Improving music education using biofeedback-based cognitive tutoring

DARTMOUTH

Jordan Sanz

Honors Thesis

David Kraemer, Advisor

Xia Zhou, Advisor

Program in Quantitative Social Science

Dartmouth College

May 17, 2022

Contents

Acknowledgements

I would first like to begin by recognizing the Neukom Institute for Computational Science and the Neukom Scholars Program for their generous grant. Without their support, this study would not have been possible.

Next, I would like to thank Professor Kraemer for his invaluable support. From countless emails and many Zoom calls, I genuinely could not have written this thesis without you. I cannot thank you enough for supporting me and believing in me throughout the entire process. I'd also like to thank Professor Zhou for her valuable advice at the start of this process in creating the application, as well as the entire Cognitive Neuroscience of Learning Lab team and the Dartmouth Reality and Robotics Lab, especially Megan Hillis, Yeongji Lee, Professor Balkcom, and Julien Blanchet. I'd additionally like to thank Professor Herron for his advice throughout the entire Quantitative Social Science thesis process and Professor Mahoney for his guidance in the revision process. I'd also like to thank the Warbler team for working with me on the original version of the web app and rhythm game, including Jacob Donoghue, Shane Hewitt, Gia Kim, and Sathvi Korandla.

To Professor Tim Tregubov and Professor Vasanta Lakshmi Kommineni, thank you both for believing in me and inspiring my love of the intersection between computer science, statistics, and education. Without the two of you, my Dartmouth career would have looked much more bleak and unexciting, so I owe much to the both of you.

Finally, I would like to thank my friends and family for their great support during the three terms of this process. You helped remind me why I was spending so much time on this important topic when I forgot, and you helped reassure me that I was capable of conducting research and development at this level.

Abstract

With the rise of intelligent tutoring systems, otherwise known as cognitive tutors, education has seen a new and effective format for humanless, individualized instruction. However, most cognitive tutoring fails to implement information called biofeedback, or being able to sense human emotion as students learn. Additionally, cognitive tutors exist primarily within STEM fields and have not been studied in-depth within music education. Previous research has not created nor studied a cognitive tutor that utilizes biofeedback to provide personalized instruction in teaching rhythm. To add to the literature, I create an affective cognitive tutor as a web app that utilizes the webcam to teach drumming notation rhythms by pressing keys on the keyboard on specified beats as a proof of concept. I test 59 participants in a 45 minute, randomized controlled trial to answer two questions: first, does displayed affect correlate with rhythm performance; and second, does incorporating biofeedback in a cognitive tutor to teach music aid in the process of student learning? There are three main findings: the emotions of fear and surprise are correlated the strongest with rhythm performance, both correlated in the positive direction; incorporating biofeedback from a webcam into cognitive tutors for rhythm education had little effect on overall performance in the short-term; and a cognitive tutor within the field of music education seems to be successful and helpful to student learning.

1 Introduction

With the advancement of technology, many successful tutoring applications have emerged within education (Ritter et al., 2015, Camilleri and Camilleri, 2019). Applications such as Duolingo have been shown to be effective teachers (Loewen et al., 2019, Ajisoko, 2020) and students are adopting these applications quickly (Badri, 2021). Thus, technology is becoming not only a helping tool in learning, but technology is becoming the tutor itself. Many tutoring applications are known as intelligent tutoring systems, or "cognitive tutors", a new format for effective, humanless, and individualized instruction (Anderson et al., 1995; Koedinger et al., 1997; Ritter et al., 2007; VanLehn, 2011; Ma et al., 2014; Pane et al., 2014; Supekar et al., 2015; Marouf and Abu-Naser, 2019; Mousavinasab et al., 2021). However, the literature about cognitive tutors is limited by two main factors: a lack of focus on student biofeedback and the lack of non-STEM-focused cognitive tutors.

A stark difference between many cognitive tutoring applications and human, face-to-face, in-person tutoring is utilizing biofeedback—being able to sense students' emotions as the students learn (Schwartz and Andrasik, 2017). Most cognitive tutors will utilize artificial intelligence to adapt their instruction to a student's learning pace where the instruction rate is determined solely by the student's performance on questions; for example, if the student performs more accurately on questions, the pace of teaching will accelerate (Anderson et al., 1995). However, in-person tutoring allows for modifying a learning pace not just by how accurately a student performs with the material, but also by watching and considering the student's emotional expressions (Alexander et al., 2008). The literature has studied a few cognitive tutors that utilize biofeedback: for instance, one tutor understands affect through speech (Grawemeyer et al., 2017) and another understands affect through webcam video (Zakharov, 2007; Spaulding et al., 2016). Still, the literature lacks cognitive tutors that can utilize biofeedback signals from the student to adapt personalized instruction in real-time, possibly limiting how effective a cognitive tutoring application can be as a substitute for

in-person learning.

Additionally, cognitive tutors mainly specialize and have been studied mostly in STEM fields (Ritter et al., 2007; Ma et al., 2014; Pane et al., 2014; Supekar et al., 2015; Marouf and Abu-Naser, 2019). The literature has not studied cognitive tutors within music education, a field of education with pronounced biofeedback such as anxiety (Egilmez, 2012; Wristen, 2013; Patston and Osborne, 2016). With visible affect like anxiety emerging in students, music education presents a rich resource of biofeedback for affect-based cognitive tutors.

The current study will contribute to the literature by determining the relationship between biofeedback and music education performance—namely, rhythm performance. This study will explore whether or not the relationship between biofeedback and music education can be utilized to create more effective and human-like cognitive tutors. The current study has two main research questions: first, does the amount of webcam biofeedback emitted from a student learning rhythm correlate with performance in music education, and if so, how do specific emotions displayed in a student correlate with the student's rhythm performance? Second, will a cognitive tutoring system that incorporates biofeedback signals from a webcam into its personalized instruction algorithm improve student performance within music education?

To answer these questions, I build a cognitive tutor that teaches musical rhythm while also monitoring biofeedback through webcam data, using the student's emotional biofeedback to construct an individualized learning plan. I then conduct a randomized controlled trial experiment on 39 subjects, allowing the subjects to learn from the new cognitive tutoring application over the span of thirty-five minutes, tracking their accuracy, timing error, and emotional biofeedback from the webcam. I analyze this data using a series of mixed models and correlations to first determine the relationship between a student's displayed affect of 6 emotions and the student's rhythm performance. Then, I test 20 participants by including biofeedback within the cognitive tutoring model to allow for weighted emotion data to help

determine a student's understanding of a lesson. I analyze this data using a two-tailed t-test to understand if including emotional biofeedback signals from a webcam into the cognitive tutoring algorithm can improve and expedite student learning within music education. Understanding the effectiveness of including biofeedback in cognitive tutoring in areas outside of STEM pushes education to continue revising and improving online learning, providing stronger and more effective resources to students beyond face-to-face tutoring.

2 Literature Review

2.1 Cognitive Tutor

Cognitive tutoring is a method of individualized instruction based on the Adaptive Control of Thought (ACT) theory of cognition (Anderson et al., 1995; Ritter et al., 2007; Anderson, 2013). Cognitive tutoring aims to utilize a student's skill profile to choose problems that emphasize the skills and subskills where the student performs the weakest (Pane et al., 2014). In this way, the goal of cognitive tutors is to be as effective as intelligent, human, face-to-face tutors (Anderson et al., 1985). The first two cognitive tutors built attempted to replicate effective human tutors with one tutor focusing on high school geometry, and another on programming; from limited testing, the tutors were effective in both teaching the subject and convincing students to enjoy the learning subject (Anderson et al., 1985).

Beginning with these first two cognitive tutors and expanding to many more, cognitive tutors showed relative and immediate success compared to their human-tutoring counterparts (Council et al., 2003; VanLehn, 2011; Ma et al., 2014). Three types of cognitive tutors were tested in their effectiveness of teaching mathematics: one tutor used answer-based learning (a student simply enters in an answer), one tutor used step-based learning (a student must enter in each step required to find an answer), and one tutor used substep-based learning (a student must enter in each substep to each step required to find an answer); all three tutors were tested against human tutoring and no tutoring (VanLehn, 2011). When compared to

no tutoring, human tutoring had an effect size of 0.79, while step-based cognitive tutors had an effect size of 0.76, showing almost the same effectiveness as an in-person human tutor (VanLehn, 2011). Intelligent tutors were also found to have greater achievement levels in teaching mathematics than teacher-led large group instruction, computer-based instruction without a cognitive tutor, and textbooks or workbooks (Ma et al., 2014). Furthermore, there was no significant difference between an intelligent tutoring system and individualized human tutoring (Ma et al., 2014).

Many cognitive tutors have been created, particularly within the STEM fields (Koedinger et al., 1997; Ritter et al., 2007; Pane et al., 2014; Supekar et al., 2015; Marouf and Abu-Naser, 2019). Many of these cognitive tutors focused on teaching mathematics and Algebra I, following the legacy of the original cognitive tutor (Koedinger et al., 1997; Ritter et al., 2007; Pane et al., 2014; Supekar et al., 2015), while some focused on programming (Marouf and Abu-Naser, 2019). One cognitive tutor teaching Algebra I improved the median student's performance by almost eight percentile points and had a statistically significant effect on high school learners' performance levels (Pane et al., 2014). Cognitive tutors also outperformed human-taught classes by 15% on standardized tests when tutoring 9th grade algebra during a real school year in three Pittsburgh high schools (Koedinger et al., 1997). Even more so, cognitive tutors generally significantly outscored peers on standardized tests in Algebra I (Ritter et al., 2007). In yet another example, one cognitive tutor helped 3rd grade students with basic mathematics, finding a decreased level of math anxiety after 8 weeks of cognitive tutoring (Supekar et al., 2015). The literature clearly shows the success of cognitive tutors within mathematics, and these findings support why cognitive tutoring in Algebra I has significant enough effects to meet the highest standards of the National Research Council (Council et al., 2003).

Cognitive tutors also have found success in teaching computer science (Marouf and Abu-Naser, 2019; Mousavinasab et al., 2021). Multiple findings from utilizing cognitive

tutors in a domain like computer science are promising: first, students using the tutor felt that the cognitive tutor was very interesting, easy to use, and useful (Marouf and Abu-Naser, 2019). Second, the professors utilizing the tutor to supplement their classes felt that the system explained material well and could expand to other subject domains (Marouf and Abu-Naser, 2019). Cognitive tutors focused on computer science span a large section of the literature: of 53 chosen papers for a meta-anlysis on cognitive tutor effectiveness, 37.73% of the tutors focused on general computer science, not including other categories such as mathematics, health/medical tutors, AI, and physics (Mousavinasab et al., 2021). Furthermore, the overall meta-analysis found that cognitive tutors "deliver adaptive guidance and instruction, evaluate learners, define and update the learner's model, and classify or cluster learners" very successfully (Mousavinasab et al., 2021). Cognitive tutors thus have found success in both the fields of mathematics and computer science.

While cognitive tutoring has found much success in mathematics and computer science, it has failed to greatly expand beyond STEM fields (Mousavinasab et al., 2021). Cognitive tutors have been built to mainly teach the fields of health, computer science, and mathematics with over 67% of cognitive tutors focusing on one of these three subjects; physics and AI are the next most popular subjects with about 4-6% of the cognitive tutor population (Mousavinasab et al., 2021). The one exception to the trend of STEM-focused cognitive tutors is language: 7.54% of the studied tutors focused on teaching some sort of language, like English (Mousavinasab et al., 2021). However, the literature fails to mention fields such as music: in fact, I am unable to find a cognitive tutor anywhere in the literature that specializes in teaching some sort of music education.

2.2 Cognitive Tutors and Music

Though cognitive tutors exist primarily in just STEM fields, intelligent tutoring systems could be extremely useful in the field of music. While cognitive tutors have yet to be created in the field of music, some tutors have been successful in the field of language

learning–a field with similar cognitive processes to music learning (C.-M. Chen and Li, 2010; Gordon et al., 2015; Vinchurkar and Sasikumar, 2015). One cognitive tutor successfully taught the English vocabulary through being "context-aware": keeping track of information such as total student learning time and the student's current vocabulary abilities in English (C.-M. Chen and Li, 2010). This cognitive tutor was shown to be more effective than basic tutoring applications that are not "context-aware" (C.-M. Chen and Li, 2010). Another tutor helped teach English grammar to 12-16 year old students by listening to a student's sentence in English, and providing proper feedback when grammar is incorrect, improving mean scores on English grammar tests by 25% as compared to a traditional method of learning grammar (Vinchurkar and Sasikumar, 2015). Importantly, rhythm and grammar utilize shared cognitive mechanisms and neural resources (Gordon et al., 2015). Since cognitive tutoring has shown success in teaching language, cognitive tutoring could be successful in teaching music. Therefore, teaching rhythm holds promise in cognitive tutoring, as teaching rhythm with a cognitive tutor could provide similar results to teaching language with a cognitive tutor. This study aims to contribute to this literature by testing the effectiveness of teaching rhythm by using a cognitive tutor.

Additionally, cognitive tutors could also be useful in music education by decreasing levels of music anxiety in students. Cognitive tutors have already aided in decreasing math anxiety in students (Supekar et al., 2015); the field of music also causes anxiety within students (Egilmez, 2012; Wristen, 2013; Patston and Osborne, 2016). Students learning music experience high amounts of music performance anxiety as children between ages 10 to 17 and as adults (Patston and Osborne, 2016). The intensity of anxiety also seems to increase as age increases (Patston and Osborne, 2016). During university, a "considerable number of [music] students reported symptoms indicative of anxiety or depression" (Wristen, 2013), and during musical examinations felt heightened anxiety as well (Egilmez, 2012). Additionally, many students feel that music is a talent-based skill that one has at birth (Woody, 2020) which could lead to heightened anxiety. Because music learning can cause extreme anxiety in

many ages, and cognitive tutoring has decreased other forms of academic anxiety, cognitive tutors may be poised to help decrease music anxiety.

2.3 Biofeedback and Emotion

In human tutoring sessions, a student's perceived affect is often useful to the tutor, allowing them to adapt the rest of the session by including how the student feels (Alexander et al., 2008; Lehman et al., 2008). This process can be referred to as biofeedback: a system that utilizes an individual's physiological processes to further understand, gain awareness of and change their current state (Gilbert and Moss, 2003; Schwartz and Andrasik, 2017). Biofeedback is most often used in medicine and therapy as a way for a patient to gain awareness of their physiological processes (Gilbert and Moss, 2003; Schwartz and Andrasik, 2017). However, biofeedback could be utilized effectively by a cognitive tutor. For instance, biofeedback is currently used by human tutors to investigate and understand a student's emotions (Alexander et al., 2008; Lehman et al., 2008). Human tutors will notice a student's emotions and whether or not a student seems interested in the material, as well as whether or not the student seems to understand the material; from this emotion, a tutor can effectively change the course of the session to better fit the particular student's needs (Alexander et al., 2008). Obviously, this process is done naturally in one-on-one tutoring without technology to detect biofeedback, but cognitive tutors could utilize measurements from technological tools such as electroencephalogram (EEG) readings, heart rate measures, and webcam-detected affect to alter the course of a tutoring session.

In creating a cognitive tutor utilizing biofeedback, many technologies could be used, such as EEG, heart rate measures, webcams, and more (Dimberg, 1982; Paranjape et al., 2001; Bradley et al., 2008; Magdin et al., 2016; Rathod et al., 2016; Madan et al., 2018). EEG has been used for many years to detect brain activity through electric signals (Lindsley, 1952) and has shown to be very effective in assessing anxiety and emotion (Thakor and Tong, 2004; Liu and Sourina, 2013). EEG has also been utilized in advertising to sense a viewer's

emotions and adapt an advertising strategy based on the emotions recognized in real-time (Liu et al., 2013). Additionally, EEG has been used to create a music-therapy web-player that plays music based on the user's emotions, sensed by EEG in real-time (Sourina and Liu, 2013). In sum, EEG is a powerful tool that can accurately detect and utilize emotions within web-based activities such as a cognitive tutor.

Heart rate has also been used to detect biofeedback effectively (Bradley et al., 2008; Vora et al., 2020). Heart rate has been shown to effectively measure emotional arousal and autonomic activation, as compared to other biofeedback measures such as skin conductance and pupillary changes (Bradley et al., 2008). Additionally, heart rate has been proposed as a useful measurement to non-invasively determine a user's attentive state in real-time (Vora et al., 2020). Furthermore, heart rate was suggested as a useful tool for a real-time cognitive tutor to investigate a user's attentive state (Vora et al., 2020); this, along with the fact that heart rate can measure emotional arousal (Bradley et al., 2008) shows proof of concept for detecting emotion in real-time for a cognitive tutor.

Another useful technology to measure biofeedback is the webcam, utilizing the human face to sense facial affect and emotion (Ekman et al., 2002; Magdin et al., 2016; Rathod et al., 2016; Madan et al., 2018; Vora et al., 2020). Using the human face as an indicator of emotion has long been an informal and common process; however, quantifying facial movement into emotion became possible with the use of the Facial Action Coding System (FACS) (Ekman et al., 2002). This system breaks down facial movement and emotion into many different "Action Units", or movements of the face, such as raising the outer brow, wrinkling the nose, or pulling in the corner of one's lip (Ekman et al., 2002). These units are then analyzed and grouped into 6 primary emotions: joy, sadness, fear, disgust, surprise, and anger, along with a catch-all "normal" or neutral emotion (Ekman et al., 2002). The FACS framework is used by many researchers to investigate human emotion through facial movement and a webcam. Emotion has been evaluated through a webcam within e-learning, showing an

overall accuracy of recognizing emotions to be 78% (Magdin et al., 2016). However, that number has increased with the usage of new technology, growing to 87% average accuracy for some emotions (Rathod et al., 2016). The webcam has also been compared to heart rate and has been shown to effectively detect psychophysiological changes such as emotion at the same accuracy as heart rate (Madan et al., 2018). Webcams have also been proposed as a viable method of collecting a user's biofeedback in cognitive tutors (Vora et al., 2020). Thus, the webcam is another useful method of recognizing and understanding biofeedback.

Because biofeedback and utilizing emotion can be used effectively in human tutoring to adapt a tutoring approach, if a cognitive tutor is to attempt to replicate human tutoring effectively, it should include emotion recognition as well. To do this, a cognitive tutor should utilize biofeedback signals from the student; I propose the most effective way to do so is with the webcam. The webcam has been shown to collect biofeedback signals effectively (Magdin et al., 2016; Rathod et al., 2016; Madan et al., 2018). Furthermore, it seems webcams are much more common, lightweight, and inexpensive than EEG or heart rate sensors in the common classroom. Thus, the webcam is a useful technology that allows cognitive tutors to sense and react to biofeedback.

2.4 Affect-Aware Learning Technologies (AALTs)

As seen, human tutors adapt to the student's emotion within a tutoring session: cognitive tutors may be able to adapt to the student's displayed affect as well. Cognitive tutors combined with an emotional, affective aspect that originates from biofeedback are called Affective Tutoring Systems (ATS), or more broadly, Affect-Aware Learning Technologies (AALT) (Zakharov, 2007; Calvo et al., 2015). If human tutors utilize student emotion to successfully adapt and improve their teaching methodology in real-time, then for a cognitive tutor to truly be as effective as a human tutor, it should theoretically become an affective tutoring system. While affective tutoring systems are much less common in the literature than general cognitive tutors, there are some examples of successful affect-aware learning

technologies and affective tutoring systems (Alexander et al., 2006; D'Mello et al., 2007; Zakharov, 2007; Cabada et al., 2012; Faghihi et al., 2013). Furthermore, there are some investigations into the promise and the future of affect-aware cognitive tutors (Woolf et al., 2009; Calvo and D'Mello, 2012; Landowska, 2014; Calvo et al., 2015; Kołakowska et al., 2015; Grawemeyer et al., 2017). Through these few examples and discussions, affect-aware cognitive tutors are a novel and somewhat unexplored territory.

Affect-Aware Learning Technologies, or "educational technologies that compute affect in addition to cognition", originate from the general domain of affective computing (Calvo et al., 2015). There are two main types of AALT's: reactive systems that analyze and respond to affect from the student after it occurs and proactive systems that attempt to induce a specific affect upon the student (Calvo et al., 2015). In a proactive system, the tutor simply attempts to respond to the student in a certain manner to induce emotion, perhaps by being stern, supportive, or energetic in nature (Alexander et al., 2006; Calvo et al., 2015). In a reactive system, the student's affect can be discovered through some sort of technology, like biofeedback, or it can be inferred based on "affective clues" (Woolf et al., 2009). Affective Tutoring Systems are AALTs with a reactive system, as they must recognize a learner's affective state and respond to it (Kołakowska et al., 2015). While some ATSs have been created, the field of affective cognitive tutoring is far from discovered, especially when considering e-learning (Landowska, 2014).

Of the AALTs that exist currently, there are some proactive and some reactive tutors, many of which have been shown to be successful (Alexander et al., 2006; D'Mello et al., 2007; Zakharov, 2007; Cabada et al., 2012; Faghihi et al., 2013). One extreme example of a proactive AALT is CELTS, a cognitive tutor agent built to assist with operating the Canadarm2 robotic arm on the International Space Station (Faghihi et al., 2013). The cognitive tutor would send messages to the user in case the tutor became "worried" about certain events occurring, like the arm being too close to a dangerous object; overall learning

for the user was found to be successful (Faghihi et al., 2013). While proactive AALTs have merit, reactive AALTs seem best used for situations where the student displays strong affect; music and math, due to the strong anxiety associated with each, could be successful domains for reactive AALTs.

While no affective tutoring systems exist within the music field, some affective tutoring systems in other fields have shown great promise; four are relevant to this study (Alexander et al., 2006; D'Mello et al., 2007; Zakharov, 2007; Cabada et al., 2012;). One tutor called AutoTutor helped students learn physics by conversing with the student aloud in their natural language, tracking the student's facial affect, body posture, and dialogue features; the affective states of confusion and delight were tracked effectively and accurately by facial expression, a nod to the effectiveness of webcam-sensed biofeedback (D'Mello et al., 2007). Another tutor, named FERMAT, taught primary-school math while obtaining emotional features "by sensors that are monitoring the user's emotions" (Cabada et al., 2012). While the study is not specific as to how the user's emotions are detected, FERMAT was successful in teaching (Cabada et al., 2012). Easy with Eve is another cognitive tutor that senses emotion in real-time through facial expressions to teach mathematics; it analyzes a student's video feed to determine the student's current affect, then responds appropriately depending on the affect detected (Alexander et al., 2006). The final affective tutoring system utilizes a live video stream to detect how the student is feeling, and if the tutor notices negative thoughts, responds with messaging to try and dispel those thoughts (Zakharov, 2007). 3 of these 4 tutors utilize the Ekman emotional framework (Alexander et al., 2006; D'Mello et al., 2007; Zakharov, 2007). Through these examples, affective tutoring systems have shown success, especially in detecting facial expressions in real-time by utilizing a webcam to separate emotions into 6 primary emotions.

While AALTs have been created and have displayed success, few are affective tutoring systems that react to the student's emotion like in human tutoring, and none are utilized in

the promising field of music education. For these reasons, this study aims to build a webcam-based affective tutoring system that teaches rhythm, utilizing the emotion displayed through facial expressions to aid students in better understanding rhythm.

2.5 The Current Study: Webcam-based Affective Cognitive Tutor for Rhythm

From reviewing the literature, there are many areas where cognitive tutors can be further researched. In this study, I create an affective cognitive tutor aimed at teaching rhythm to students, collecting affect with a webcam by evaluating facial expressions in real-time and separating them into Ekman's 6 emotions (Ekman et al., 2002). Then, using the cognitive tutor, I test whether or not utilizing affect to adapt a tutoring plan for teaching rhythm expedites and increases the efficacy of the teaching itself. The current study aligns with and expands upon the current literature in four ways: first, the study builds upon the previous successes of cognitive tutors; second, the study produces a system that accurately collects affect through facial expressions and the use of a webcam; third, the study utilizes the reactive system of Affect-Aware Learning Technology to create an application that responds to user affect; and fourth, the current study advances the use of a cognitive tutor to the unexplored field of music, relying on the fact that rhythm and grammar are similar cognitively.

The study at hand contributes to the literature in multiple ways while also following the successes of the current literature. Cognitive tutors have been shown to be very successful in teaching math, computer science, and other STEM subjects (Council et al., 2003; VanLehn, 2011; Ma et al., 2014). However, cognitive tutors have yet to explore the field of music. Rhythm proves to possibly be a great subject for cognitive tutors, since cognitive tutors have shown success in teaching language (C.-M. Chen and Li, 2010) and language and rhythm have similar cognitive and neurological processes (Gordon et al., 2015). Thus, this study first adds to the literature by creating a cognitive tutor that explores music education as a

field of study for cognitive tutors.

Additionally, affective tutoring systems have shown much success in teaching, but again, only within the STEM fields (Alexander et al., 2006; D'Mello et al., 2007; Cabada et al., 2012; Faghihi et al., 2013; Landowska, 2014). Affective tutoring systems have yet to enter the sphere of music, somewhere with much pronounced affect such as anxiety (Egilmez, 2012; Wristen, 2013; Patston and Osborne, 2016). Therefore, this study also contributes to the literature by adding music to the list of educational fields that affective tutoring systems have explored. Finally, affect-aware learning technologies have successfully used webcams to collect and understand a student's affect, and the AALTs have then separated that affect into 6 emotions using Ekman's emotional theory (Ekman et al., 2002; Alexander et al., 2006; D'Mello et al., 2007; Zakharov, 2007). Thus, the cognitive tutor for this study both adds to the literature in multiple ways, as well as follows the past successes and rules of previous cognitive and affective cognitive tutor studies.

3 Methodology

3.1 Overview

This study was funded by a Neukom Scholar Grant from the Neukom Institute of Technology at Dartmouth College, as well as the Dartmouth Cognitive Neuroscience of Learning Laboratory. Additionally, this study was approved by the Institutional Board of Dartmouth College. I employ a randomized controlled trial experiment with two experimental conditions to determine the relationship between biofeedback and music education performance, as well as the viability of biofeedback in cognitive tutoring within the music education sphere. For this experiment, all participants first took a survey inquiring about any previous musical knowledge, such as being able to read sheet music. Then, all participants learned drumming notation from an affect-aware cognitive tutor, learning various rhythms. Participants were asked to tap out rhythms using three keys on their keyboard, each relating to a different

drum in their "drum kit". The drumming process is expanded on below.

After each lesson, all participants were scored on three different factors in their performance: the accuracy of the participant's hits, the timing error of the participant's hits, and the participant's facial affect that they display during the lesson. These 3 scores—accuracy, timing, and affect—were then used to determine whether or not a participant continues to the next lesson or repeats the current lesson. The calculation of the final determination of whether to advance or not separates the control condition and two experimental conditions.

Control Condition - No Affect Score. In the control condition, only the accuracy and timing scores are used to decide whether the user advances or not by taking the average of the binned accuracy and timing scores.

Experimental Condition - Intelligent Affect Score. In the experimental condition, the affect score was used to determine if the participant should continue to the next lesson in addition to the accuracy and timing scores. The affect score was calculated by computing a performance score based on a mixed-effect model run on the results of the control experiment. This mixed-effect model uses fear and surprise to predict emotion.

3.2 Participants

For this experiment, 59 participants were tested, with 20 in the experimental condition and 39 in the control condition. Participants were recruited utilizing Amazon Mechanical Turk, a program allowing users to take the experiment offline without a moderator present and receiving payment at the end (Paolacci et al., 2010), Crowston, 2012). Though doubts exist about the reliability of participants recruited through Amazon Mechanical Turk due to a lack of monitoring participants (Crowston, 2012), this study leaves little room for the participant to become disengaged. Each activity requires the user to try and succeed in order to continue, rather than allowing mindless engagement like a simple question survey. Furthermore, participants that were recruited needed at least 99% of previous assignments

accepted, as well as at least 500 previously completed assignments. Additionally, each participant was required to have a webcam and microphone that could be enabled during the study. Each participant was paid eight dollars upon successful completion of the experiment, regardless of their success.

In addition to participants recruited on Amazon Mechanical Turk, participants were also recruited in-person at Dartmouth College through Sona Systems, an online research participation software that allows students and community members to participate in a study for either monetary compensation or, if applicable, some form of credit for a research component in some Psychology department courses, such as Introduction to Psychology ("Research and participant management made easy in the cloud", n.d.). In-person testing allows for monitoring participants to further ensure their focus and drive to succeed and allows for a comparison between participants recruited from Amazon Mechanical Turk and participants that took the experiment in-person. Each participant again was paid ten dollars upon successful completion of the study or 1 T-point, the units of credit for the Psychology and Education Departments at Dartmouth College.

3.3 Pre-Survey and Post-Survey

Each participant was administered a pre-survey immediately prior to beginning the drumming notation lessons section of the study. The pre-survey collected participant's demographic data and asked for consent to both participate in the study and have video and audio recorded during the study. Importantly, the pre-survey asked each participant about their previous musical ability, whether or not the participant knew basic sheet music, and what musical abilities they had practiced in the past and for how long (i.e. drums for two years, singing for one year, etc.). The data from the pre-survey is reported in Appendix A.

Each participant was also administered a post-survey immediately after finishing the drumming section of the study. Each participant noted any technical issues they had with

the study and whether all sounds of the application played with correct timing. Additionally, the post-survey asked how comfortable the participant felt with basic drumming notation having completed the study, asked in detail what the participant learned during the study, and asked whether or not the participant would participant would enter a study like this one again. This data is reported in Appendix A.

3.4 Tutor Format and Lesson Format

Regardless of treatment condition, each participant utilized an affect-aware cognitive tutor built specifically for this study. The affect-aware cognitive tutor was built using React, a web development framework allowing for users to take the study directly in their browser on their personal computer ("React – a JavaScript library for building user interfaces", n.d.). All participants used the cognitive tutor in Google Chrome to standardize browser conditions for the study.

The cognitive tutor first began with an introduction screen, instructing users to turn on their webcams and microphones to ensure proper affect recording during each lesson. Once the user read the instructions, they could click on the start button, where two events occurred: first, a 35-minute timer began. When this timer ended, the study immediately ended, ensuring that each user had the same maximum amount of time to spend taking the lessons. Second, the participant was shown the first lesson of the experiment.

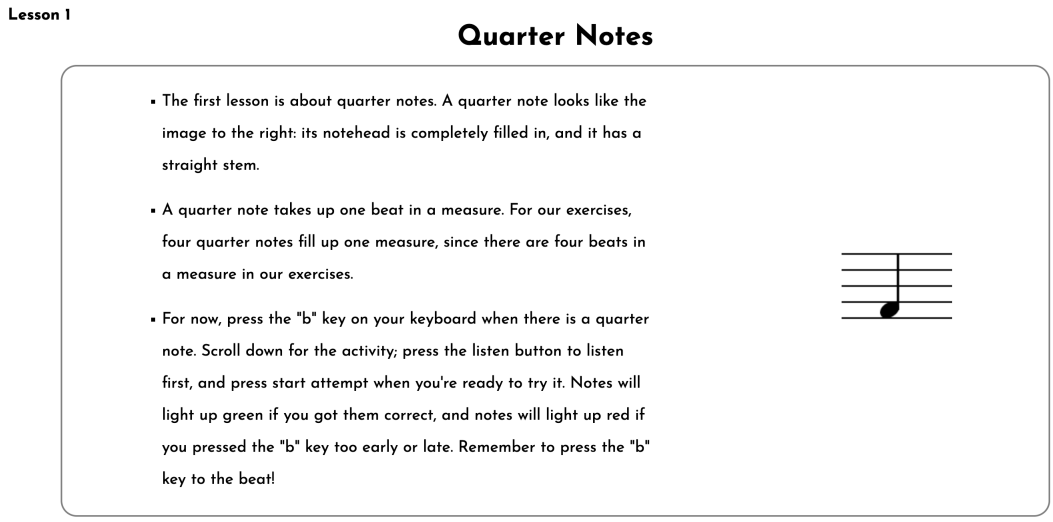

Figure 1: Instructions for the first lesson.

In the cognitive tutor, there are a total of 18 lessons. Each lesson is pre-made to where participants would take the same 18 lessons in the same order after completing the previous lesson successfully. The list of lessons is described in A3. Participants learned quarter notes, half notes, whole notes, and eighth notes. Each lesson began with a textbox of notes about the information being introduced; an example of the first lesson instructions are shown in Figure 1. During each lesson, participants were asked to play back 12 measures of a displayed rhythm utilizing the types of notes they have learned so far. The tempo at which the measures were played varied between lessons, shown in Table A3. Participants could first press a "Listen" button to listen to the measures played for them, and when the participant was ready to attempt the lesson, the participant could press "Start Attempt" to play the rhythms themselves. The lesson screen is shown in Figure 2.

For the participant to attempt the rhythms for a lesson, after pressing "Start Attempt", the participant had to press the correct button on the keyboard corresponding to a specific instrument. When a keyboard button was pressed, a drum sound corresponding to the button played. There were three buttons total used in the experiment: the "B" button, which corresponded to a bass drum sound; the "F" button, which corresponded to a floor-tom drum

sound; and the "J" button, which corresponded to a hi-hat drum sound. Participants began by only pressing the "B" button, and as lessons progressed, participants were introduced to notes that corresponded to the "F" button and "J" button, as shown in Table A3.

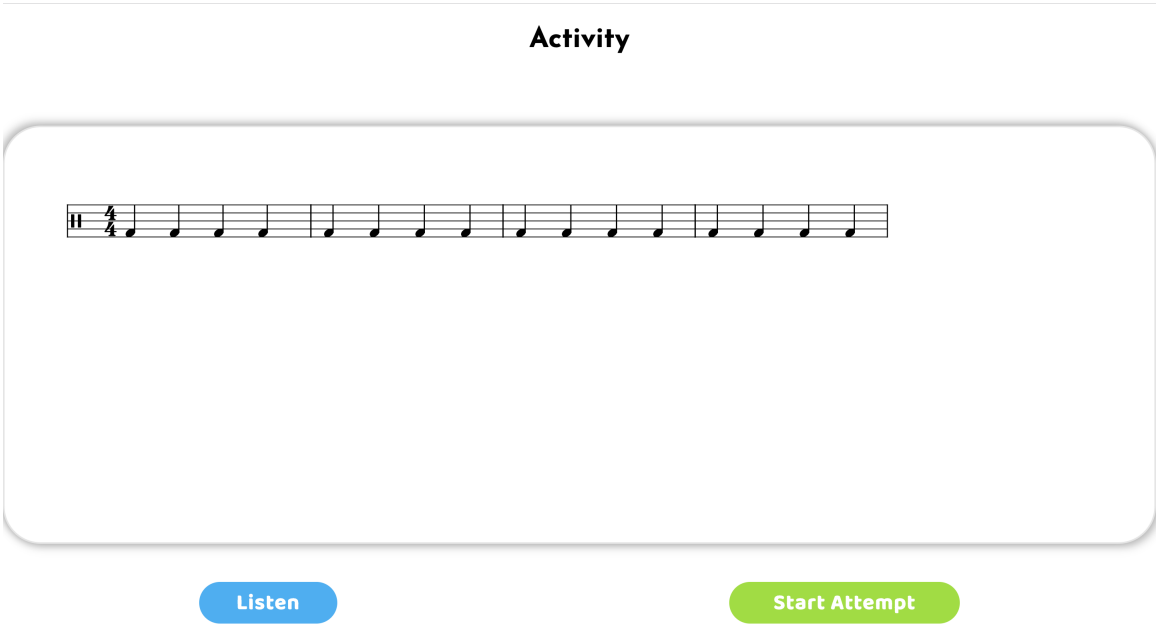

Figure 2: The starting interface for a lesson's activity.

When a keyboard button was pressed, feedback was given to the participant, and the participant's timing and accuracy were recorded for that specific note. Before the participant pressed a button, if a specific note was within the range of being pressed, the note would light up with the color blue. When the participant pressed a button, the note would change colors depending on how "well" the user performed. If the participant pressed the wrong button (the user should have pressed the "B" button, but pressed the "J" button instead), the note would light up as red. Likewise, if the participant pressed the button more than 350 milliseconds early or 350 milliseconds late, the note would light up as red. Finally, if the participant pressed the correct button within 350 milliseconds of when the button should be pressed based off of the correct rhythm, the note would light up as green. This process continued for each note in the 12-measure lesson. For each press, the number of milliseconds away the user was from the correct time was recorded, as well as whether or not the note

became green or red, indicating if the participant answered the note correctly or incorrectly. An example of the notes changing colors is shown in Figure 3.

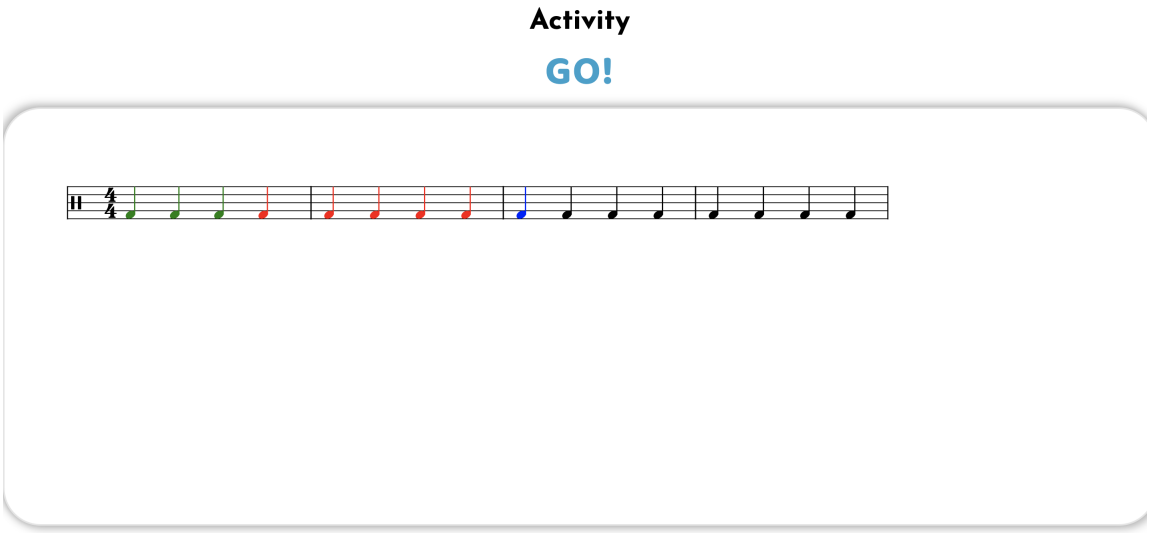

Figure 3: A lesson activity in progress.

While the participant attempted the lesson, two other events occurred: another timer began to track how long the participant spent on that single attempt, and the participant's webcam began recording a video stream of the participant's face. The timer compiled the amount of time spent on all of the lesson attempts, showing the total amount of minutes and seconds the participant spent attempting rhythms (since the timer only began during each attempt, the total time amount excluded time spent listening to rhythms without attempting and time spent reading instructions). At the end of the attempt, the participant's video stream was then analyzed for facial affect. The participant then clicked on the next button to see the results of their performance. At this point, the cognitive tutor would consider this an "attempt" for that lesson, and the number of attempts for each lesson was likewise recorded for each participant. After each attempt, the participant was then notified as to whether or not they have advanced to the next lesson. If the participant was in the control condition, if the average of their overall computed accuracy and timing from the most recent lesson attempt exceeded a specified boundary, then the participant continue to the next

lesson; otherwise, the participant repeated the current lesson. If the participant was in the experimental condition, if the average of their overall computed accuracy, timing, and facial affect from the most recent lesson attempt exceeded a specified boundary, then the participant continued to the next lesson; otherwise, the participant repeated the current lesson. The computation of the overall accuracy, timing, and facial affect, as well as the collection of the facial affect score, is now explained.

3.5 Measures

There are a couple of main measures of data for this study. First, for calculating overall performance, there are three measures: accuracy, timing, and highest lesson. These three measures are then combined together to create an overall performance measure for the lesson. The calculation and meaning of each of the three performance measures is described below.

Accuracy. Accuracy is measured for each lesson attempt by counting the percentage of notes the participant responds to correctly, both in time and with the correct instrument (key). This is done by counting the number of "green" versus "red" notes in the attempt accuracy array, with "red" constituting an incorrect button press or pressing the correct button 350 milliseconds early or late for any given note. The mean attempt's accuracy is then taken and kept as the participant's overall accuracy percent. Thus, accuracy is a percentage from 0-100% for each participant.

Timing. Timing is measured for each lesson attempt by calculating the magnitude of milliseconds the participant's button press on any given note is away from the correct time to react to the note. If a user fails to react to any given note, that note's error is recorded as 500 milliseconds; oppositely, if the user reacts perfectly to any given note, that note's error is recorded as 0 milliseconds. The mean attempt's error is then taken and divided by 500 milliseconds to create a percentage error. We then subtract this number from 1 in order

to create the timing measure, putting it on the same valence scale as accuracy (a higher accuracy percent means that a participant is doing better; likewise, by subtracting from 1, a higher timing percent means that the participant is doing better). This measure is then kept as a participant's overall timing percent, on a scale of 0-100%.

Highest lesson. Finally, the highest lesson is measured for each participant. This measure simply records the highest lesson that the participant reaches, out of 18 possible lessons. The highest lesson is then divided by 18 in order to put this measure on a 0-100% scale, with 100% meaning that the participant reached every level possible. Thus, a higher percentage for highest lesson means that the participant has performed better.

Accuracy, timing, and highest lesson are the three performance measures tracked for each participant to determine a participant's overall success with learning throughout the study. Performance is then analyzed with regard to the participant's overall affect measures, the collection and meaning of which are described in the next section.

3.6 Affect Analysis and Collection

In order to create an affective cognitive tutor, affect is monitored and analyzed throughout each lesson. There are three important pieces of how this study analyzes affect: the collection method, the real-time summarizing of the affect into emotions, and the calculation of the affect percent measure for each participant.

3.6.1 Collection of Affect

In order to collect affect for this study, the cognitive tutor utilizes the user's webcam. The webcam will begin recording the user's face as soon as the user clicks to start an attempt for any given lesson, and the video recording will end when after the metronome clicking ends for the last measure in a lesson. The recording of the participant's face is then sent to be analyzed and stored in a private database hosted by Google Cloud Storage. Collecting

the user's affect through their facial expressions seen by a webcam allows the application to be a lightweight and real-time web application while still producing accurate results (Ekman et al., 2002; Magdin et al., 2016; Madan et al., 2018; Rathod et al., 2016; Vora et al., 2020).

3.6.2 Summarizing Affect into Emotions

Once the user's affect has been collected for a lesson attempt, it is then sent as an .mp4 file to a server built specifically for this study to analyze the user's displayed affect. The server is built in a Python-based framework called Flask (Van Rossum and Drake, 2009; Grinberg, 2018), allowing the server to communicate with web applications while using Python packages.

The server used a special package called Py-feat in order to analyze the webcam-monitored affect of each participant (Cheong et al., 2021). Py-feat is an open-source package available on GitHub written for and in Python3 (Cheong et al., 2021). The package allows for an .mp4 file to be analyzed in real-time and split into emotions based off of the specified models used for detecting faces, facial landmarks, action units, and emotions. For this study, the model used to detect faces was the FaceBoxes model (Guo et al., 2018; Guo et al., 2020), for landmark detection was the MobileNet model (Howard et al., 2017), for action units was a Random Forest based model (J. Chen et al., 2017), and for emotions was the Residual Masking Network model (Luan et al., 2020). The combination of these four models resulted in high accuracy as well as about a 20 second processing time for each attempt, which allowed the study to run fast enough in real-time while still maintaining accuracy. The Py-feat package will analyze every 98th frame of the input video to determine the participant's levels of displayed emotion during that frame, analyzing only every 98th frame in order to run fast enough for real-time purposes.

From each video, a dataframe of seven emotions is produced with the level of displayed emotion displayed in each frame analyzed included in the dataframe. There are seven emotions included in the output: happiness, sadness, anger, disgust, fear, surprise, and a

neutral emotion, all based off of Ekman's theory of emotions (Ekman et al., 2002). The level of displayed emotion is a percentage of that emotion in comparison to the other displayed emotions in that frame: for instance, if happiness receives a rating of 97%, there is a value of 3% left that is split between the other emotions in that frame. The mean value of each emotion is then calculated for each of the seven emotions during that attempt, and these values are used to calculate the affect percentage for that participant's lesson attempt.

3.6.3 Affect as a Measure

In addition to performance measures, affect is the final measure of importance to the overall analysis of the study. The mean percent for each emotion for each attempt is averaged across every attempt to calculate an overall percent affect score for each participant for each emotion. Thus, there are seven percents per participant: one for happiness, anger, sadness, disgust, fear, surprise, and neutral emotion. All seven of these percentages are then used to create mixed effect models to determine the effectiveness of utilizing affect to predict performance.

While the participant is taking each lesson, to determine whether or not the participant should continue with the next lesson or repeat the same lesson, an affect percent score is calculated and averaged with the accuracy and timing percent scores. If the participant is in the control group, then the affect score is calculated by taking the mean of the negative affect emotion scores: sadness, anger, disgust, and fear. If the participant is in the experimental group, however, the affect score is calculated by utilizing a model for predicting performance outlined in the first results section of this study.

4 Results

4.1 Overview

To answer the first research question of whether or not emotion is correlated with performance on rhythm activities, I utilize a correlation matrix, a principal component analysis, and mixed-effects models. First, I use Pearson's correlation coefficient to create a correlation matrix of the seven emotions—happiness, sadness, anger, disgust, fear, surprise, and neutral—to determine the correlation between each emotion. Then, I use a principal component analysis to reduce the dimensionality of each of the seven affect emotions and determine which emotions are key to explaining the data. Finally, I then create a mixed-effects model to predict overall performance with each emotion as its own fixed effect, including subject and the type of performance measure as random effects. I create a model of all 6 emotions together (excluding neutral), as well as a model of each of the 7 emotions individually, both on the raw emotion scores and the z-scored versions of the emotion scores.

To answer the second research question, I run multiple two-tailed t-tests to determine if there was a significant difference between the performance of the control group and experimental group. I also analyze the overall performance over time of each group as well as the performance of each group on the highest lesson each participant reached. I additionally investigate to see if the number of attempts per lesson changes by treatment group.

4.2 Correlation of Affect Variables with Performance

I first ran a cross-correlation matrix to calculate Pearson's correlation coefficient between all of the emotions of affect, displayed in Table 1. Unsurprisingly, many emotions were correlated with one another, and neutral emotion was correlated with every other emotion with the largest p-value being .01.

I also run a cross-correlation matrix to calculate Pearson's correlation coefficient be-

Table 1: Correlation of emotions.

	sadness	anger	fear	disgust	happiness	neutral
sadness						
anger	0.24****					
fear	-0.01	-0.23****				
disgust	0.03	0.29****	-0.11*			
happiness	-0.12*	-0.03	0.08	-0.05		
neutral	-0.59****	-0.31****	-0.61****	-0.16**	-0.20****	
surprise	-0.11*	-0.07	0.06	-0.09	0.19***	-0.28****

* $p < .05$, ** $p < .01$, *** $p < .001$ .001, **** $p < .0001$
Notes: Pearson's r coefficient is displayed as the correlation coefficient.

tween all three measures of performance: accuracy, timing, and highest lesson per participant. The results of this can be seen in Table 2. All three performance measures are strongly correlated, with accuracy and timing having a correlation of .96, accuracy and highest lesson having a correlation of .83, and timing and highest lesson having a correlation of .81. These all are significantly correlated at a level of $p < .0001$.

Table 2: Performance metrics correlation.

	Accuracy	Timing
Accuracy		
Timing	0.96****	
Highest Lesson	0.83****	0.81****

* $p < .05$, ** $p < .01$, *** $p < .001$ **** $p < .0001$
Notes: Pearson's r coefficient is displayed as the correlation coefficient.

I then ran a principal component analysis to understand the dimensions of each emotion among each other. According to the analysis, three clusters emerge: fear, happiness, and surprise as one cluster; anger, sadness, and disgust as another cluster; and neutral as a third cluster. The results of the PCA can be seen in Figure 4, Figure 5, and Figure 6, as well as in 3.

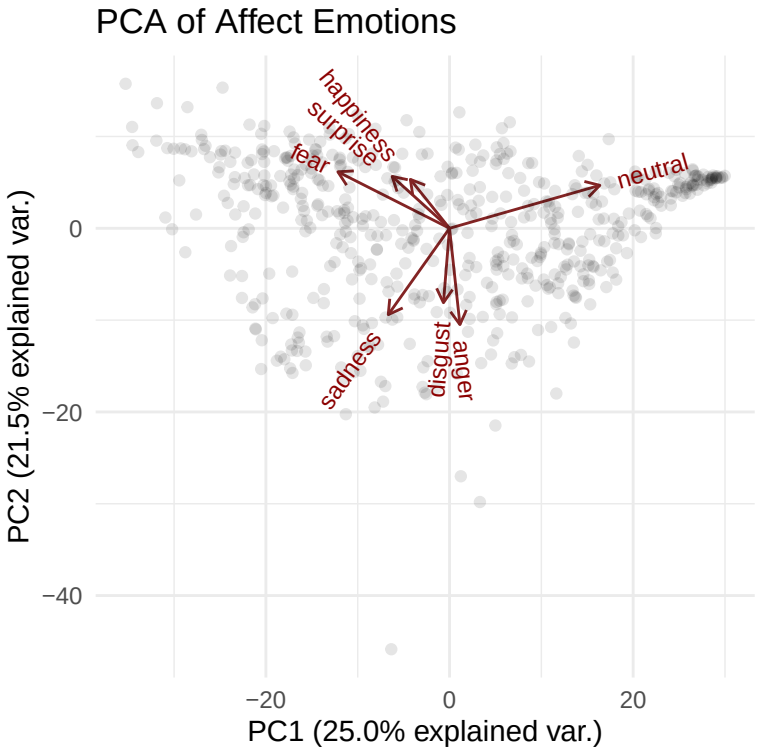

Figure 4: Principal Component Analysis of Seven Affect Emotions

Table 3: PCA of Emotions.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.3218	1.2259	1.0887	0.9826	0.9115	0.8766	0.0000
Proportion of Variance	0.2496	0.2147	0.1693	0.1379	0.1187	0.1098	0.0000
Cumulative Proportion	0.2496	0.4643	0.6336	0.7715	0.8902	1.0000	1.0000

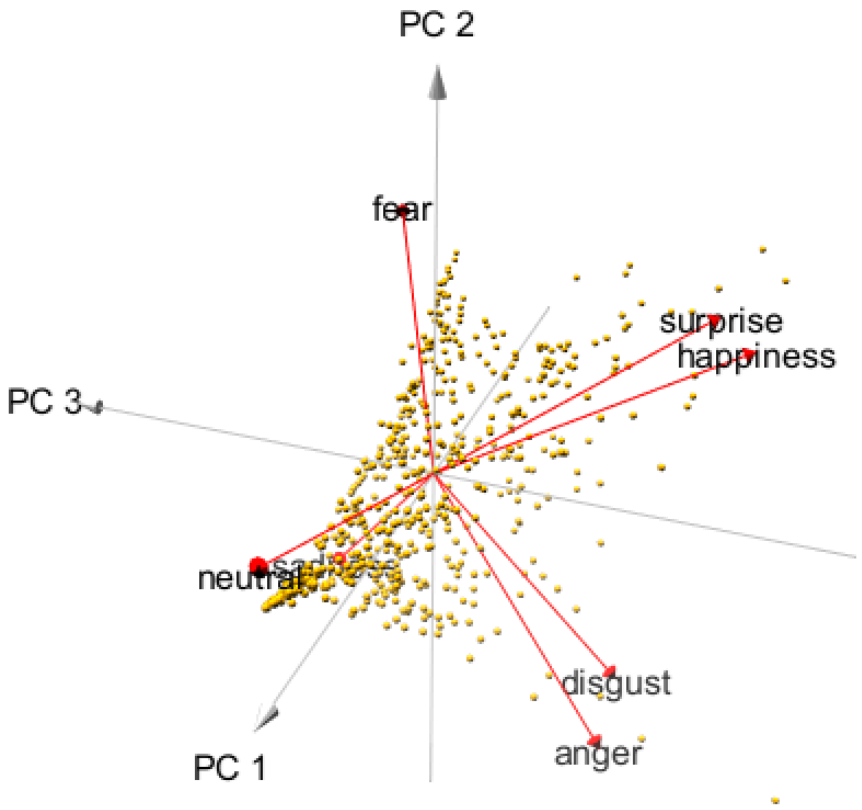

Figure 5: 3D Principal Component Analysis of Seven Affect Emotions

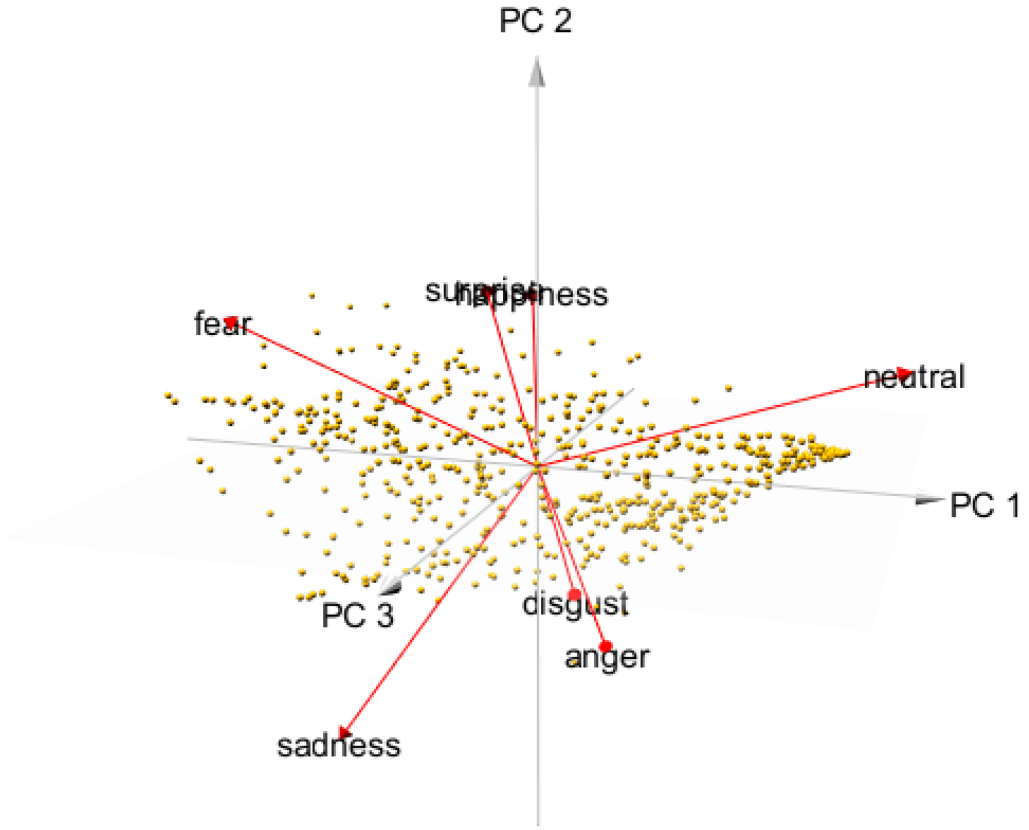

Figure 6: 3D Principal Component Analysis of Seven Affect Emotions

I then created a mixed-effects model for each emotion, using emotion to predict performance. Figures 7 through 13 show the mixed-effects models for each emotion and their effects on performance for the three measures of performance. Disgust and happiness have very limited ranges, and neutral seems to have a small effect on performance. Furthermore, happiness, surprise, and fear seem to have positive effects on performance. The betas for each individual mixed-effects model can be found in Table 4.

Table 4: Results of Individual Mixed-Effects Models.

	Estimate	Std. Error	t value
intercept	0.59	0.07	8.21
anger	-0.19	0.35	-0.54
intercept	0.60	0.08	7.90
sadness	-0.18	0.21	-0.85
intercept	0.59	0.07	9.07
disgust	-2.00	1.00	-2.00
intercept	0.53	0.07	7.54
fear	0.24	0.19	1.27
intercept	0.55	0.07	7.56
happiness	0.76	1.18	0.65
intercept	0.54	0.07	7.61
surprise	0.61	0.54	1.11
intercept	0.59	0.09	6.24
neutral	-0.03	0.14	-0.24

Notes: Each mixed-effects model is listed here, with the intercept for the model listed first and the emotion's intercept listed second. The standard errors and t-values for each model's intercepts are also listed.

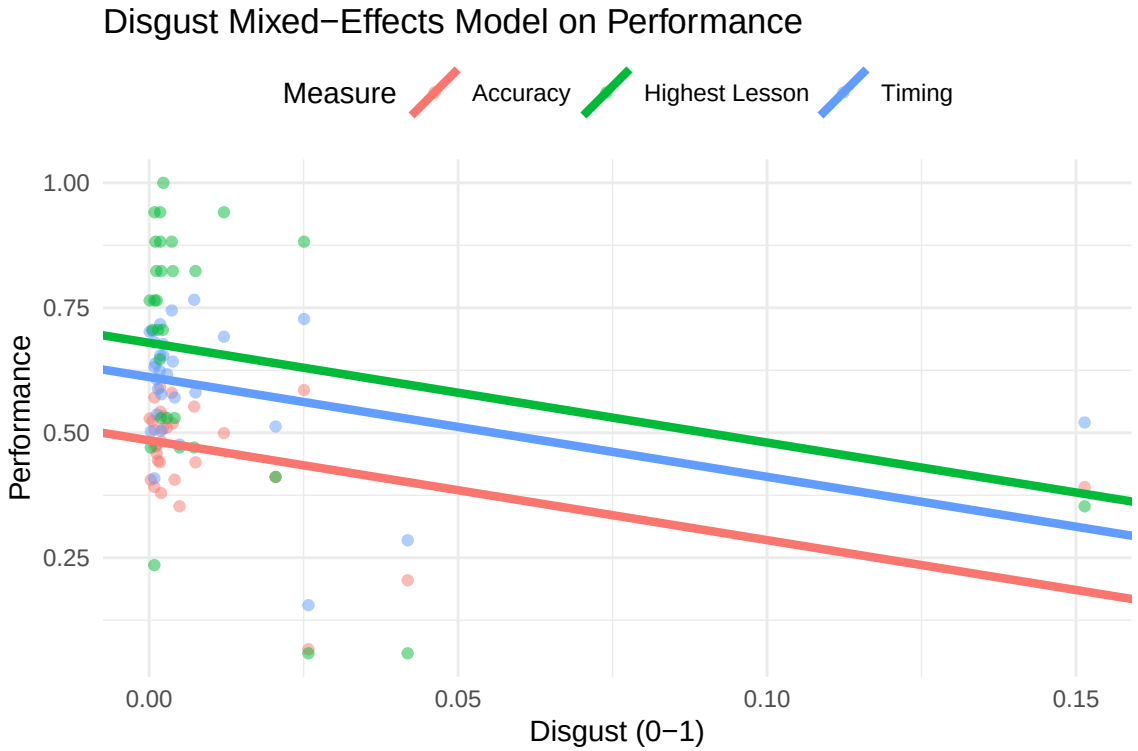

Figure 7: Disgust Mixed-Effects Model for Performance.

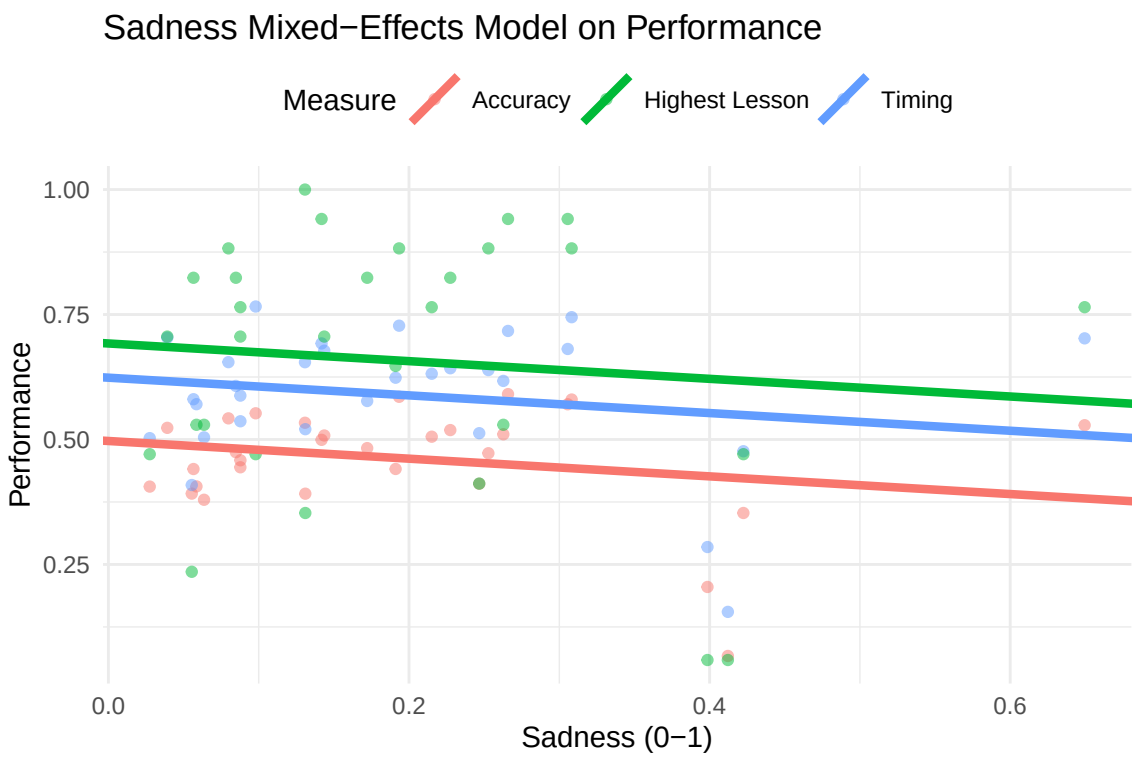

Figure 8: Sadness Mixed-Effects Model for Performance

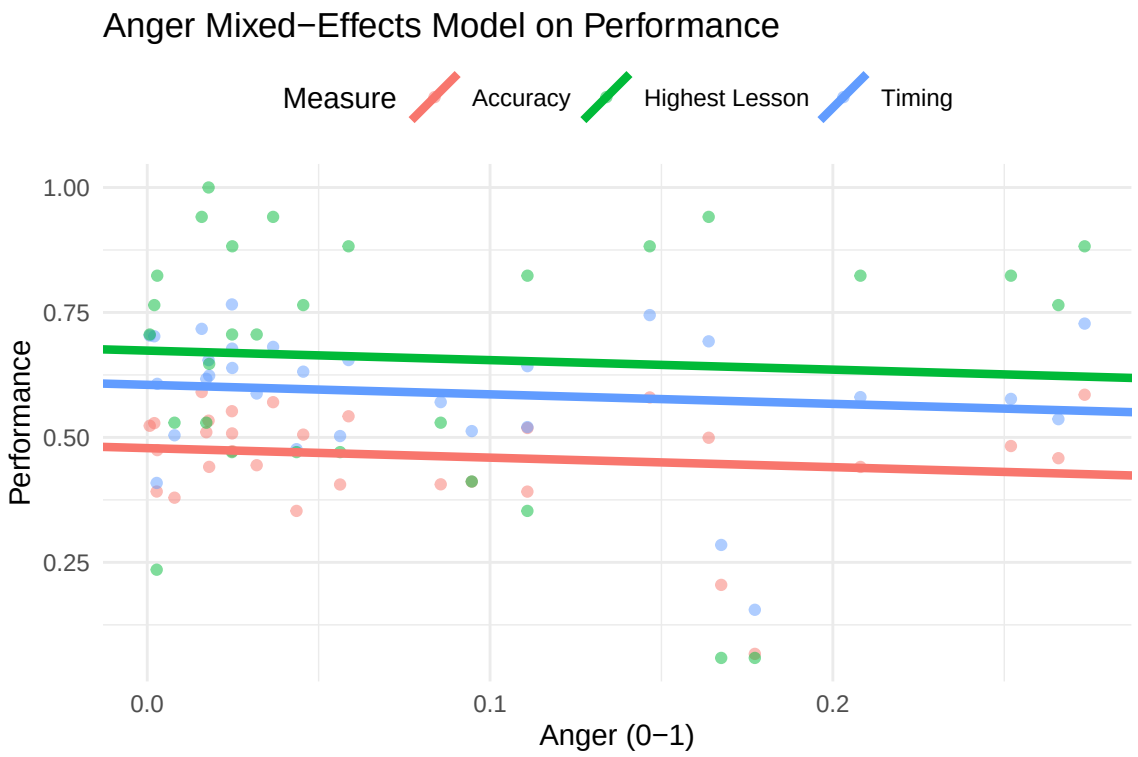

Figure 9: Anger Mixed-Effects Model for Performance.

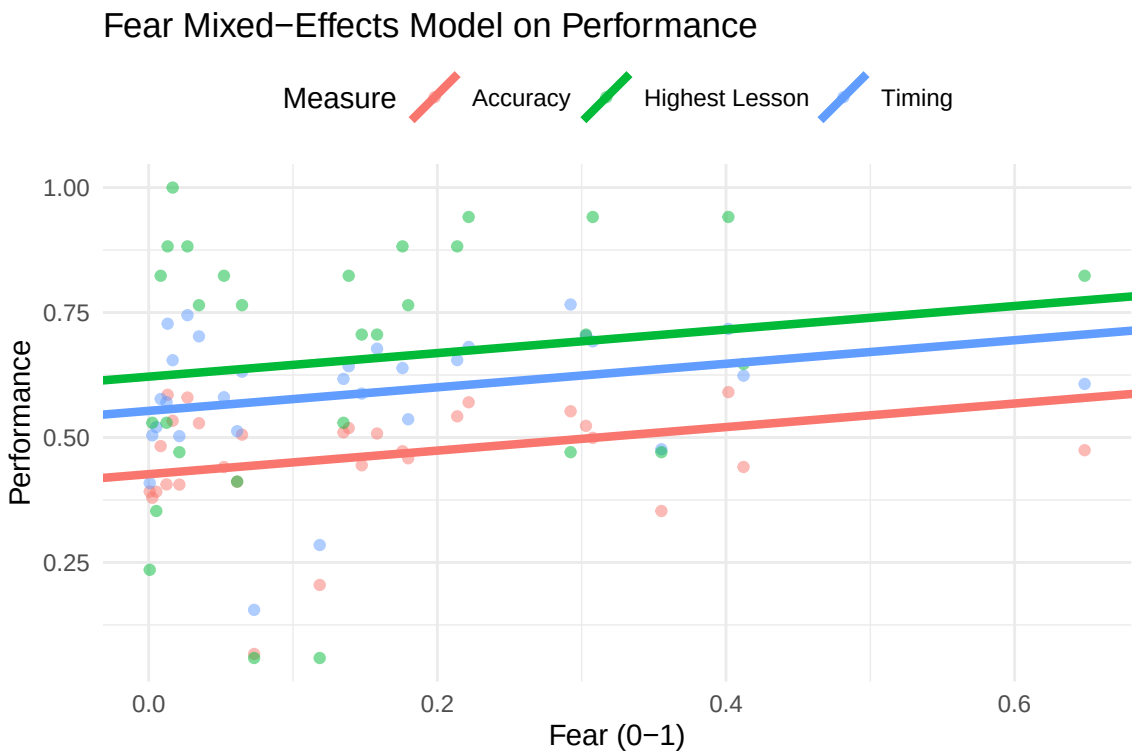

Figure 10: Fear Mixed-Effects Model for Performance.

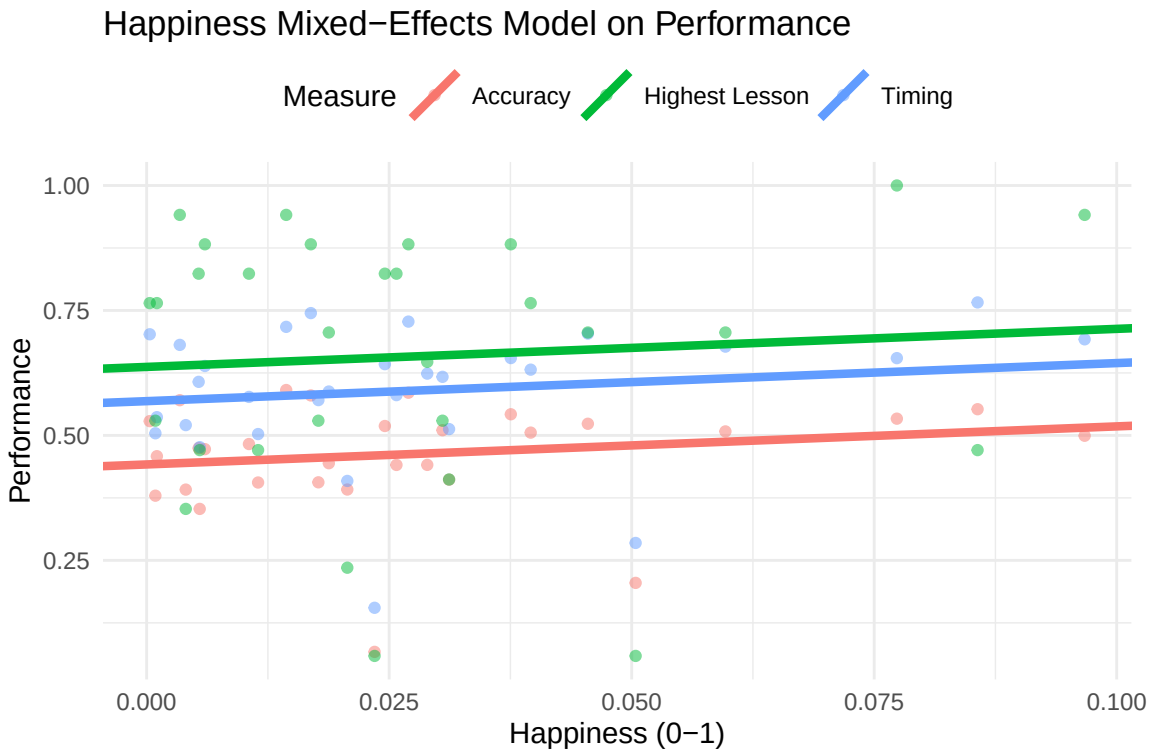

Figure 11: Happiness Mixed-Effects Model for Performance.

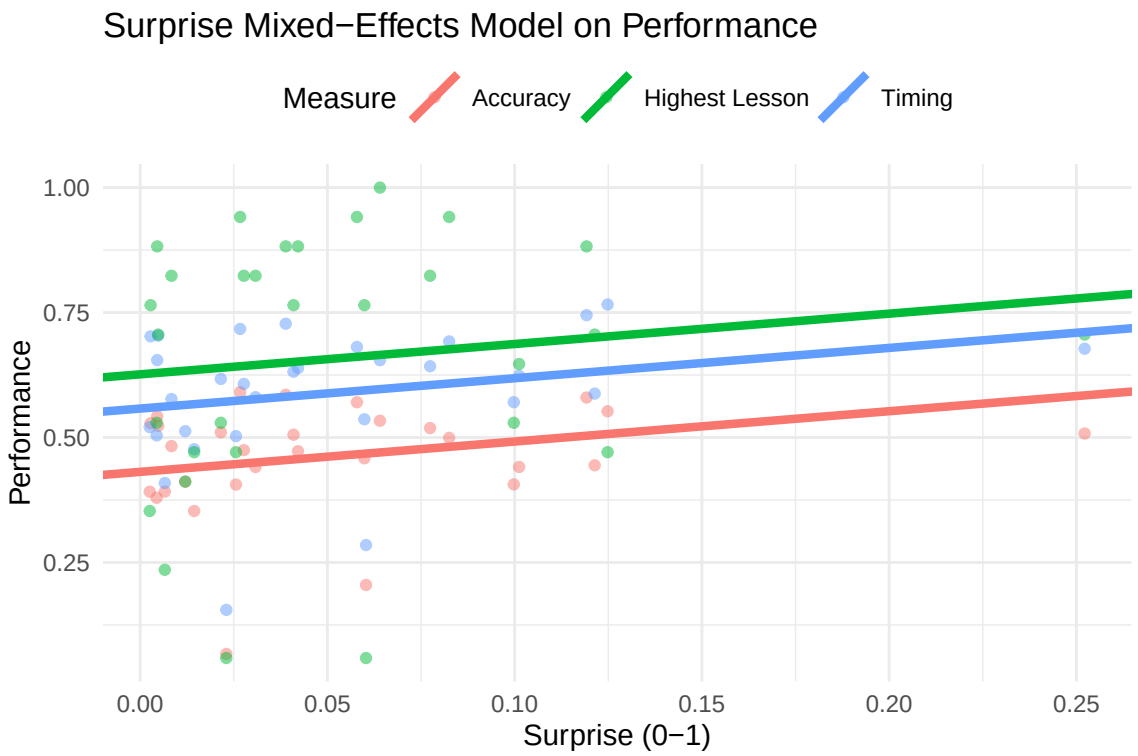

Figure 12: Surprise Mixed-Effects Model for Performance.

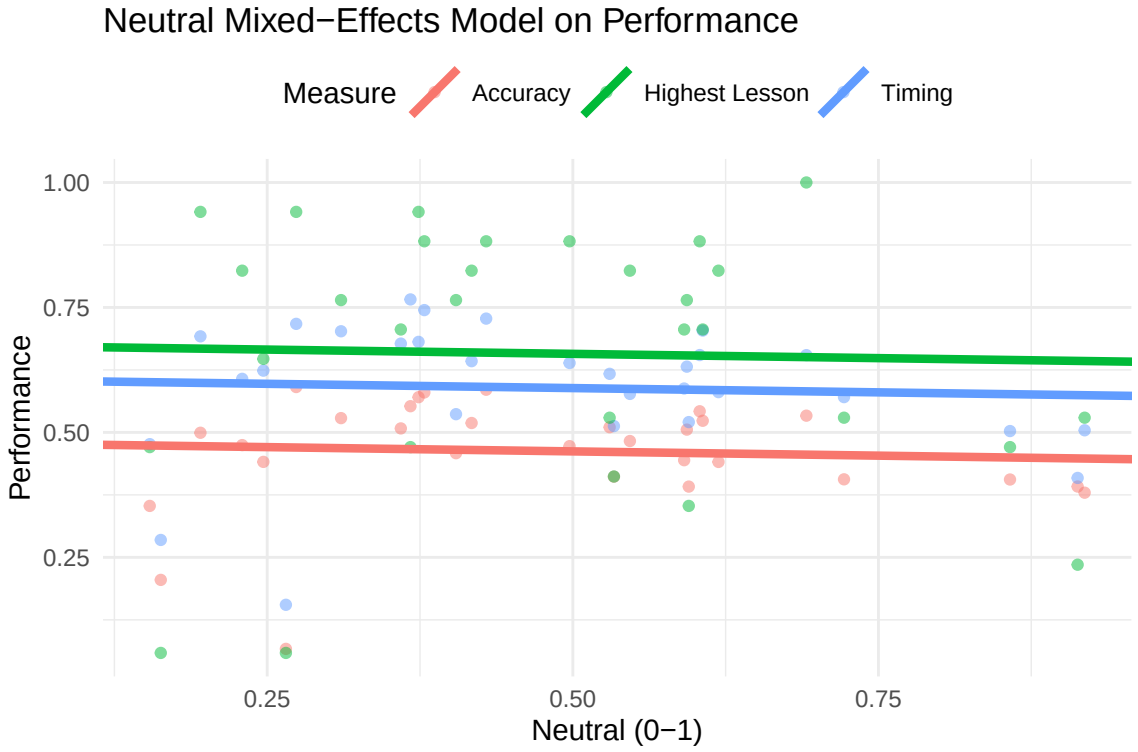

Figure 13: Neutral Mixed-Effects Model for Performance.

Additionally, I create another mixed-effects model to predict performance using all emotions. This model's results are displayed in Table 5. From this, fear and surprise have the largest positive effects on performance with betas of $\beta_{fear} = 0.16$ and $\beta_{surprise} = 0.35$.

Table 5: Results of Mixed-Effects Model with All Emotions.

	Estimate	Std. Error	t value
(Intercept)	0.57	0.10	5.54
sadness	-0.14	0.21	-0.66
anger	0.05	0.38	0.13
fear	0.16	0.20	0.78
disgust	-1.69	1.12	-1.51
happiness	0.04	1.33	0.03
surprise	0.35	0.62	0.57

Notes: The intercept for the model, as well as 6 emotions, are displayed in the first column of the table. The standard error and t-values are displayed. Note that neutral emotion is not included in this model, as the rank of the model was too low in order to proceed with running the model.

Table 6: Mixed-effects Model with Fear and Surprise Only.

	Estimate	Std. Error	t value
(Intercept)	0.51	0.08	6.76
surprise	0.55	0.54	1.00
fear	0.22	0.19	1.17

Notes: The intercept for the model, as well as the intercepts for surprise and fear are displayed in the first column of the table. The standard error and t-values are displayed.

I then run a final mixed-effects model that just uses fear and surprise to predict performance, displayed in Table 6. The betas here were 0.55 for surprise and 0.22 for fear, meaning that with every 1 unit increase in surprise, there is a 0.55 increase in performance, as well as for every 1 unit increase in fear, there is a 0.22 percent increase in performance. Thus, the final model for predicting performance is seen in Equation 1, where $\beta_{fear} = 0.22$ and $\beta_{surprise} = 0.55$, x_{fear} and $x_{surprise}$ are the values of fear and surprise displayed by the student respectively, a is the intercept, and ϵ is the term accounting for any noise:

$$\boxed{y_{performance} = a + \beta_{fear} x_{fear} + \beta_{surprise} x_{surprise} + \epsilon} \quad (1)$$

4.3 Inclusion of Affect Variables in Affective Cognitive Tutor

I first ran a two-tailed t-test between the control and experimental groups to determine if including affect into the decision-making of the cognitive tutor had an effect on any of the three performance measures for students. A t-test was run between the groups for overall accuracy, overall timing, and overall highest lesson values for each participant. The results are shown below in Table 7. As shown in the table, there was no significant difference between groups in accuracy, timing, or highest lesson. The effect size of each t-test was negligible, with the Cohen's d value being .18, .19, and -.10 for Accuracy, Timing, and Highest Lesson. In addition, I plot out all of the overall performance values per participant in each group, shown in Figure 14.

Table 7: T-test on Performance Metrics between Control and Experimental Groups.

	Metric	n1	n2	statistic	df	p	d	Effect mag.
1	Accuracy	30	20	-0.66	47.61	0.52	0.18	negligible
2	Timing	30	20	-0.67	47.99	0.51	0.19	negligible
3	Highest Lesson	30	20	0.35	42.23	0.72	-0.10	negligible

* $p < .05$, ** $p < .01$, *** $p < .001$ .001, **** $p < .0001$.

Notes: Cohen's d was used to calculate effect size. The column "n1" refers to the number of data points analyzed in the control group, and "n2" refers to the number of data points analyzed in the experimental group.

I then ran another set of two-tailed t-tests between the control and the experimental groups to determine if including affect into the decision-making of the cognitive tutor had an effect on accuracy or timing measures, looking just at the accuracy and timing from each participant's highest level attempted. For this, I took the mean of the accuracy scores from each attempt on every participant's highest lesson; I then took the mean of the timing scores from each attempt on every participant's highest lesson. Finally, I tested the difference between the highest lessons for each group. The results of the three t-tests are shown below

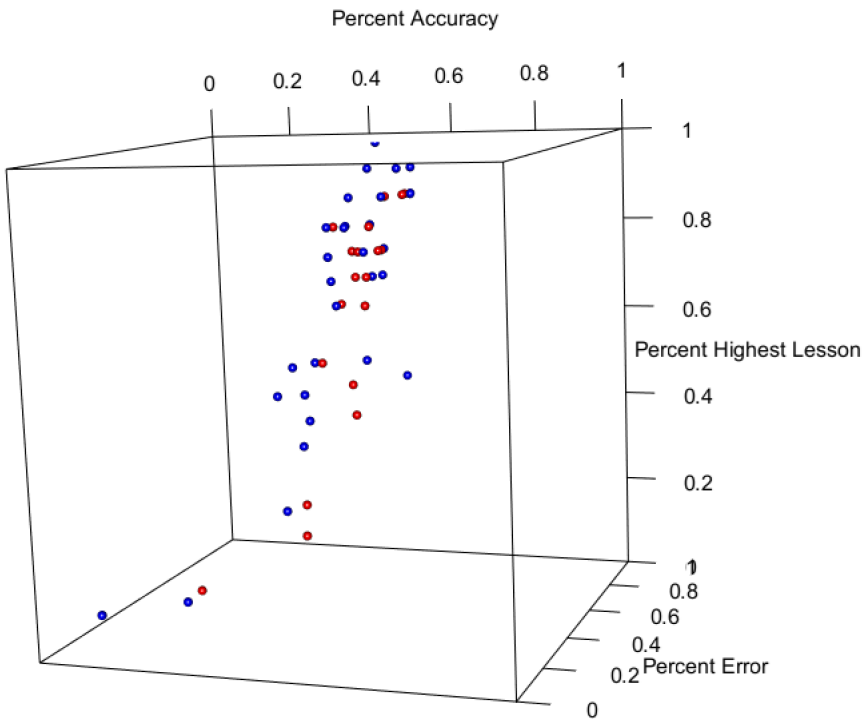

Figure 14: Performance of each group plotted on 3 axes.

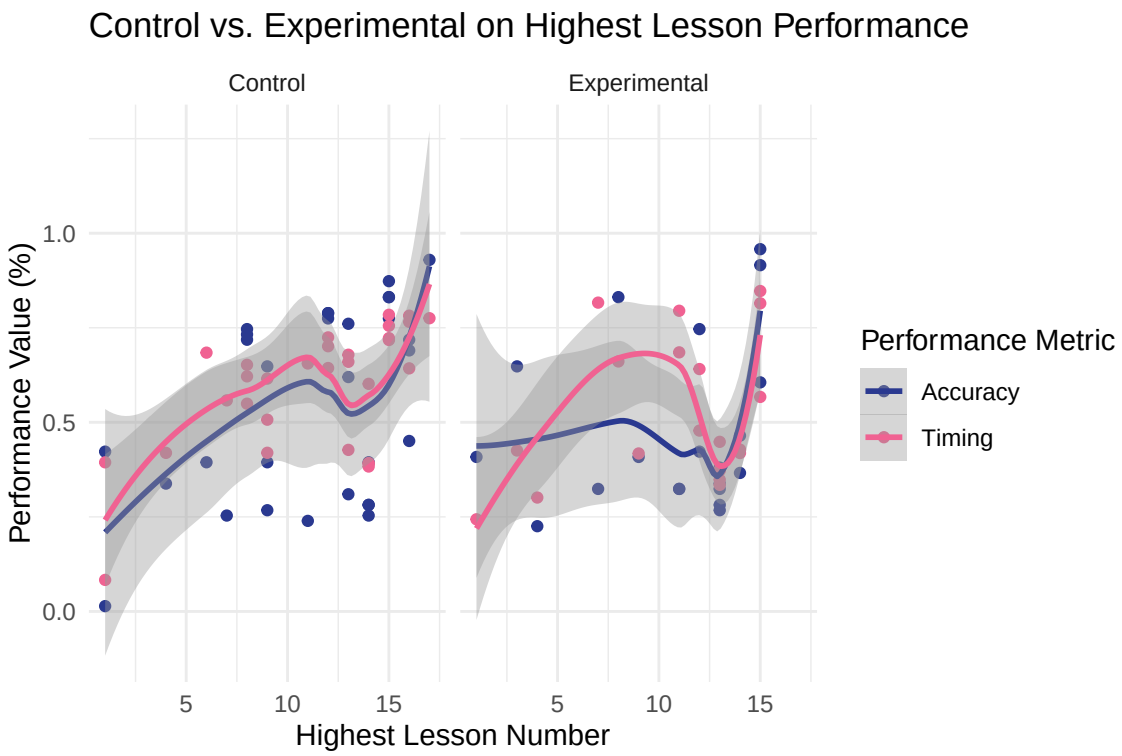

Figure 15: Performance on highest lesson per participant per group.

in Table 8. None of the t-tests were statistically significant. The effect size of the treatment on accuracy for the highest lesson was .31, the effect size of the treatment on timing for the highest lesson was .39, and the effect size of the treatment on highest lesson was .10. The accuracy and timing scores for each participant's highest lesson are shown in Figure 15.

Table 8: T-test on performance metrics during last lesson between control and experimental Groups.

	Metric	n1	n2	statistic	df	p	d	Effect mag.
1	Accuracy	30	20	1.07	43.77	0.29	0.31	small
2	Timing	30	20	1.33	36.40	0.19	0.39	small

* $p < .05$, ** $p < .01$, *** $p < .001$ .001, **** $p < .0001$.

Notes: Cohen's d was used to calculate effect size. The column "n1" refers to the number of data points analyzed in the control group, and "n2" refers to the number of data points analyzed in the experimental group.

Next, I investigated the results to see if there was a difference in accuracy and timing over all lessons for each group. To do this, I found the mean accuracy for each lesson for each participant, as well as the mean timing for each lesson for each participant. The results are

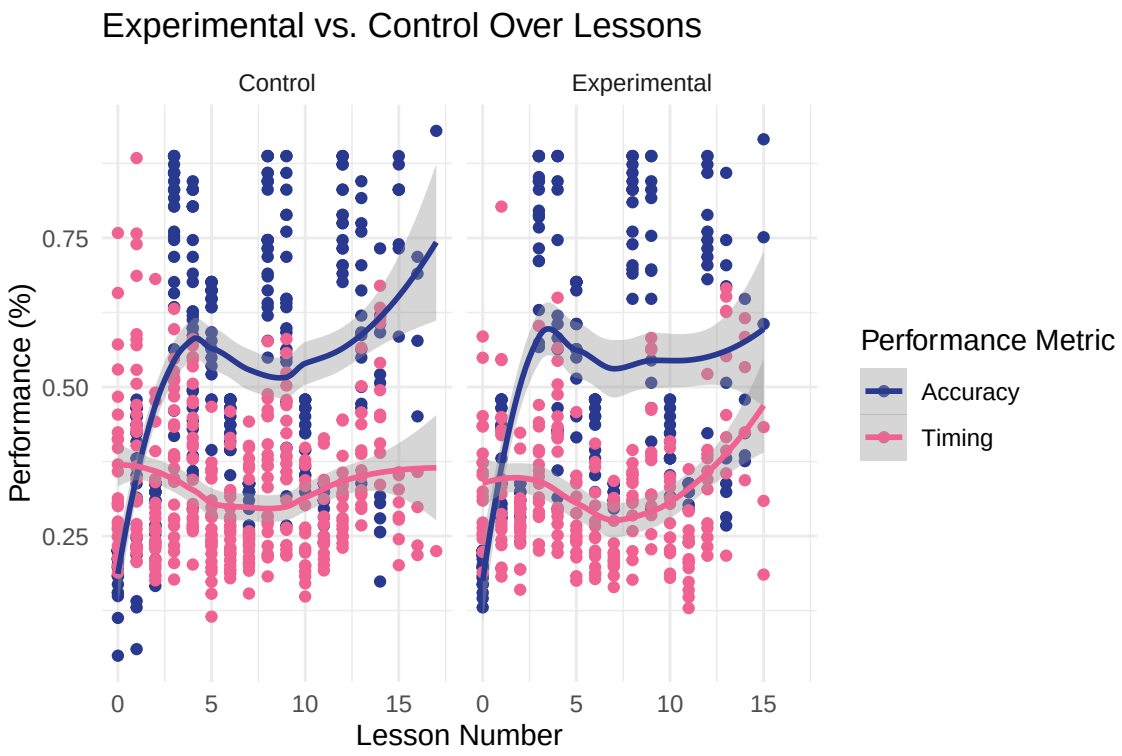

Figure 16: Each treatment group's performance over time.

displayed in Figure 16 for each treatment group. As displayed, both groups seem to perform almost equally as well over time on the lessons.

I then tested to see if there was any difference between the control and the experimental in the number of attempts each student spent on each lesson. To do this, I summed the number of attempts each participant spent on each lesson, and then, to reduce heterogeneity, I grouped each student into a high performing or a low performing category. To create the high performing and low performing categories, I found the mean highest lesson for all students combined to be 12.5; students who advanced to level 13 or higher were placed in the high performing group, while students who did not advance past level 12 were placed in the low performing group. The figure plotting these results can be found in Figure 17. Additionally, I created a mixed-effects model to test for any interaction between the treatment and the lesson on the number of attempts spent on that lesson, with the participant being held as a random effect. The results of this mixed-effects model can be seen in Table

9. While the model does not have a high enough t-value to show statistical significance, Figure 17 shows participants in the low performing group spending more attempts on earlier lessons in the treated condition than the control condition.

Table 9: Mixed-effects model for the interaction between lesson and treatment group on number of attempts per lesson.

	Estimate	Std. Error	t value
(Intercept)	1.78	0.29	6.14
Treatment Group	0.22	0.46	0.47
Lesson Number	0.02	0.02	1.30
Interaction	-0.02	0.03	-0.72

Notes: This model holds the specific participant as a random effect to attempt to account for heterogeneity of participants. The interaction specified is the interaction between the treatment group (control or experimental) and the specific lesson the participant is currently attempting.

5 Discussion

5.1 Interpretation

The results from the first part of the study support the hypothesis that emotions translated from displayed affect collected from a webcam does correlate with rhythm performance. From the principal component analyses in Figures 4 through 6, three main clusters appear of emotions when considering their correlations on performance: sadness, disgust, and anger are negatively correlated with performance; happiness, fear, and surprise are positively correlated with performance; and neutral seems to have little positive or negative correlation to performance. The 2D PCA displayed in Figure 4 shows these three clusters of emotions, but the 3D PCA displayed in Figures 5 and 6 then reveals that disgust and anger are on the opposite side of an axis from fear, and neutral is on the opposite side of the axis from surprise and happiness. Sadness as well splits off from disgust and anger when a third component is used to examine the data.

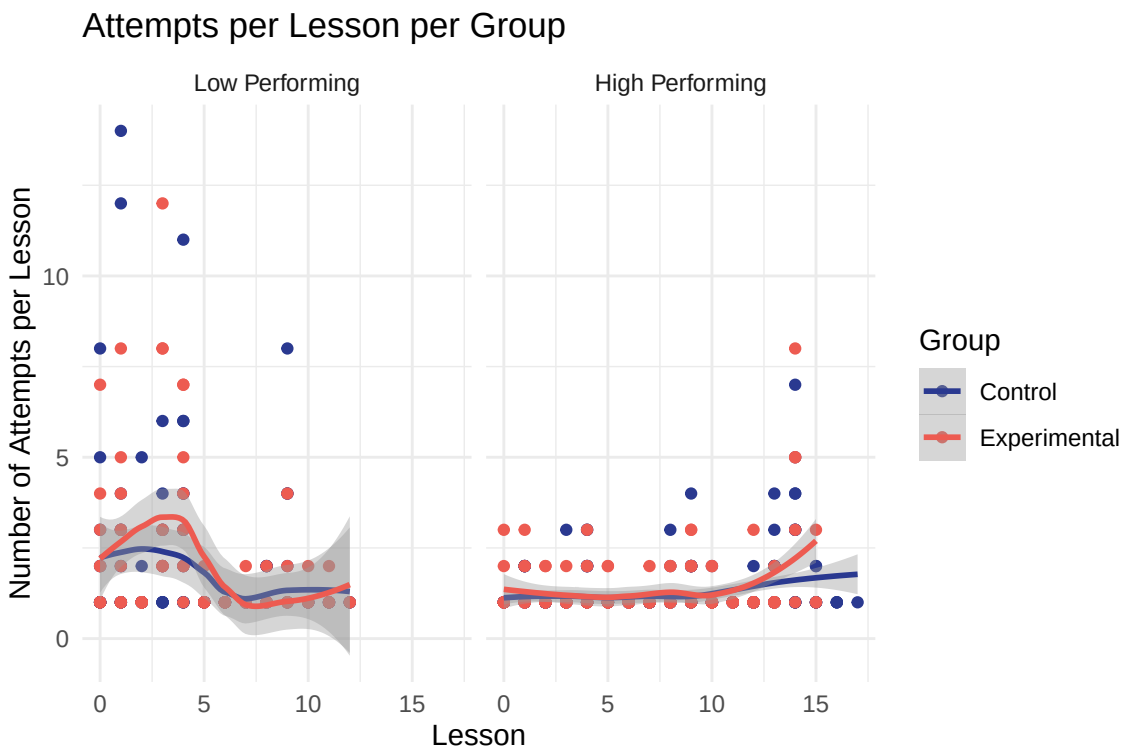

Figure 17: Attempts per Lesson per Group, Faceted by Overall Performance.

From the individual mixed-effects models in Table 4, fear and surprise seem to predict performance the best. Disgust has a small range of 0 to 0.15 only; this range makes it much more difficult to use as a predictor of performance. Additionally, happiness has an even smaller range of 0-0.1; it also seems to show the same information as surprise according to the 3D PCA. The remaining models show a larger range of scores overall, where fear and surprise remain as the emotions that predict positive performance. Finally, looking at the attempts per lesson per group in Figure 17, the low performing group seems to spend more attempts on earlier lessons in the experimental condition than the treated condition. This may suggest that the affective cognitive tutor is rightfully preventing low-performers from continuing until they understand the basics well, as opposed to spending more attempts later on during more difficult lessons.

These results have implications for the future of cognitive tutors. Using emotions, cognitive tutors could be able to predict and monitor performance and perhaps future per-

formance, at least for rhythm-learning applications, by monitoring levels of fear and surprise. Thus, cognitive tutors can then adapt their teaching style to the learner, similar to how human tutors adapt their approach mid-session (Alexander et al., 2008; Lehman et al., 2008). Furthermore, this rhythm-teaching cognitive tutor can be programmed to look specifically for surprise and fear, rather than other emotions that seem like they might have been useful like anger or sadness.

The second part of the results of this study found no significant difference between the control and experimental groups in terms of their performance overall or performance on their highest lesson; additionally, their performance overtime seems similar. There are many pieces to interpret from these results. The results suggest that utilizing emotional affect from a webcam in the model in the current study does not significantly affect performance. There are many possible reasons for this outcome. First, because the experimental condition still relies on the accuracy and timing scores, it is expected that the groups have similar outcomes overall. Additionally, the results shown could just be due to the heterogeneity of the participants, as their experience with music could differ, as well as their familiarity with utilizing a computer system tutor like the one in the present study. Furthermore, since each group only tests within a time-span of 35 minutes, it is difficult to measure performance and improvement over time; a longitudinal study would add to the current study's findings greatly. Another possible implication is that the model to predict performance must be tweaked. When reviewing the data from the experimental group, few participants ever displayed enough surprise and fear combined to receive a beneficial affect score; this could mean that the binning for affect scores must be tweaked in order to give more weight to very small values of fear and surprise.

Overall, however, participants seem to show some sign of improvement over time and over each highest lesson regardless of group. This result supports the hypothesis that a cognitive tutor in the field of music can be successful in and helpful to student learning. As

this is the first cognitive tutor documented to focus on teaching music, and this cognitive tutor was successful in its teaching, this study raises the need and opportunity for more cognitive tutors to be created in the field of music. Additionally, this study serves to present a successful cognitive tutor in a non-STEM field. Regardless of affect, the cognitive tutor in the current study pushes the literature to explore the possibilities of cognitive tutors further.

5.2 Limitations

There are some limitations within the current study. First, the number of total participants was small, especially when calculating the betas to be used in the experimental condition. Thus, the model for predicting performance using displayed affect could be strengthened and modified with more participants. Regardless, more participants would have increased the sample size and the power of the statistical methods utilized in the study. While this does limit the statistical power of the study, the study can still act as a proof of concept that there are trends of benefits to utilizing affect within a cognitive tutor as well as using a cognitive tutor in the field of music.

Another limitation is the recruiting of participants. Of the participants tested in-person, a large number of them were Dartmouth College students; this fact could cause the study to be less generalizable to other populations. Dartmouth students could be generally faster learners, display more or less affect than the average student, or could be more talented with learning rhythm, causing them to succeed more or less than the average student. While this is a limitation of the study, further work can be done to test other populations with an affective cognitive tutor in the field of music as this study has shown a valid proof of concept for the cognitive tutor. Furthermore, there could be differences between taking the study in-person and taking the study on Amazon MTurk as attention levels could be decreased, or the amount of time spent actively focused on the study could be lower. While this could cause innate differences between participants in-person and on MTurk, it does allow for imagining the cognitive tutor being used by a variety of students with different attention

spans, helping to paint an idea of the tutor's use within a music classroom.

Continuing, the number of frames analyzed of each attempt's video of the participant presents an additional limitation for the tutor's effectiveness. In order for the application to run quickly and in real-time, only every 98th frame can be analyzed to create the table of emotions for that participant's attempt of any given lesson; thus, frames with important affect data could be missed, quick emotional changes. This may lead to less accurate calculated affect overall. Additionally, some affect could have been misdiagnosed; for instance, concentration could have been labeled as neutral emotion, or furrowed brows in concentration could have been identified as anger. These misidentifications could have led to slightly inaccurate affect data for any given attempt. Finally, after a participant missed multiple attempts in a row, they could begin to be frustrated with the lesson; in these cases, it could have been better for the application to move onto another lesson rather than stay on the same one, regardless of performance.

5.3 Future Work

While this study helps show important trends in administering an affective cognitive tutor within the field of music, there is still much research to be done on creating new affective cognitive technologies. First, the way that emotions are collected within the study could be adapted and refined to create an even better cognitive tutor. For instance, while utilizing webcam technologies to track facial expressions has been useful and effective in past cognitive tutors and in this cognitive tutor (Zakharov, 2007; Spaulding et al., 2016), webcams could pose an issue in collecting facial affect when a user's face is not in frame and the face cannot be detected. Furthermore, webcams have been shown to cause privacy concerns, especially within schools (Squelch and Squelch, 2005). Thus, creating an affective cognitive tutor that utilizes other or multiple forms of biofeedback, such as EEG or heart-rate, could be beneficial to the effectiveness of the tutor. Additionally, heart-rate wearables like the Apple Watch are becoming more popular and common and provide very reliable

heart-rate data (Hernando et al., 2018); this could make cognitive tutors more accurate through a more available form of biofeedback monitoring as well. Other technologies and algorithms to collect biofeedback, such as analyzing the student's verbal affect through tone analysis, could as well be implemented in the future.

The cognitive tutor could also be improved by tweaking its analysis and modeling of emotion. First, enabling the tutor to recognize higher-level emotions such as confusion, boredom, or concentration could be effective. While Ekman's emotions are useful and have been used prior, being able to categorize action units into higher-level emotions such as boredom or concentration could help further analyze the student's emotional state and understanding. Another tweak could be weighting the emotional data for each attempt based on the time at which the emotion is displayed by the user. For instance, users may display a great amount of emotion immediately after the attempt, but not during the attempt; the emotion displayed immediately after might be more important, as the user could sigh with relief, grunt with frustration, or smile with confidence. Thus, further testing into how accurately certain times of displayed facial affect correlate with performance could be useful to the effectiveness of an affective cognitive tutor.

Another addition to the cognitive tutor could be implementing a calibration feature that calibrates the student's displayed affect levels before the student starts the first lesson. Students may display differing levels of facial affect: one student could be very expressive while another student could show very little expression on their face. Because the current cognitive tutor uses the student's displayed facial affect to help decide whether or not the student should continue to the next lesson, students who naturally display less facial affect could be at a disadvantage as it may be more difficult to proceed to the next lesson. However, integrating a calibration feature into the cognitive tutor before the student begins taking lessons could help evaluate students whom naturally display minimal affect. A calibration feature would allow the cognitive tutor to first measure the student's naturally displayed

affect; the cognitive tutor could then utilize the calibration results to evaluate the student's displayed affect on lessons. For instance, if Student A shows a maximum of 30% fear during the calibration session and Student B shows a maximum of 80% fear on the calibration session, and both students then show 20% fear on a lesson later on, student A would have a higher affect score than student B, due to the personalized calibration of the cognitive tutor. Thus, incorporating calibration of student affect into the cognitive tutor could create an even more accurate algorithm for analyzing student performance.

Finally, cognitive tutors should continue to expand into other subject fields as well as test with real students within classrooms. While this affective cognitive tutor is the first of its kind for the field of music, STEM fields are still the most prevalent fields of education for cognitive tutors (Ritter et al., 2007, Ma et al., 2014, Pane et al., 2014, Supekar et al., 2015, Marouf and Abu-Naser, 2019); exploring other educational fields would provide a further understanding of the effectiveness of cognitive tutors. Additionally, the current study did not test participants longitudinally over multiple sessions; thus, the current study could not evaluate whether students retained information taught from the affective cognitive tutor. Further research should implement a longitudinal version of the tutor, allowing the correlation between retention of knowledge and displayed facial affect to be analyzed. Finally, bringing an affective cognitive tutor like the one in the current study into the classroom to be used with actual music students could help show the effectiveness of the cognitive tutor within the classroom. Data on an affective cognitive tutor's success inside of an actual classroom setting could provide insight on the tutor's true effectiveness and usefulness to the field of education.

6 Conclusion

This study utilized 59 participants to determine the relationship between displayed facial affect and rhythm performance as well as to understand whether an affective cognitive

tutor can be useful and effective within the educational field of music. A cognitive tutoring application was created that collects biofeedback through usage of a webcam and translating displayed facial affect to emotions. Mixed-effects models were created to analyze the relationship between emotions collected by webcam and three measures of performance. The strongest model was then used to create an experimental condition that allowed the cognitive tutor to utilize affect in determining a student's performance on each attempt of a lesson. The effectiveness of this affective cognitive tutor was then compared against a non-affective cognitive tutor with a two-tailed t-test.

The results found that the emotions of fear and surprise are correlated the strongest with the three measures of performance; the results also found three separate clusters of emotions based on their effects on performance, with fear, surprise, and happiness yielding positive effects, neutral emotion having no effect, and sadness, anger, and fear yielding negative effects on performance. Second, the results found no significant difference between the control and experimental groups in terms of performance. While this seems unsatisfying, this still has great implications for future work as well as affective cognitive tutors in general. Through the current study, gaps in the literature have been filled in, including creating a cognitive tutor within the field of music, as well as creating a new affective cognitive tutor that utilizes facial affect data in real-time to guide a tutoring session. With the findings of this study, much future work remains into the field of cognitive tutors and affect, as the goal of creating an even more human-like cognitive tutor is well within reach.

References

Ajisoko, P. (2020). The use of duolingo apps to improve english vocabulary learning. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(7), 149–155.

Alexander, S., Sarrafzadeh, A., Hill, S., et al. (2006). Easy with eve: A functional affective tutoring system. *workshop on motivational and affective issues in ITS. 8th international conference on ITS*, 5–12.

Alexander, S., Sarrafzadeh, A., & Hill, S. (2008). Foundation of an affective tutoring system: Learning how human tutors adapt to student emotion. *International journal of intelligent systems technologies and applications*, *4*(3-4), 355–367.

Anderson, J. R. (2013). *The architecture of cognition*. Psychology Press.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, *228*(4698), 456–462.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, *4*(2), 167–207.

Badri, M. (2021). Adoption of innovations online tutoring apps on high school students. *Journal of Physics: Conference Series*, *1823*(1), 012026.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602–607.

Cabada, R. Z., Estrada, M. L. B., Garcìa, C. A. R., Pérez, Y. H., et al. (2012). Fermat: Merging affective tutoring systems with learning social networks. *2012 IEEE 12th International Conference on Advanced Learning Technologies*, 337–339.

Calvo, R. A., & D'Mello, S. (2012). Frontiers of affect-aware learning technologies. *IEEE Intelligent Systems*, *27*(6), 86–89.

Calvo, R. A., D'Mello, S., Gratch, J. M., & Kappas, A. (2015). *The oxford handbook of affective computing*. Oxford Library of Psychology.

Camilleri, A. C., & Camilleri, M. A. (2019). Mobile learning via educational apps: An interpretative study. *Proceedings of the 2019 5th International Conference on Education and Training Technologies*, 88–92. https://doi.org/10.1145/3337682.3337687

Chen, C.-M., & Li, Y.-L. (2010). Personalised context-aware ubiquitous learning system for supporting effective english vocabulary learning. *Interactive Learning Environments*, *18*(4), 341–364.

Chen, J., Zhang, M., Xue, X., Xu, R., & Zhang, K. (2017). An action unit based hierarchical random forest model to facial expression recognition. *ICPRAM*, 753–760.

Cheong, J. H., Xie, T., Byrne, S., & Chang, L. J. (2021). Py-feat: Python facial expression analysis toolbox. *arXiv preprint arXiv:2104.03509*.

Council, N. R. et al. (2003). *Strategic education research partnership*. National Academies Press.

Crowston, K. (2012). Amazon mechanical turk: A research tool for organizations and information systems scholars. *Shaping the future of ict research. methods and approaches* (pp. 210–221). Springer.

Dimberg, U. (1982). Facial reactions to facial expressions. *Psychophysiology*, *19*(6), 643–647.

D'Mello, S., Picard, R. W., & Graesser, A. (2007). Toward an affect-sensitive autotutor. *IEEE Intelligent Systems*, *22*(4), 53–61.

Egilmez, H. O. (2012). Music education students' views related to the piano examination anxieties and suggestions for coping with students' performance anxiety. *Procedia-Social and behavioral sciences*, *46*, 2088–2093.

Ekman, P., Friesen, W., & Hager, J. (2002). The facial action coding system: A technique for the measurement of facial movement. a human face. *I-Tech Education and Publishing, Vienna*.

Faghihi, U., Fournier-Viger, P., & Nkambou, R. (2013). Celts: A cognitive tutoring agent with human-like learning capabilities and emotions. *Intelligent and adaptive educational-learning systems* (pp. 339–365). Springer.

Gilbert, C., & Moss, D. (2003). Biofeedback and biological monitoring. *Handbook of mind-body medicine in primary care: Behavioral and physiological tools*, 109–122.

Gordon, R. L., Jacobs, M. S., Schuele, C. M., & McAuley, J. D. (2015). Perspectives on the rhythm–grammar link and its implications for typical and atypical language development. *Annals of the New York Academy of Sciences*, *1337*(1), 16–25.

Grawemeyer, B., Mavrikis, M., Holmes, W., Gutiérrez-Santos, S., Wiedmann, M., & Rummel, N. (2017). Affective learning: Improving engagement and enhancing learning with affect-aware feedback. *User Modeling and User-Adapted Interaction*, *27*(1), 119–158.

Grinberg, M. (2018). *Flask web development: Developing web applications with python.* " O'Reilly Media, Inc.".

Guo, J., Zhu, X., & Lei, Z. (2018). 3ddfa.

Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., & Li, S. Z. (2020). Towards fast, accurate and stable 3d dense face alignment. *Proceedings of the European Conference on Computer Vision (ECCV)*.

Hernando, D., Roca, S., Sancho, J., Alesanco, Á., & Bailón, R. (2018). Validation of the apple watch for heart rate variability measurements during relax and mental stress in healthy subjects. *Sensors*, *18*(8), 2619.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., Mark, M. A., et al. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30–43.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wróbel, M. R. (2015). Modeling emotions for affect-aware applications. *Information Systems Development and Applications*, 55–69.

Landowska, A. (2014). Affective learning manifesto-10 years later. *European Conference on e-Learning*, 281.

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What are you feeling? investigating student affective states during expert human tutoring sessions. *International conference on intelligent tutoring systems*, 50–59.

Lindsley, D. B. (1952). Psychological phenomena and the electroencephalogram. *Electroencephalography & Clinical Neurophysiology*.

Liu, Y., & Sourina, O. (2013). Eeg databases for emotion recognition. *2013 international conference on cyberworlds*, 302–309.

Liu, Y., Sourina, O., & Hafiyyandi, M. R. (2013). Eeg-based emotion-adaptive advertising. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 843–848.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A duolingo case study. *ReCALL*, *31*(3), 293–311.

Luan, P., Huynh, V., & Tuan Anh, T. (2020). Facial expression recognition using residual masking network. *IEEE 25th International Conference on Pattern Recognition*, 4513–4519.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of educational psychology*, *106*(4), 901.

Madan, C. R., Harrison, T., & Mathewson, K. E. (2018). Noncontact measurement of emotional and physiological changes in heart rate from a webcam. *Psychophysiology*, *55*(4), e13005.

Magdin, M., Turcani, M., & Hudec, L. (2016). Evaluating the emotional state of a user using a webcam.

Marouf, A. M., & Abu-Naser, S. S. (2019). Intelligent tutoring system for teaching computer science i in al-azhar university, gaza. *International Journal of Academic and Applied Research (IJAAR)*, *3*(3), 31–53.

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, *29*(1), 142–163.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra i at scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Paranjape, R., Mahovsky, J., Benedicenti, L., & Koles, Z. (2001). The electroencephalogram as a biometric. *Canadian Conference on Electrical and Computer Engineering 2001. Conference Proceedings (Cat. No. 01TH8555)*, *2*, 1363–1366.

Patston, T., & Osborne, M. S. (2016). The developmental features of music performance anxiety and perfectionism in school age music students. *Performance Enhancement & Health*, *4*(1-2), 42–49.

Rathod, P., George, K., & Shinde, N. (2016). Bio-signal based emotion detection device. *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 105–108.

React – a javascript library for building user interfaces. (n.d.). https://reactjs.org/

Research and participant management made easy in the cloud. (n.d.). https://www.sona-systems.com/

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic bulletin & review*, *14*(2), 249–255.

Ritter, S., Carlson, R., Sandbothe, M., & Fancsali, S. E. (2015). Carnegie learning's adaptive learning products. *Educational Data Mining*, *2015*, 8th.

Schwartz, M. S., & Andrasik, F. (2017). *Biofeedback: A practitioner's guide*. Guilford Publications.

Sourina, O., & Liu, Y. (2013). Eeg-enabled affective applications. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 707–708.

Spaulding, S., Gordon, G., & Breazeal, C. (2016). Affect-aware student models for robot tutors.

Squelch, J., & Squelch, A. (2005). Webcams in schools: A privacy menace or a useful monitoring tool? *Australia and New Zealand Journal of Law and Education*.

Supekar, K., Iuculano, T., Chen, L., & Menon, V. (2015). Remediation of childhood math anxiety and associated neural circuits through cognitive tutoring. *Journal of Neuroscience*, *35*(36), 12574–12583.

Thakor, N. V., & Tong, S. (2004). Advances in quantitative electroencephalogram analysis methods. *Annu. Rev. Biomed. Eng.*, *6*, 453–495.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, *46*(4), 197–221.

Vinchurkar, D. P., & Sasikumar, M. (2015). Intelligent tutoring system for voice conversion in english. *2015 IEEE 15th International Conference on Advanced Learning Technologies*, 314–316.

Vora, K., Shah, S., Harsoda, H., Sheth, J., & Thakkar, A. (2020). Necessary precautions in cognitive tutoring system. *Intelligent communication, control and devices* (pp. 445–452). Springer.

Woody, R. (2020). Dispelling the die-hard talent myth: Toward equitable education for musical humans. *The American Music Teacher*, *70*(2), 22–25.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognising and responding to student affect. *International Journal of Learning Technology*, *4*(3-4), 129–164.

Wristen, B. G. (2013). Depression and anxiety in university music students. *Update: Applications of Research in Music Education*, *31*(2), 20–27.

Zakharov, K. (2007). Affect recognition and support in intelligent tutoring systems.

Supplementary Materials

Appendix A: Supplementary Tables

Table A1: Demographics Table

Demographic Category	Control(N=39)	Experimental(N=20)	Total(N=59)
Gender			
Female	18 (46.2%)	9 (45.0%)	27 (45.8%)
Male	21 (53.8%)	11 (55.0%)	32 (54.2%)
Age			
18-24	27 (69.2%)	20 (100%)	47 (79.7%)
25-34	4 (10.3%)	0 (0%)	4 (6.8%)
35-44	6 (15.4%)	0 (0%)	6 (10.2%)
45-54	2 (5.1%)	0 (0%)	2 (3.4%)
Ethnicity			
Hispanic or Latinx	3 (7.7%)	3 (15.0%)	6 (10.2%)
Not Hispanic or Latinx	36 (92.3%)	16 (80.0%)	52 (88.1%)
Prefer not to say	0 (0%)	1 (5.0%)	1 (1.7%)
Race			
Asian	2 (5.1%)	3 (15.0%)	5 (8.5%)
Black or African American	2 (5.1%)	1 (5.0%)	3 (5.1%)
Indigenous American or Alaska Native	1 (2.6%)	0 (0%)	1 (1.7%)
Prefer not to say	1 (2.6%)	1 (5.0%)	2 (3.4%)
White	33 (84.6%)	15 (75.0%)	48 (81.4%)

Table A2: Demographics Table: Can you read basic sheet music?

Response	Control(N=39)	Experimental(N=20)	Total(N=59)
Yes	23 (59.0%)	14 (70.0%)	37 (62.7%)
No	13 (33.3%)	4 (20.0%)	17 (28.8%)
Other	3 (7.7%)	2 (10.0%)	5 (8.5%)

Table A3: Lesson Format

Lesson Number	Lesson Notes	BPM	Number of Instruments
1	Quarter	60	1
2	Quarter, Half	60	1
3	Quarter, Half, Whole	60	1
4	Quarter, Half, Whole, Eighth	60	1
5	Quarter, Half, Whole, Eighth	75	1
6	Quarter	60	2
7	Quarter, Half	60	2
8	Quarter, Half, Whole	60	2
9	Quarter, Half, Whole, Eighth	60	2
10	Quarter, Half, Whole, Eighth	75	2
11	Quarter, Half	60	3
12	Quarter, Half, Whole	60	3
13	Quarter, Half, Whole, Eighth	60	3
14	Quarter, Half, Whole, Eighth	75	3
15	Quarter, Half, Whole, Eighth	100	3
16	Quarter, Half, Whole, Eighth	60	3
17	Quarter, Half, Whole, Eighth	60	3
18	Quarter, Half, Whole, Eighth	60	3

Notes: For each lesson, the types of notes are displayed in the lesson notes column, the BPM that the activity is played with is shown, and the number of keys used (called number of instruments) is shown. The lessons progress in order from the first lesson to the eighteenth lesson.

Table A4: Time spent attempting lessons per group.

	Condition	Time spent attempting (min)
1	Control	12.70
2	Experimental	15.00

Table A5: Mean of percentage of lessons reached.

Group	Mean of percentage of lessons reached.
Control	0.62
Experimental	0.64

Table A6: Perceived comfort after using the tutor.

Perceived Comfort	Group	n
Extremely comfortable	Control	11 (28.2%)
Extremely comfortable	Experimental	9 (45.0%)
Somewhat comfortable	Control	22 (56.4%)
Somewhat comfortable	Experimental	8 (40.0%)
Neither comfortable nor uncomfortable	Control	1 (2.6%)
Neither comfortable nor uncomfortable	Experimental	1 (5.0%)
Somewhat uncomfortable	Control	3 (7.7%)
Somewhat uncomfortable	Experimental	2 (10.0%)
Extremely uncomfortable	Control	2 (5.1%)
Extremely uncomfortable	Experimental	0 (0.0%)

Appendix B: Supplementary Figures

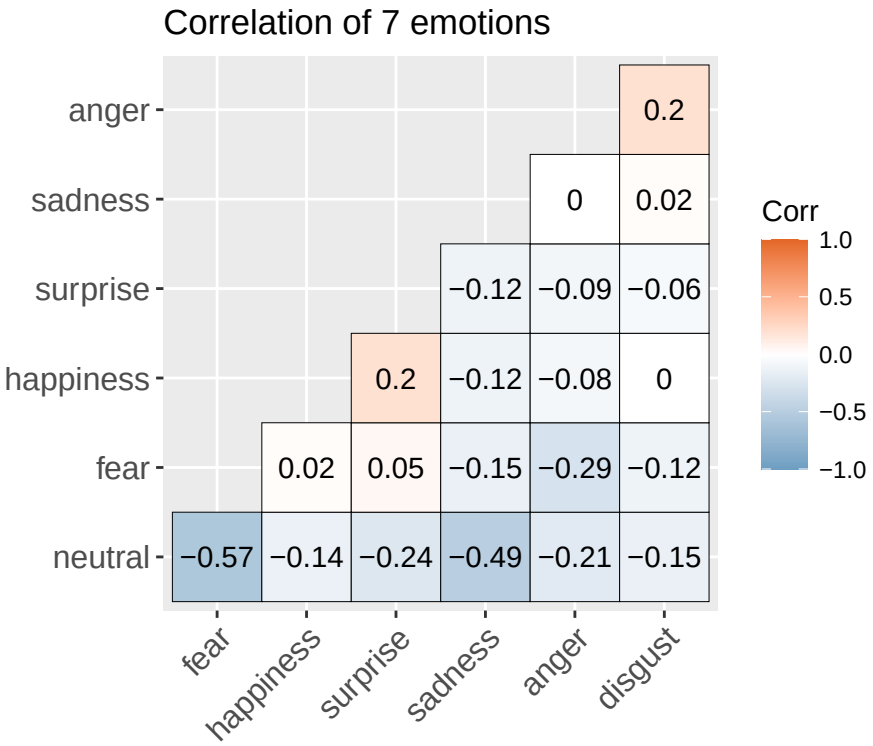

Figure B1: Correlation of 7 emotions in current study.

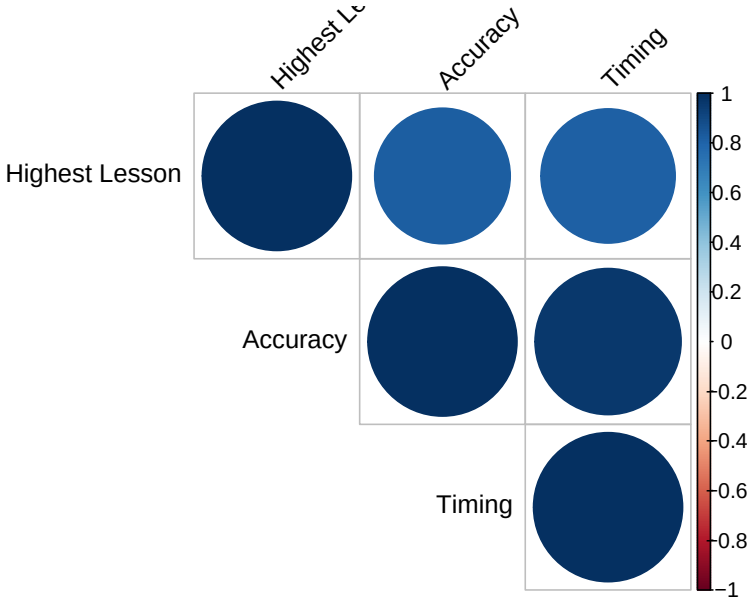

Figure B2: Correlation Matrix for 3 Performance Measures.

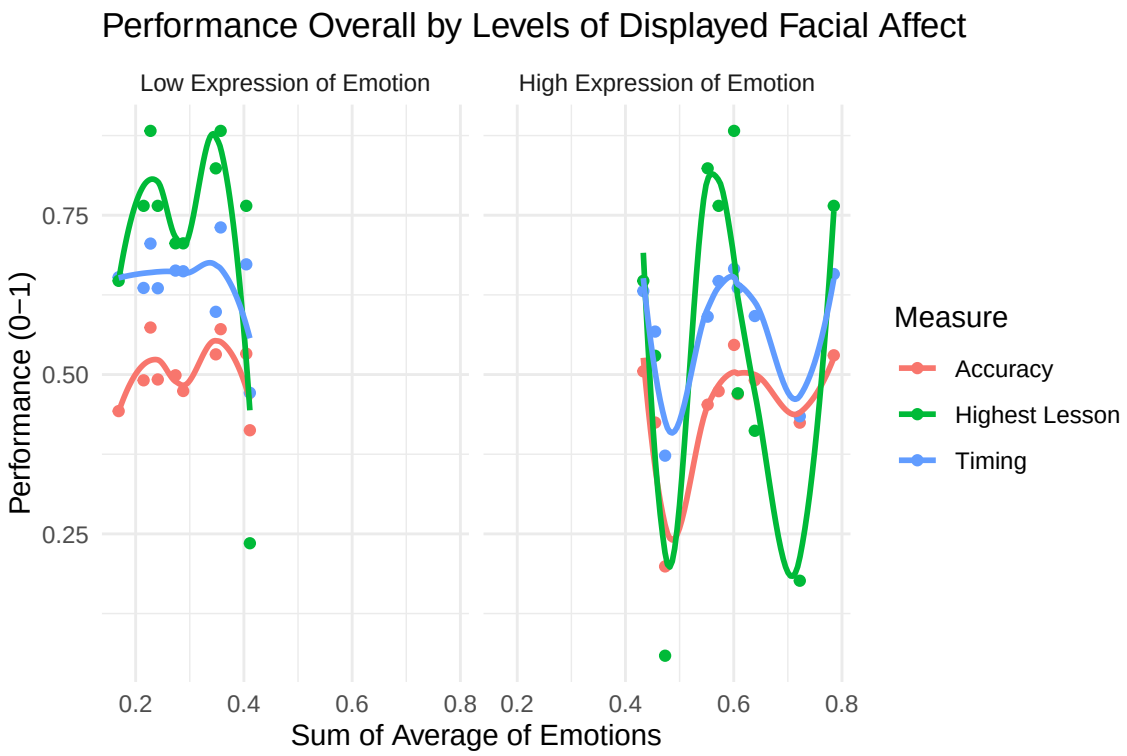

Figure B3: Performance Overall by Levels of Displayed Facial Affect.

Figure B3 shows a microanalysis of the correlation between the overall emotion displayed by a student and their overall performance for the treated condition, with each performance metric measured on a scale of 0-1. The overall emotion value was calculated for each participant by first taking the participant's average emotion scores from each level for each emotion, excluding neutral emotion. The average emotion scores from each level were then averaged together to result in 6 scores, one for each of the 6 remaining emotions. These were then summed to create the sum of average of emotions value displayed in Figure B3. The cutoff between Low Expression of Emotion and High Expression of Emotion was 0.422, the mean value of emotion expression shown by participants. Figure B3 shows much more variability of performance in students with high expression of emotion as opposed to students with low expression of emotion.