, P. J., & Wubbels, T. (2017). Positive teacher–student relationships go beyond the classroom, problematic ones stay inside. *Journal of Educational Research*, *110*(5), 478–493. https://doi.org/10.1080/00220671.2015.1129595

Costa, J., Adams, A. T., Jung, M. F., Guimbetière, F., & Choudhury, T. (2016). EmotionCheck: Leveraging bodily signals and false feedback to regulate our emotions. In *UbiComp 2016 - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 758–769). Association for Computing Machinery, Inc. https://doi.org/10.1145/2971648.2971752

Croft, R. J., Gonsalvez, C. J., Gander, J., Lechem, L., & Barry, R. J. (2004). Differential relations between heart rate and skin conductance, and public speaking anxiety. *Journal of Behavior Therapy and Experimental Psychiatry*, *35*(3), 259–271.

Daly, J. (1991). Understanding communication apprehension: An introduction for language educators. *Language Anxiety: From Theory and Research to Classroom Implications*, *9*(1), 3–13.

Daly, J. A., & Stafford, L. (1984). Correlates and consequences of social-communicative anxiety. *Avoiding Communication: Shyness, Reticence, and Communication Apprehension*, 125–143.

Daly, J. A., Vangelisti, A. L., & Lawrence, S. G. (1989). Self-focused attention and public speaking anxiety. *Personality and Individual Differences*, *10*(8), 903–913.

Daly, J. A., Vangelisti, A. L., & Weber, D. J. (1995). Speech anxiety affects how people prepare speeches: A protocol analysis of the preparation processes of speakers. *Communications Monographs*, *62*(4), 383–397.

Daly, S. (1978). Behavioural correlates of social anxiety. *British Journal of Social and Clinical Psychology*, *17*(2), 117–120. https://doi.org/10.1111/j.2044-8260.1978.tb00252.x

Damian, I., Tan, C. S. S., Baur, T., Schöning, J., Luyten, K., & André, E. (2015). Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems* (pp. 565–574).

Dawson, M., Schell, A., & Filion, D. (2017). The electrodermal system. Retrieved from https://psycnet.apa.org/record/2016-44948-010

De, T., Alcantara, S., Ferreira, J., & Maurer, F. (2013). *Interactive Prototyping of Tabletop and Surface Applications*.

Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., & Thies, W. (2012). " Yours is better!" participant response bias in HCI. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1321–1330).

Dermody, F. (2016). Multimodal positive computing system for public speaking with real-Time feedback. In *ICMI 2016 - Proceedings of the 18th ACM International*

Conference on Multimodal Interaction (pp. 541–545). New York, New York, USA: Association for Computing Machinery, Inc. https://doi.org/10.1145/2993148.2997616

Dewaele, J.-M., & Li, C. (2020). Emotions in second language acquisition: A critical review and research agenda. *Foreign Language World*, *196*(1), 34–49.

Dewaele, J., Petrides, K. V, & Furnham, A. (2008). Effects of trait emotional intelligence and sociobiographical variables on communicative anxiety and foreign language anxiety among adult multilinguals: A review and empirical investigation. *Language Learning*, *58*(4), 911–960.

Di Lascio, E., Gashi, S., Krasic, D., & Santini, S. (2017). In-classroom self-tracking for teachers and students: Preliminary findings from a pilot study. In *UbiComp/ISWC 2017 - Adjunct Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (pp. 865–870). Association for Computing Machinery, Inc. https://doi.org/10.1145/3123024.3125504

Di Lascio, E., Gashi, S., & Santini, S. (2018). Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–21. https://doi.org/10.1145/3264913

Dimberg, U., & Thunberg, M. (2007). Speech anxiety and rapid emotional reactions to angry and happy facial expressions. *Scandinavian Journal of Psychology*, *48*(4), 321–328.

Dixon, L. Q., Zhao, J., Shin, J.-Y., Wu, S., Su, J.-H., Melike, R. B.-B., … Snow, C.

(2012). What We Know About Second Language Acquisition: A Synthesis From Four Perspectives, *82*(1), 5–60. https://doi.org/10.3102/0034654311433587

Dove, G., Hansen, N. B., & Halskov, K. (2016). An Argument For Design Space Reflection. https://doi.org/10.1145/2971485.2971528

Duvall, E., Robbins, A., Graham, T., & Divett, S. (2014). Exploring filler words and their impact. *Schwa. Language & Linguistics*, *11*, 35–49.

Ebert, D. D., Mortier, P., Kaehlke, F., Bruffaerts, R., Baumeister, H., Auerbach, R. P., … Lochner, C. (2019). Barriers of mental health treatment utilization among first-year college students: First cross-national results from the WHO World Mental Health International College Student Initiative. *International Journal of Methods in Psychiatric Research*, *28*(2), e1782.

Echeverría, V., Avendaño, A., Chiluiza, K., Vásquez, A., & Ochoa, X. (2014). Presentation skills estimation based on video and kinect data analysis. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (pp. 53–60).

Egger, M., Ley, M., & Hanke, S. (2019). Emotion Recognition from Physiological Signal Analysis: A Review. *Electronic Notes in Theoretical Computer Science*, *343*, 35–55. https://doi.org/10.1016/j.entcs.2019.04.009

Ekkekakis, P. (2013). *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge University Press.

Ekman, P. (1973). Cross-cultural studies of facial expression. *Darwin and Facial Expression: A Century of Research in Review*, *169222*(1).

Elkhafaifi, H. (2005). The effect of prelistening activities on listening comprehension in

Arabic learners. *Foreign Language Annals*, *38*(4), 505–513.

Engel, F., Keary, A., Berwind, K., Bornschlegl, M. X., & Hemmje, M. (2017). The role of reproducibility in affective computing. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2008–2014). IEEE.

Epp, C. D. (2016). English language learner experiences of formal and informal learning environments. In *ACM International Conference Proceeding Series* (Vol. 25-29-April-2016, pp. 231–235). Association for Computing Machinery. https://doi.org/10.1145/2883851.2883896

Eysenck, M. W. (1979). Anxiety, learning, and memory: A reconceptualization. *Journal of Research in Personality*, *13*(4), 363–385. https://doi.org/10.1016/0092-6566(79)90001-1

Fairclough, S. (2014). Physiological data must remain confidential. *Nature News*, *505*(7483), 263.

Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual framework. *Educational Researcher*, *47*(4), 235–245.

Fiori, M., Antonietti, J.-P., Mikolajczak, M., Luminet, O., Hansenne, M., & Rossier, J. (2014). What is the ability emotional intelligence test (MSCEIT) good for? An evaluation using item response theory. *PloS One*, *9*(6), e98827.

Fleiss, J. L., Levin, B., & Paik, M. C. (1981). The measurement of interrater agreement. *Statistical Methods for Rates and Proportions*, *2*(212–236), 22–23.

Friard, O., & Gamba, M. (2016). BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and*

Evolution, *7*(11), 1325–1330.

Fung, M., Jin, Y., Zhao, R., & Hoque, M. E. (2015). ROC speak: Semi-Automated personalized feedback on nonverbal behavior from recorded videos. In *UbiComp 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 1167–1178). New York, New York, USA: Association for Computing Machinery, Inc. https://doi.org/10.1145/2750858.2804265

Gao, N., Saiedur Rahaman, M., Shao, W., & Salim, F. D. (2021). Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (pp. 237–242).

Gao, N., Shao, W., Rahaman, M. S., & Salim, F. D. (2020). n-Gage: Predicting in-class emotional, behavioural and cognitive engagement in the wild. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(3), 1–26.

Gardner, H. (1999). Who owns intelligence. *The Atlantic Monthly*, *283*(2), 67–76.

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., & Gil, Y. (2013). Quantifying reproducibility in computational biology: the case of the tuberculosis drugome. *PloS One*, *8*(11), e80278.

Garner, P. W. (2010, May 2). Emotional competence and its influences on teaching and learning. *Educational Psychology Review*. Springer. https://doi.org/10.1007/s10648-010-9129-4

Gashi, S., Lascio, E. Di, & Santini, S. (2018). Using students' physiological synchrony to quantify the classroom emotional climate. In *UbiComp/ISWC 2018 - Adjunct Proceedings of the 2018 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2018 ACM International Symposium on Wearable Computers* (pp. 698–701). Association for Computing Machinery, Inc. https://doi.org/10.1145/3267305.3267693

Gay, G. (2018). *Culturally responsive teaching: Theory, research, and practice*. Retrieved from https://books.google.com/books?hl=en&lr=&id=uD9qDwAAQBAJ&oi=fnd&pg=PP1&dq=+Gay,+G.+(2010).+Culturally+responsive+teaching:+Theory,+research,+and+practice,+2nd+edition.+NY:+Teachers++College+Press&ots=v_BA1dWiae&sig=xqNVRmzIuwRDE1p-aUDH88mSs80

Gilkinson, H. (1942). Social fears as reported by students in college speech classes. *Communications Monographs*, *9*(1), 141–160.

Giraud, T., Soury, M., Hua, J., Delaborde, A., Tahon, M., Jauregui, D. A. G., … Martin, J. C. (2013). Multimodal expressions of stress during a public speaking task: Collection, annotation and global analyses. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (pp. 417–422). https://doi.org/10.1109/ACII.2013.75

Glendinning, S. (2018). A new rootedness? Education in the technological age. *Studies in Philosophy and Education*, *37*(1), 81–96.

Goering, S., & Yuste, R. (2016). On the necessity of ethical guidelines for novel neurotechnologies. *Cell*, *167*(4), 882–885.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, *44*(5), 293–307.

Greenberg, M. T., Domitrovich, C. E., Weissberg, R. P., & Durlak, J. A. (2017). Social and emotional learning as a public health approach to education. *The Future of Children*, 13–32.

Gregersen, T. (2020). Dynamic properties of language anxiety. *Studies in Second Language Learning and Teaching*, *10*(1), 67–87.

Gregersen, T., & Horwitz, E. K. (2002). Language learning and perfectionism: Anxious and non-anxious language learners' reactions to their own oral performance. *The Modern Language Journal*, *86*(4), 562–570.

Gregersen, T., Macintyre, P. D., & Meza, M. D. (2014). The motion of emotion: Idiodynamic case studies of learners' foreign language anxiety. *Modern Language Journal*, *98*(2), 574–588. https://doi.org/10.1111/modl.12084

Gregersen, T. S. (2005). Nonverbal Cues: Clues to the Detection of Foreign Language Anxiety. *Foreign Language Annals*, *38*(3), 388–400. https://doi.org/10.1111/j.1944-9720.2005.tb02225.x

Gregersen, T. S. (2007). Language learning beyond words: Incorporating body language into classroom activities. *Reflections on English Language Teaching*, *6*(1), 51–64.

Guba, E. G., & Lincoln, Y. S. (1981). *Effective evaluation: Improving the usefulness of evaluation results through responsive and naturalistic approaches.* Jossey-Bass.

Gulliver, A., Griffiths, K. M., & Christensen, H. (2010). Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. *BMC Psychiatry*, *10*(1), 1–9.

Gupta, A., & Pathania, P. (2021). To study the impact of Google Classroom as a platform of learning and collaboration at the teacher education level. *Education and Information Technologies*, *26*(1), 843–857. https://doi.org/10.1007/s10639-020-10294-1

Güvendir, E., Oya, O., & Dündar, S. (2020). The Influence of Counsellor Trainee Support on Public Speaking and Foreign Language Speaking Anxiety in the Class Setting. *International Journal of Psychology and Educational Studies*, *7*(1), 11–26.

Hagelskamp, C., Brackett, M. A., Rivers, S. E., & Salovey, P. (2013). Improving Classroom Quality with The RULER Approach to Social and Emotional Learning: Proximal and Distal Outcomes. *American Journal of Community Psychology*, *51*(3–4), 530–543. https://doi.org/10.1007/s10464-013-9570-x

Hamre, B. K., & Pianta, R. C. (2005). Can Instructional and Emotional Support in the First-Grade Classroom Make a Difference for Children at Risk of School Failure? *Child Development*, *76*(5), 949–967. https://doi.org/10.1111/j.1467-8624.2005.00889.x

Hansen, E. I. K., & Iversen, O. S. (2013). You are the real experts! - Studying teenagers' motivation in Participatory Design. *ACM International Conference Proceeding Series*, 328–331. https://doi.org/10.1145/2485760.2485826

Hargreaves, A. (1998). The emotional practice of teaching. *Teaching and Teacher Education*.

Henning, J. H. (1934). A Study of Stage Fright Through the Comparison of Students' Reactions and Instructor's Observations Durng the Speech Situation. Northwestern University.

Hernandez, J., Lovejoy, J., McDuff, D., Suh, J., O'Brien, T., Sethumadhavan, A., … Czerwinski, M. (2021). Guidelines for Assessing and Minimizing Risks of Emotion Recognition Applications. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1–8). IEEE.

Hernandez, J., Riobo, I., Rozga, A., Abowd, G. D., & Picard, R. W. (2014). Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 307–317). ACM.

Hilleson, M. (1996). I want to talk with them, but I don't want them to hear": An introspective study of second language anxiety in an English-medium school. *Voices from the Language Classroom: Qualitative Research in Second Language Education*, 248–275.

Ho, M. R., Smyth, T. N., Kam, M., & Dearden, A. (2009). Human-computer interaction for development: The past, present, and future. *Information Technologies & International Development*, *5*(4), pp-1.

Horwitz, E. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*, 112–126.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal*, *70*(2), 125–132.

Howell, N., Chuang, J., De Kosnik, A., Niemeyer, G., & Ryokai, K. (2018). Emotional biosensing: Exploring critical alternatives. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–25.

Howell, N., Devendorf, L., Vega Gálvez, T. A., Tian, R., & Ryokai, K. (2018). Tensions of Data-Driven Reflection, 1–13. https://doi.org/10.1145/3173574.3174005

Hyson, M. C., & Lee, K. M. (1996). Assessing early childhood teachers' beliefs about emotions: Content, contexts, and implications for practice. *Early Education and Development*, *7*(1), 59–78. https://doi.org/10.1207/s15566935eed0701_5

Jackson, J. M., & Latané, B. (1981). All alone in front of all those people: Stage fright as a function of number and type of co-performers and audience. *Journal of Personality and Social Psychology*, *40*(1), 73.

Jiang, B., Annis, K., & Lim, W. (2016). Developing Culturally Responsive First-Year Teachers for Urban Schools. *Journal of Classroom Interaction*, *51*(1), 18–30. Retrieved from https://www.jstor.org/stable/26174347?seq=1&cid=pdf-reference#references_tab_contents

Joffe, H. (2012). Thematic analysis. *Qualitative Research Methods in Mental Health and Psychotherapy*, *1*.

Jun Zhang, L. (2001). Exploring variability in language anxiety: Two groups of PRC students learning ESL in Singapore. *RELC Journal*, *32*(1), 73–91.

Kagan, N. (1967). Studies in Human Interaction, Interpersonal Process Recall Stimulated By Videotape. Retrieved from http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED017946

Kakoulli Constantinou, E. (2021). *Using the G Suite for Education in Language Teacher Education: Benefits and Challenges*. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 12784 LNCS). Springer International Publishing. https://doi.org/10.1007/978-3-030-77889-7_21

Kamiya, N. (2019). What Factors Affect Learners' Ability to Interpret Nonverbal Behaviors in EFL Classrooms? *Journal of Nonverbal Behavior*, *43*(3), 283–307.

Kaneko, D., Stuldreher, I., Reuten, A. J. C., Toet, A., Van Erp, J. B. F., & Brouwer, A.-M. (2021). Comparing explicit and implicit measures for assessing cross-cultural food experience. *Frontiers in Neuroergonomics*, *2*, 5.

Kaur, R. &. (2011). Technical oral presentations in English: Qualitative analysis of Malaysian engineering undergraduates' sources of anxiety. *Procedia-Social and Behavioral Sciences*, *29*, 1436–1445.

Keltner, D., & Anderson, C. (2000). Saving face for Darwin: The functions and uses of embarrassment. *Current Directions in Psychological Science*, *9*(6), 187–192.

Kendall, M. (1948). Rank correlation methods. Retrieved from https://psycnet.apa.org/record/1948-15040-000

Khan, Z. A., & Zafar, S. (2010). The Effects of Anxiety on Cognitive Processing in English Language Learning. *English Language Teaching*, *3*(2), 199–209.

Kimani, E., & Bickmore, T. (2019). Addressing Public Speaking Anxiety in Real-time Using a virtual public speaking coach and physiological sensors. In *IVA 2019 - Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 260–263). Association for Computing Machinery, Inc. https://doi.org/10.1145/3308532.3329409

Kirsch, J., Siltanen, C., Zhou, Q., Revzin, A., & Simonian, A. (2013). Biosensor technology: recent advances in threat agent detection and medicine. *Chemical Society Reviews*, *42*(22), 8733–8768.

Kleinmann, H. H. (1977). AVOIDANCE BEHAVIOR IN ADULT SECOND

LANGUAGE ACQUISITION [1]. *Language Learning*, *27*(1), 93–107. https://doi.org/10.1111/j.1467-1770.1977.tb00294.x

Knowles, B., & Hanson, V. L. (2018). The wisdom of older technology (non) users. *Communications of the ACM*, *61*(3), 72–77.

Kothgassner, O. D., Felnhofer, A., Hlavacs, H., Beutl, L., Palme, R., Kryspin-Exner, I., & Glenk, L. M. (2016). Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment. *Computers in Human Behavior*, *62*, 124–135. https://doi.org/10.1016/j.chb.2016.03.081

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*, 14–41. https://doi.org/10.1016/j.biopsycho.2010.03.010

Kritikos, J., Tzannetos, G., Zoitaki, C., Poulopoulou, S., & Koutsouris, P. D. (2019). Anxiety detection from Electrodermal Activity Sensor with movement interaction during Virtual Reality Simulation. In *International IEEE/EMBS Conference on Neural Engineering, NER* (Vol. 2019-March, pp. 571–576). IEEE Computer Society. https://doi.org/10.1109/NER.2019.8717170

Kye, S., Moon, J., Lee, J., Choi, I., Cheon, D., & Lee, K. (2017). Multimodal data collection framework for mental stress monitoring. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers* (pp. 822–829).

Lakshmi, K. R., Nelakurthi, H., Kumar, A. S., & Rudraraju, A. (2017). Oral fluid-based biosensors: A novel method for rapid and noninvasive diagnosis. *Indian Journal of Dental Sciences*, *9*(1), 60.

Lane, H. C., & D’Mello, S. K. (2019). Uses of physiological monitoring in intelligent learning environments: A review of research, evidence, and technologies. *Mind, Brain and Technology*, 67–86.

Leddy, J. J., Moineau, G., Puddester, D., Wood, T. J., & Humphrey-Murto, S. (2011). Does an emotional intelligence test correlate with traditional measures used to determine medical school admission? *Academic Medicine : Journal of the Association of American Medical Colleges*, *86*(10 Suppl), 4–7. https://doi.org/10.1097/acm.0b013e31822a6df6

Lee, H., & Kleinsmith, A. (2019). Public Speaking Anxiety in a Real Classroom: Towards Developing a Reflection System. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (p. LBW2417).

Li, H., Kwong, S., Yang, L., Huang, D., & Xiao, D. (2011). Hilbert-Huang transform for analysis of heart rate variability in cardiac health. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *8*(6), 1557–1567.

Lim, Y. K., Stolterman, E., & Tenenberg, J. (2008). The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction*, *15*(2). https://doi.org/10.1145/1375761.1375762

Liu, M. (2006). Anxiety in EFL classrooms: Causes and consequences. *TESL Reporter*, *39*(1), 13–32.

Lomas, C. W. (1934). A Study of Stage Fright as Measured by Student Reactions to the Speech Situaton. Northwestern University.

Lopez, F. A. (2012). Moderators of language acquisition models and reading achievement for English language learners: The role of emotional warmth and instructional support. *Arizona.Pure.Elsevier.Com*. Retrieved from https://arizona.pure.elsevier.com/en/publications/moderators-of-language-acquisition-models-and-reading-achievement

Luo, H. (2014). Foreign language speaking anxiety: A study of Chinese language learners. *Journal of the National Council of Less Commonly Taught Languages*, *15*(1), 99–117.

Ma, L. P. F. (2012). Strengths and weaknesses of NESTS and NNESTs: Perceptions of NNESTs in Hong Kong. *Linguistics and Education*, *23*(1), 1–15.

MacIntyre, P. D., & Gardner, R. C. (1991). Methods and results in the study of anxiety and language learning: A review of the literature. *Language Learning*, *41*(1), 85–117.

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language Learning*, *44*(2), 283–305.

Mackay, W. E. (1995). Ethics, lies and videotape…. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 138–145).

MacLean, D., Roseway, A., & Czerwinski, M. (2013). MoodWings: a wearable biofeedback device for real-time stress intervention. In *Proceedings of the 6th international conference on PErvasive Technologies Related to Assistive Environments* (pp. 1–8).

Mahn, H., & John-Steiner, V. (2002). The gift of confidence: A Vygotskian view of emotions. *Learning for Life in the 21st Century*, 46–58.

Mairi, S., Febriani, S., & Solusia, C. (2018). An Analysis of Chinese Universities

Students' Attitudes towards Native and Non-Native English-Speaking Teachers (NESTs/NNESTs). *Online Submission*, *5*(2), 86–96.

Mandinach, E. B., Parton, B. M., Gummer, E. S., & Anderson, R. (2015). Ethical and appropriate data use requires data literacy. *Phi Delta Kappan*, *96*(5), 25–28.

Mayer, J. D., DiPaolo, M., & Salovey, P. (1990). Perceiving affective content in ambiguous visual stimuli: A component of emotional intelligence. *Journal of Personality Assessment*, *54*(3–4), 772–781.

Mayer, J. D., Salovey, P., & Caruso, D. R. (2002). Mayer-Salovey-Caruso emotional intelligence test (MSCEIT) item booklet.

McCaughtry, N., & Rovegno, I. (2003). Development of pedagogical content knowledge: Moving from blaming students to predicting skillfulness, recognizing motor development, and understanding emotion. *Journal of Teaching in Physical Education*, *22*(4), 355–368.

McCroskey, J. C. (1970). Measures of communication-bound anxiety.

McEvoy, L. K., Smith, M. E., & Gevins, A. (2000). Test–retest reliability of cognitive EEG. *Clinical Neurophysiology*, *111*(3), 457–463.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

McStay, A., & Pavliscak, P. (2019). Emotional artificial intelligence: Guidelines for ethical use. *COMEST/UNESCO*.

Meehan, M., Insko, B., Whitton, M., & Brooks Jr, F. P. (2002). Physiological measures of presence in stressful virtual environments. *Acm Transactions on Graphics (Tog)*, *21*(3), 645–652.

Menzel, W., Herron, D., Bonaventura, P., & Morton, R. (2000). Automatic detection and correction of non-native English pronunciations. *Proceedings of INSTILL*, 49–56.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Merrill, N., Chuang, J., & Cheshire, C. (2019). Sensing is believing: what people think biosensors can reveal about thoughts and feelings. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 413–420).

Mestanlk, M., Vlsnovcova, Z., & Tonhajzerova, I. (2014). The assessment of the autonomic response to acute stress using electrodermal activity. *Acta Medica Martiniana*, *14*(2), 5–9.

Metallinou, A., & Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (pp. 1–8). IEEE.

Meyer, D. K., & Turner, J. (2007). Chapter 14. Scaffolding Emotions in Classrooms Assessment issues View project Edited Book on Teaching Motivation for Student Engagement View project. https://doi.org/10.1016/B978-012372545-5/50015-0

Mierswa, I., & Klinkenberg, R. (2019). Rapidminer studio.

Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017). Put your thinking cap on: Detecting cognitive load using EEG during learning. In *ACM International Conference Proceeding Series* (pp. 80–89).

Association for Computing Machinery. https://doi.org/10.1145/3027385.3027431

Miyakawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain*. Springer.

Moskowitz, G. (1965). The Fearsome Foreign Language Hour. *The French Review*, *38*(6), 781–786.

Munn, N. L. (1940). The effect of knowledge of the situation upon judgment of emotion from facial expressions. *The Journal of Abnormal and Social Psychology*, *35*(3), 324.

Murali, P., Trinh, H., Ring, L., & Bickmore, T. (2021). A Friendly Face in the Crowd: Reducing Public Speaking Anxiety with an Emotional Support Agent in the Audience. *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents, IVA 2021*, 156–163. https://doi.org/10.1145/3472306.3478364

Na, Z. (2007). A study of high school students' English learning anxiety. *The Asian EFL Journal*, *9*(3), 22–34.

Neale, J. (2016). Iterative categorization (IC): a systematic technique for analysing qualitative data. *Addiction*, *111*(6), 1096–1106. https://doi.org/10.1111/add.13314

Neumann, E., & Blanton, R. (1970). The early history of electrodermal research. *Psychophysiology*, *6*(4), 453–475.

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, *2*(2), 92–105.

Nieters, J. E., Ivaturi, S., & Ahmed, I. (2007). Making personas memorable. *Conference on Human Factors in Computing Systems - Proceedings*, 1817–1824.

https://doi.org/10.1145/1240866.1240905

Norman, D., & Draper, S. W. (1986). User centered design: new perspectives on human-computer interaction. *Hillsdale, NJ: Lawrence Erlbaurn Associates*.

O'Connor, R. C., Rasmussen, S., & Hawton, K. (2010). Predicting depression, anxiety and self-harm in adolescents: The role of perfectionism and acute life stress. *Behaviour Research and Therapy*, *48*(1), 52–59.

Onwuegbuzie, A. J., Bailey, P., & Daley, C. E. (1999). *Factors associated with foreign language anxiety*. *Applied Psycholinguistics* (Vol. 20).

Pali, M., Jagannath, B., Lin, K.-C., Upasham, S., Sankhalab, D., Upashama, S., … Prasad, S. (2021). CATCH (Cortisol Apta WATCH):'Bio-mimic alarm'to track Anxiety, Stress, Immunity in human sweat. *Electrochimica Acta*, *390*, 138834.

Park, G. P., & French, B. F. (2013). Gender differences in the Foreign Language Classroom Anxiety Scale. *System*, *41*(2), 462–471. https://doi.org/10.1016/j.system.2013.04.001

Park, S., Shoemark, P., & Morency, L. P. (2014). Toward crowdsourcing micro-level behavior annotations: The challenges of interface, training, and generalization. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (pp. 37–46). https://doi.org/10.1145/2557500.2557512

Parkinson, B., Totterdell, P., Briner, R. B., & Reynolds, S. (1996). *Changing moods: The psychology of mood and mood regulation*. Longman London.

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, *9*(DEC), 660. https://doi.org/10.3389/fnhum.2015.00660

Parthasarathy, S., Cowie, R., & Busso, C. (2016). Using agreement on direction of change to build rank-based emotion classifiers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2108–2121.

Pataranutaporn, P., Jain, A., Johnson, C. M., Shah, P., & Maes, P. (2019). Wearable lab on body: combining sensing of biochemical and digital markers in a wearable device. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 3327–3332). IEEE.

Patton, M. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Retrieved from https://books.google.com/books?hl=en&lr=&id=ovAkBQAAQBAJ&oi=fnd&pg=PP1&dq=patton,+2015,+qualitative+research+and+evaluation+method&ots=ZQ2-4ryJIZ&sig=NcVeoMexPEf970cQbCZSNhw-gGA

Persico, D., Manca, S., & Pozzi, F. (2014). Adapting the technology acceptance model to evaluate the innovative potential of e-learning systems. *Computers in Human Behavior*, *30*, 614–622. https://doi.org/10.1016/j.chb.2013.07.045

Pertaub, D. P., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and Virtual Environments*, *11*(1), 68–78. https://doi.org/10.1162/105474602317343668

Petrovica, S., Anohina-Naumeca, A., & Ekenel, H. K. (2017). Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science*, *104*, 437–444.

Phillips, E. M. (1992). The effects of language anxiety on students' oral test performance and attitudes. *The Modern Language Journal*, *76*(1), 14–26.

Phuong-Mai, N., Terlouw, C., & Pilot, A. (2005). Cooperative learning vs Confucian heritage culture's collectivism: confrontation to reveal some cultural conflicts and mismatch. *Asia Europe Journal*, *3*(3), 403–419.

Pianta, R., Paro, K. La, & Hamre, B. (2008). *Classroom Assessment Scoring System^{TM}: Manual K-3.* Retrieved from https://psycnet.apa.org/record/2007-18799-000

Picard, R. W., Fedor, S., & Ayzenberg, Y. (2016). Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review*, *8*(1), 62–75.

Pijeira-Díaz, H. J., Drachsler, H., Järvelä, S., & Kirschner, P. A. (2016). Investigating collaborative learning success with physiological coupling indices based on electrodermal activity. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 64–73).

Plattner, H., Meinel, C., & Weinberg, U. (2009). *Design-thinking*. Springer.

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, *57*(5), 1243–1252.

Poulou, M. S. (2020). Students' adjustment at school: The role of teachers' need satisfaction, teacher–student relationships and student well-being. *School Psychology International*, *41*(6), 499–521.

Pozo-Rico, T., & Sandoval, I. (2020). Can Academic Achievement in Primary School Students Be Improved Through Teacher Training on Emotional Intelligence as a Key Academic Competency? *Frontiers in Psychology*, *10*, 2976.

Price, M. Lou. (1991). The subjective experience of foreign language anxiety: Interviews with highly anxious students. *Language Anxiety: From Theory and Research to Classroom Implications*, 101–108.

Qu, Z., Chen, J., Li, B., Tan, J., Zhang, D., & Zhang, Y. (2020). Measurement of high-school students' trait math anxiety using neurophysiological recordings during math exam. *IEEE Access*, *8*, 57460–57471.

Quezada, R. L., Talbot, C., & Quezada-Parker, K. B. (2020). From bricks and mortar to remote teaching: A teacher education program's response to COVID-19. *Journal of Education for Teaching*, *46*(4), 472–483.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Richards, L. (1999). Data Alive! The Thinking Behind NVivo. *Qualitative Health Research*, *9*(3), 412–428. https://doi.org/10.1177/104973239900900310

Richardson, S. (2020). Affective computing in the modern workplace. *Business Information Review*, *37*(2), 78–85.

Richmond, V. P., McCroskey, J. C., & Hickson, M. (2008). *Nonverbal behavior in interpersonal relations*. Allyn & Bacon.

Riedl, R., Fischer, T., & Léger, P.-M. (2017). A Decade of NeuroIS Research: Status Quo, Challenges, and Future Directions. In *ICIS*.

Roggema, R. (2013). *The design charrette: ways to envision sustainable futures*. Springer Science & Business Media.

Rosiek, J. (2003). Emotional scaffolding: An exploration of the teacher knowledge at the intersection of student emotion and the subject matter. *Journal of Teacher*

Education, *54*(5), 399–412.

Roulston, K. (2014). *Reflective Interviewing: A Guide to Theory and Practice*. *Reflective Interviewing: A Guide to Theory and Practice*. SAGE Publications Ltd. https://doi.org/10.4135/9781446288009

Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *Interactions*, *3*(1), 76–85. https://doi.org/10.1145/223500.223514

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102.

Sanches, P., Janson, A., Karpashevich, P., Nadal, C., Qu, C., Roquet, C. D., … Sas, C. (2019). HCI and Affective Health: Taking stock of a decade of studies and chart-ing future research directions, 17. https://doi.org/10.1145/3290605.3300475

Sano, A., & Picard, R. W. (2013). Stress recognition using wearable sensors and mobile phones. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013* (pp. 671–676). https://doi.org/10.1109/ACII.2013.117

Santamaria, L. J. (2009). Culturally Responsive Differentiated Instruction: Narrowing Gaps Between Best Pedagogical Practices Benefiting All Learners. *Teachers College Record*, *111*(1), 214–247.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: reactions to tests. *Journal of Personality and Social Psychology*, *46*(4), 929.

Sato, H., & Kawahara, J. ichiro. (2011). Selective bias in retrospective self-reports of negative mood states. *Anxiety, Stress and Coping*, *24*(4), 359–367. https://doi.org/10.1080/10615806.2010.543132

Schlenker, B. R., & Leary, M. R. (1982a). Audiences' reactions to self-enhancing, self-denigrating, and accurate self-presentations. *Journal of Experimental Social Psychology*, *18*(1), 89–104.

Schlenker, B. R., & Leary, M. R. (1982b). Social anxiety and self-presentation: A conceptualization model. *Psychological Bulletin*, *92*(3), 641.

Schneider, J., Börner, D., Van Rosmalen, P., & Specht, M. (2015). Presentation trainer, your public speaking multimodal coach. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 539–546). New York, New York, USA: Association for Computing Machinery, Inc. https://doi.org/10.1145/2818346.2830603

Schoenberg, P. L. A., & David, A. S. (2014). Biofeedback for psychiatric disorders: a systematic review. *Applied Psychophysiology and Biofeedback*, *39*(2), 109–135.

Schütz, R. (2007). Stephen Krashen's Theory of Second Language Acquisition. *English Made in Brazil*, *2*(2), 1–3. https://doi.org/10.1186/1752-0509-4-28

Schwarzer, R. (2013). *Self-related cognitions in anxiety and motivation*. Psychology Press.

Schwerdtfeger, A. (2004). Predicting autonomic reactivity to public speaking: don't get fixed on self-report data! *International Journal of Psychophysiology*, *52*(3), 217–224.

Scovel, T. (1978). THE EFFECT OF AFFECT ON FOREIGN LANGUAGE LEARNING: A REVIEW OF THE ANXIETY RESEARCH. *Language Learning*, *28*(1), 129–142. https://doi.org/10.1111/j.1467-1770.1978.tb00309.x

Sefelin, R., Tscheligi, M., & Giller, V. (2003). Paper prototyping - What is it good for? A comparison of paper- and computer-based low-fidelity prototyping. In *Conference on Human Factors in Computing Systems - Proceedings* (pp. 778–779). https://doi.org/10.1145/765891.765986

Selwyn, N. (2019). *Should robots replace teachers?: AI and the future of education*. John Wiley & Sons.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2009). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine*, *14*(2), 410–417.

Sevinç, Yesim. (2018). Language anxiety in the immigrant context: Sweaty palms? *International Journal of Bilingualism*, *22*(6), 717–739.

Sevinç, Yeşim. (2017). Language anxiety in the immigrant context: Sweaty palms? *International Journal of Bilingualism*, 1367006917690914.

Sevinç, Yeşim, & Backus, A. (2019). Anxiety, language use and linguistic competence in an immigrant context: a vicious circle? *International Journal of Bilingual Education and Bilingualism*, *22*(6), 706–724. https://doi.org/10.1080/13670050.2017.1306021

Sharma, K., Papamitsiou, Z., & Giannakos, M. (2019). Building pipelines for educational data using AI and multimodal analytics: A “grey-box” approach. *British Journal of Educational Technology*, *50*(6), 3004–3031.

Shields, S. A. (1984). Distinguishing between emotion and nonemotion: Judgments about experience. *Motivation and Emotion*, *8*(4), 355–369.

Shim, J. M. (2019). Towards a more nuance understanding of white teachers of English learners: Increasing capacity and capabilities. *Journal of Language Teaching and Research*, *10*(1), 1.

Shim, J. M., & Shur, A. M. (2018). Learning from ELLs' Perspectives: Mismatch between ELL and Teacher Perspectives on ELL Learning Experiences. *English Language Teaching*, *11*(1), 21–32.

Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., … Yang, X. (2018, July 1). A review of emotion recognition using physiological signals. *Sensors (Switzerland)*. MDPI AG. https://doi.org/10.3390/s18072074

Shweta, Bajpai, R. C., & Chaturvedi, H. K. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, *41*(Special Issue 3), 20–27.

Siegman, A. W., & Pope, B. (1965). Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. *Journal of Personality and Social Psychology*, *2*(4), 522–530. https://doi.org/10.1037/h0022491

Singh, R. K., Agrawal, S., Sahu, A., & Kazancoglu, Y. (2021). Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda. *The TQM Journal*.

Spielberger, C. D. (1983). State-trait anxiety inventory for adults.

Steiner, D. D. (1989). Immediate and Delayed Primacy and Recency Effects in Performance Evaluation. *Article in Journal of Applied Psychology*. https://doi.org/10.1037/0021-9010.74.1.136

Steinert, S., & Friedrich, O. (2020). Wired emotions: Ethical issues of affective brain–computer interfaces. *Science and Engineering Ethics*, *26*(1), 351–367.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research, and Evaluation*, *9*(1), 4.

Stewart, D. W., Shamdasani, P. N., & Rook, D. W. (2007). Analyzing focus group data. *Focus Groups: Theory and Practice*, *20*.

Takac, M., Collett, J., Blom, K. J., Conduit, R., Rehm, I., & De Foe, A. (2019). Public speaking anxiety decreases within repeated virtual reality training sessions. *PloS One*, *14*(5).

Tang, T. T. T., Reilly, J., & Dickson, J. M. (2012). Attitudes toward seeking professional psychological help among Chinese students at a UK university. *Counselling and Psychotherapy Research*, *12*(4), 287–293.

Tanveer, M. (2007). Investigation of the factors that cause language anxiety for ESL/EFL learners in learning speaking skills and the influence it casts on communication in the target language. *University of Glasgow, Scotland*.

Tanveer, M. I., Lin, E., & Hoque, M. E. (2015). Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 286–295). ACM.

Tanveer, M. I., Zhao, R., Chen, K., Tiet, Z., & Hoque, M. E. (2016). AutoManner: An automated interface for making public speakers aware of their mannerisms. In *International Conference on Intelligent User Interfaces, Proceedings IUI* (Vol. 07-10-Marc, pp. 385–396). Association for Computing Machinery. https://doi.org/10.1145/2856767.2856785

Tarafdar, M., DArcy, J., Turel, O., & Gupta, A. (2015). The dark side of information technology. *MIT Sloan Management Review*, *56*(2), 61.

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (Vol. 2015-Novem, pp. 1934–1937). https://doi.org/10.1109/EMBC.2015.7318762

Tobias, S. (1979). Anxiety research in educational psychology. *Journal of Educational Psychology*, *71*(5), 573–582. https://doi.org/10.1037/0022-0663.71.5.573

Torrico, D. D., Fuentes, S., Viejo, C. G., Ashman, H., & Dunshea, F. R. (2019). Cross-cultural effects of food product familiarity on sensory acceptability and non-invasive physiological responses of consumers. *Food Research International*, *115*, 439–450.

Trinh, H., Asadi, R., Edge, D., & Bickmore, T. (2017). RoboCOP. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 1–24. https://doi.org/10.1145/3090092

Trinh, Ha, Asadi, R., Edge, D., & Bickmore, T. (2017). RoboCOP: A Robotic Coach for Oral Presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(2), 27.

Umair, M. (2021). Design and Evaluation of Biofeedback: Interfaces for Awareness and Regulation of Affect. Lancaster University.

Umematsu, T., Sano, A., Taylor, S., & Picard, R. W. (2019). Improving students' daily life stress forecasting using LSTM neural networks. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (pp. 1–4). IEEE.

Usakli, A. B. (2010). Improvement of EEG Signal Acquisition: An Electrical Aspect for

State of the Art of Front End. *Computational Intelligence and Neuroscience*, *2010*. https://doi.org/10.1155/2010/630649

Vandenbroucke, L., Spilt, J., Verschueren, K., Piccinin, C., & Baeyens, D. (2018). The classroom as a developmental context for cognitive development: A meta-analysis on the importance of teacher–student interactions for children's executive functions. *Review of Educational Research*, *88*(1), 125–164.

Vashistha, A., Okeke, F., Anderson, R., & Dell, N. (2018). 'You Can Always Do Better!" The Impact of Social Proof on Participant Response Bias. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1–13).

Wagner, J., Kim, J., & André, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *2005 IEEE international conference on multimedia and expo* (pp. 940–943). IEEE.

Walqui, A. (2006). Scaffolding instruction for english language learners: A conceptual framework. *International Journal of Bilingual Education and Bilingualism*. Multilingual Matters Ltd. https://doi.org/10.1080/13670050608668639

Wang, N. (2005). Beliefs about language learning and foreign language anxiety: A study of university students learning English as a foreign language in Mainland China.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., … Campbell, A. T. (2014). StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 3–14).

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of*

Consulting and Clinical Psychology, *33*(4), 448.

Waycott, J., Vetere, F., Pedell, S., Morgans, A., Ozanne, E., & Kulik, L. (2016). Not for me: Older adults choosing not to participate in a social isolation intervention. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 745–757).

Weinreich, H., Obendorf, H., Herder, E., & Mayer, M. (2008). Not quite the average: An empirical study of Web use. *ACM Transactions on the Web (TWEB)*, *2*(1), 1–31.

Welsh, E. (2002). Dealing with data: Using NVivo in the qualitative data analysis process. *Forum Qualitative Sozialforschung*, *3*(2). https://doi.org/10.17169/fqs-3.2.865

Weyns, D. (2020). Towards a code of ethics for autonomous and self-adaptive systems. *Proceedings - 2020 IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, SEAMS 2020*, 163–165. https://doi.org/10.1145/3387939.3391567

Wörtwein, T., Morency, L. P., Chollet, M., Stiefelhagen, R., Schauerte, B., & Scherer, S. (2015). Multimodal public speaking performance assessment. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 43–50). Association for Computing Machinery, Inc. https://doi.org/10.1145/2818346.2820762

Yaikhong, K., & Usaha, S. (2012). English Language Teaching. *ERIC*, *5*(12). https://doi.org/10.5539/elt.v5n12p23

Yalçın, Ö., & İnceçay, V. (2014). Foreign language speaking anxiety: The case of spontaneous speaking activities. *Procedia-Social and Behavioral Sciences*, *116*, 2620–2624. https://doi.org/10.1016/j.sbspro.2014.01.623

Young, Dolly J. (1986). The Relationship Between Anxiety and Foreign Language Oral Proficiency Ratings. *Foreign Language Annals*, *19*(5), 439–445. https://doi.org/10.1111/j.1944-9720.1986.tb01032.x

Young, Dolly Jesusita. (1990). *An Investigation of Students' Perspectives on Anxiety and Speaking*. *Foreign Language Annals* (Vol. 23). https://doi.org/10.1111/j.1944-9720.1990.tb00424.x

Zangróniz, R., Martínez-Rodrigo, A., Pastor, J. M., López, M. T., & Fernández-Caballero, A. (2017). Electrodermal Activity Sensor for Classification of Calm/Distress Condition. *Sensors*, *17*(10), 2324.

Zhang, J., & Pelttari, C. (2014). Exploring the emotions and needs of English language learners: facilitating pre-service and in-service teachers' recognition of the tasks facing language learners. *Journal of Multilingual and Multicultural Development*, *35*(2), 179–194. https://doi.org/10.1080/01434632.2013.822505

Zhang, X., Li, W., Chen, X., & Lu, S. (2018). Moodexplorer: Towards compound emotion detection via smartphone sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(4), 1–30.

Zheng, Y., Wong, T. C. H., Leung, B. H. K., & Poon, C. C. Y. (2016). Unobtrusive and multimodal wearable sensing to quantify anxiety. *IEEE Sensors Journal*, *16*(10), 3689–3696.