ABSTRACT

GRAFSGAARD, JOSEPH F. Multimodal Affect Modeling in Task-Oriented Tutorial Dialogue. (Under the direction of Dr. Kristy Elizabeth Boyer and Dr. James C. Lester.)

Recent years have seen a growing recognition of the central role of affect and motivation in learning. In particular, learning-centered cognitive-affective states such as anxiety, confusion, engagement, and frustration may co-occur with positive or negative cycles of learning. Just as highly effective human tutors pay attention to more than whether a student is simply correct or incorrect, automated approaches may be used to identify and understand students' nonverbal behaviors. Observed channels such as posture, gesture and facial expression provide key insights into students' affective and motivational states.

As students engage in computer-mediated task-oriented dialogue with a human tutor, their affective states are expressed through posture, gesture, and facial expression. Posture and gesture were tracked using novel algorithms on Kinect depth images. This allowed for automated labeling of whether a student was shifting posture or contacting the lower face with a hand. Facial expression was recognized using an existing tool, the Computer Expression Recognition Toolbox (CERT), which identified fine-grained facial movements at each frame of video. Prior analyses within this body of work have identified associations of nonverbal behavior and learning-centered affective states, such as anxiety, effortful concentration, engagement, and frustration.

Preliminary work used hidden Markov models (HMMs) to analyze sequences of affective tutorial interaction. Descriptive HMMs were machine-learned from the combined context of facial expression, tutorial dialogue, and task progress in an unsupervised approach using the Baum-Welch algorithm. Subsequent work focused on development of multimodal differential sequence mining, a technique that handles the large state space inherent in automatically generated multimodal data streams. This novel extension of differential sequence mining was applied to the multimodal data streams of student task actions, dialogue, and nonverbal behavior, including facial expression, gesture, and posture. Multimodal sequences were found to be associated with tutorial outcomes of engagement, frustration, and learning gain. Additionally,

incoming student characteristics of general and computer science self-efficacy were linked to other multimodal sequences. Among these results, one-hand-to-face gestures were found to occur more frequently with positive phenomena of engagement, learning, and general self-efficacy, while two-hands-to-face gestures occurred more frequently with frustration. This line of research improves automated understanding of learning-centered affect, with particular insights into how affect unfolds from moment to moment during tutoring. This may result in systems that treat student affect not as transient states, but instead as interconnected links in a student's path toward learning.

Multimodal Affect Modeling in Task-Oriented Tutorial Dialogue

by
Joseph F. Grafsgaard

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2014

APPROVED BY:

_____	_____
Dr. James C. Lester Co-Chair of Advisory Committee	Dr. Kristy Elizabeth Boyer Co-Chair of Advisory Committee
_____	_____
Dr. Tiffany M. Barnes	Dr. Eric N. Wiebe

BIOGRAPHY

Joseph F. Grafsgaard spent his early childhood on Guam, enjoying visits to Talofofo Falls and the coral reef. As a youth in Minnesota, he studied Japanese, Latin, and Chinese in middle school and high school, and attended college courses full-time in his final year of high school. He earned a bachelor's degree in computer science from University of Minnesota, Twin Cities, in 2005, also completing graduate courses in software engineering. Prior to entering the graduate program at North Carolina State University, he worked with a start-up company as a software developer, providing online and point-of-sale ticket services to regional race tracks. In graduate school, his coursework focused on artificial intelligence and machine learning. During the course of his graduate career, he was president of the Computer Science Graduate Student Association and assistant academic liaison of the STARS Student Leadership Corps. He earned a Master of Science degree in Computer Science from North Carolina State University in 2012, continuing on the path to the doctoral degree.

ACKNOWLEDGEMENTS

I would like to thank my advisors, James Lester and Kristy Elizabeth Boyer, for their guidance throughout the course of my dissertation research. This dissertation would not have been possible without their support. Their complementary approaches to advising provided help whenever I needed it and modeled career excellence of both new and established faculty. Additionally, I am grateful for the insightful feedback and advice of the members of my committee, Tiffany Barnes and Eric Wiebe.

I have had the good fortune to collaborate and interact with many amazing individuals. I thank my colleagues in the Center for Educational Informatics, the IntelliMedia Group, and the LearnDialogue Group for their collaboration and input over the course of my degree. I am particularly thankful for my collaborations with Alok Baikadi, Aysu Ezen-Can, Christopher Fox, Megan Frankosky, Eun Young Ha, Seung Lee, Scott McQuiggan, Christopher Mitchell, Jonathan Rowe, Jennifer Sabourin, Lucy Shores, and Andy Smith. I would also like to thank the exceptional undergraduate researchers with whom I have worked, including Stewart Cartmell, David Dearmore, Denae Ford, Robert Fulton, Natalie Kerby, Kate Lester, Kristin Rachor, Alexandria Vail, and Joseph Wiggins. Each of these young researchers has a great future ahead. I am also grateful for the interactions I have had with those in the broader research community, including Sharifa Alghowinem, Ivon Arroyo, Roger Azevedo, Ryan Baker, Robert Bixler, Nigel Bosch, Keith Brawner, Rafael Calvo, Jeffrey Cohn, Cristina Conati, Sidney D'Mello, Arthur Graesser, Jonathan Gratch, Mirela Gutica, Zakia Hammal, Jason Harley, Neil Heffernan, Ehsan Hoque, Natasha Jaques, Jyoti Joshi, Susanne Lajoie, Blair Lehman, Stacy Marsella, Elisabeth Meier, Brent Morgan, Jack Mostow, Nicholas Mudrick, Jaclyn Ocumpaugh, Andrew Olney, Zachary Pardos, Reinhard Pekrun, Rosalind Picard, Valerie Shute, Amber Strain, Michelle Taub, Wixon, Beverly Woolf, and Marcelo Worsley.

There are many individuals in the department and university that have worked toward the betterment of our community. I would like to specifically thank Barbara

Jasmine Adams, Carol Allen, Tammy Coates, Marhn Fullmer, Ron Hartis, Sarah Heckman, Carol Miller, Trey Murdoch, Douglas Reeves, Nagiza Samatova, Andrew Sleeth, Ken Tate, David Thuente, and Mladen Vouk for their exceptional efforts. I thank my peers involved in the Computer Science Graduate Student Association for committing time and effort to improving our graduate community. I greatly appreciate and admire all of the student leaders I have worked with in NC State STARS. Together, we have provided a positive impact within and beyond the university community. I would like to thank Veronica Catete, Arpan Chakraborty, and Robinson Udechukwu for continuing this legacy of student leadership.

I would especially like to thank my family and friends for their encouragement over the years. Wisdom, knowledge, philosophical insight, strength of spirit and appreciation for creativity have all been gifts from time spent with family. My friends have provided a grounding in the world outside the dissertation, for which I am grateful. To my beloved, you have provided a haven from the demands of graduate research.

This research was supported in part by the Department of Computer Science along with the National Science Foundation through Grants DRL-1007962, IIS-0812291, CNS-0739216 and CNS-1042468. Any opinions, findings, and conclusions or recommendations expressed in this dissertation are those of the author and do not necessarily reflect the views of the National Science Foundation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

Chapter 1. Introduction

Cognitive and affective processes intertwine during learning, comprising a rich layer of emotional, or *affective*, experience. Affective states often influence progress on learning tasks, resulting in positive or negative cycles of affect that impact learning outcomes (Baker, D'Mello, Rodrigo, & Graesser, 2010; D'Mello, Craig, & Graesser, 2009; Woolf et al., 2009). Consequently, the influence of affect on learning has led to a recognized need to understand the occurrence, timing, and impact of cognitive-affective states (Baker et al., 2010; D'Mello et al., 2009; Grafsgaard, Boyer, & Lester, 2011, 2012; Woolf et al., 2009). Indeed, highly effective human tutors keep in mind students' affective and motivational states (Lepper & Woolverton, 2002). Thus, it is imperative to develop a clear understanding of how affective phenomena impact and co-occur with learning. The next generation of intelligent tutoring systems would necessarily detect and respond to students' affective states (D'Mello & Calvo, 2011).

Numerous studies have investigated learning-centered cognitive-affective states such as frustration, anxiety, boredom (or disengagement), confusion, delight, eureka (or a-ha moments), excitement, flow (or engaged concentration), and surprise. Each of these states contributes to the complex intermingling of cognitive and affective processes inherent in learning. For instance, frustration plays a central role in learning, possibly hindering it (Baker et al., 2010; Woolf et al., 2009). When students are unable to surmount difficulties during learning tasks, they may remain in a "state of stuck" (Baker et al., 2010; Kapoor, Burleson, & Picard, 2007; McQuiggan, Lee, & Lester, 2007). Similarly, if students are unable to reconcile confusion induced by new concepts, they may transition to frustration (D'Mello et al., 2009). Thus, automated approaches that recognize and understand frustration and other affective states are vital to designing affective tutorial interventions that foster learning through affective and motivational support.

1.1. Research Question and Hypotheses

This dissertation addresses the research question:

How can automated approaches be used to identify multimodal sequences of positive or negative affect during tutoring?

Toward this inquiry, analyses were conducted to examine nonverbal behaviors related to cognitive-affective states in tutoring, such as facial expression, gesture, and posture. These efforts continued in a natural progression toward increasingly sophisticated approaches that aimed to uncover more evidence regarding the relationship between nonverbal behavior and affect. This resulted in the formulation of the following hypotheses:

H1. Multimodal sequences in the tutoring corpus are significantly associated with the positive and negative affect that students report at the end of the tutoring session, both in terms of *engagement* and *frustration*.

H2. Multimodal sequences differ significantly across student groupings based on incoming characteristics, specifically *computer science self-efficacy* and *general self-efficacy*.

This research question and corresponding hypotheses are re-examined in Chapter 12.

1.2. Contributions

This dissertation work has added to the literature in affective computing, artificial intelligence in education, educational data mining, intelligent tutoring systems, and multimodal interaction. The novel contributions of this work include:

C1. **Facial Expression Annotation**: Facial expression was manually annotated to provide high quality information on facial expression and for comparison with automated techniques. Coders were trained in the standard research protocol

for facial expression annotation (the Facial Action Coding System) and inter-rater agreement was assessed.

C2. **Multimodal Data Collection**: A large human-human tutoring corpus was collected using a wide array of affect channels, including facial expression, posture, gesture, and skin conductance. This effort also involved synchronization of the data streams at millisecond precision.

C3. **Automated Facial Expression Recognition**: An automated facial expression recognition toolbox was validated against manual annotations and then applied to the large human-human tutoring corpus. The validation analysis presented a first-of-its-kind large-scale comparison of manual annotations and automated facial expression tracking output.

C4. **Novel Posture Tracking & Gesture Detection Algorithms**: New techniques were developed to allow automated tracking of posture and gesture from Kinect depth recordings in the tutoring corpus. Accuracy of these algorithms was compared against manual labels.

C5. **Statistical Analyses of Nonverbal Behavior in Tutoring**: The associations among nonverbal behavior, affect, and learning were explored through multiple analyses. These analyses provided insight into key affective states, such as confusion, engagement, and frustration, as well as tutoring phenomena including self-efficacy and learning.

C6. **Descriptive Hidden Markov Models of Affective Tutorial Interaction**: A preliminary approach toward sequential modeling of limited data streams was developed. This effort identified salient patterns of student nonverbal behavior, tutor-student collaboration, and student task progress. The resulting hidden Markov models were particularly effective at describing these patterns of interaction derived from facial expression, dialogue, and task progress data streams.

C7. **Multimodal Affect Representation and Feature Comparison**: Nonverbal behaviors are displayed across multiple data streams, including posture, gesture, and facial expression. With a large, automated dataset of nonverbal behavior, it is possible to consider a variety of features. Regression models were constructed using a conservative approach to identify the multimodal features most associated with tutorial outcomes (engagement, frustration, learning gain). The feature sets developed in this effort formed the basis for subsequent sequential analyses.

C8. **Multimodal Differential Sequence Mining**: A novel implementation of differential sequence mining extended the technique to multimodal data streams. This multimodal differential sequence mining approach enables discovery of a small set of differentially significant patterns across a large set of observations. This approach enables automated identification of multimodal event sequences that vary across two groups. The implementation of multimodal differential sequence mining will be publicly released with a future publication.

C9. **Comparison of Multimodal Sequences**: Multimodal differential sequence mining was used to examine student differences in tutorial outcomes and incoming student characteristics. Two approaches were used for comparison: 1) analyzing sequential patterns related to tutorial outcomes (engagement, frustration, and learning gain), and 2) identifying sequential patterns distinguishing between incoming characteristics, namely general self-efficacy and computer science self-efficacy. These sequential analyses identified patterns of student nonverbal behavior and tutor-student interaction that provided much-needed empirical evidence for the role of nonverbal behavior and affect in tutoring.

1.3. Motivation and Organization

This dissertation explores the fundamental human experience of emotion within the context of tutoring. Effective human tutors attend to a student's feelings and motivation in addition to cognitive performance. Thus, a comprehensive approach to tutoring necessarily considers nonverbal expressions of emotion. Accordingly, this line of research investigates posture, gesture, and facial expression as indicators of student cognitive and affective states.

The dissertation is comprised of twelve chapters. Chapter 2 provides background and related work. Chapter 3 presents the tutoring corpora that this research is based on and manual annotation of facial expression and dialogue. Chapter 4 describes the facial expression recognition tool used and validation of its output through comparison with manual labels. Chapter 5 presents analyses of affect and learning using a facial expression recognition toolbox. Chapter 6 describes posture and gesture tracking algorithms developed in the course of this research. Chapter 7 details analyses of embodied affect using the posture and gesture tracking algorithms. Chapter 8 presents results on sequential modeling of affect using hidden Markov models. Chapter 9 examines how multimodal features are predictive of tutoring outcomes, including affect and learning. Chapter 10 compares combinations of modalities to identify multimodal feature sets that are most predictive of tutorial outcomes. Chapter 11 presents multimodal differential sequence mining, a novel sequential analysis implementation that is robust to multimodal data streams. Chapter 12 concludes the dissertation with a summary.

Chapter 2. Background and Related Work

This dissertation primarily resides at the intersection of two research communities: Affective Computing and Artificial Intelligence in Education. This chapter highlights work that intersects both communities. Section 2.1 describes work on annotation, analysis, and detection of facial expression. Sections 2.2 and 2.3 describe work related to embodied affect, which include posture and gesture, respectively. Section 2.4 presents work related to affect in computer-mediated communication. Additionally, Section 2.6 presents current approaches to sequential analysis of learning-centered affect.

2.1. Facial Expression

Studies of facial expressions related to learning-centered cognitive-affective states can be categorized into one of three paradigms: 1) observation and annotation of affective behaviors; 2) investigation of facial action units involved in learning-centered affect; and 3) application of automated methods to detect affective states. These categories are relevant to this dissertation, so they are each discussed in turn.

2.1.1 Annotation of Affect from Facial Expression

The first category of studies involves observing and annotating affective behavior, and often represents a precursor to further analyses of learning-centered affect. Prior to applying automated methods to detect student affective states during interactions with Wayang Outpost (a mathematics intelligent tutoring system), Woolf et al. observed student behaviors including head nodding/leaning, postural movement, verbalization, and smiling, which supported further study of arousal and valence (Woolf et al., 2009). Afzal & Robinson studied affect in a naturalistic video corpus taken during self-study of tutorial materials and a complex mental task, multiple coders were used to label emotions (Afzal & Robinson, 2009). The coders identified confusion, happiness, interest, and surprise as the most frequent cognitive-affective states. Lastly, Baker et al.

have used classroom observations to identify and analyze moment-by-moment affect during student interactions with cognitive tutors and other educational software (Baker et al., 2010). Their observation protocol has been developed over several years, and involves viewing students through peripheral vision to interpret their posture, facial expression, gesture, speech, and eye gaze. The protocol has been applied to student populations throughout the world, and has provided key insight into student affective states and learning, such as the detrimental nature of boredom.

2.1.2 Analysis of Learning-Centered Affect and Facial Expression

The second category of studies involves investigating facial action units in learning-centered affect. These studies yield detailed data for designing affective tutoring systems. In a rich line of research, D'Mello and colleagues have compiled correlations of facial action units and self-reported and judged affect; for example, in a study of seven students' emote-alouds during interaction with AutoTutor, a natural language intelligent tutoring system, FACS coders annotated video at moments of students' emote-alouds (D'Mello et al., 2009). In the same work, multiple judges annotated affect from videos of twenty-eight students' tutoring sessions with AutoTutor. The FACS labels of both studies were compared, identifying correlations of AU1 (inner brow raising) and AU2 (outer brow raising) with frustration, and correlations of AU4 (brow lowering) and AU7 (eyelid tightening) with confusion.

2.1.3 Automated Detection of Affect from Facial Expression

There have been few studies in the third category, which focuses on automatically detecting facial expressions of learning-centered affect. Woolf et al. tracked cognitive-affective states of students interacting with Wayang Outpost using the MindReader tracking software (Arroyo et al., 2009). The MindReader system was trained on posed facial expressions and head movements of states such as interested or concentrating (el Kaliouby & Robinson, 2005). MindReader tracking of interest was found to improve

predictive models of student self-reported confidence and excitement, while tracked interest did, in fact, improve predictive models of student self-reported interest (Arroyo et al., 2009). In other work on automated detection, the authors of CERT have used it to track facial action units related to learning-centered affective states (Littlewort, Bartlett, Salamanca, & Reilly, 2011; Whitehill et al., 2011). For instance, CERT was used to track facial expressions of students interacting with a human tutor operating an iPad interface during cognitive game tasks in a Wizard-of-Oz design (Whitehill et al., 2011). Additionally, CERT has been used to track children's facial expressions during a cognitive task (Littlewort, Bartlett, et al., 2011). In both cases, CERT output was used as a relative comparison measure (i.e., the amount and type of facial movement before, during, and after performing a task). While this provides insight into facial expressions at meaningful moments, the cognitive tasks were simplified and may not have captured the full complexity of cognitive and affective processes involved during learning in an academic scenario.

2.2. Posture and Learning-Centered Affect

Nonverbal behaviors such as posture and gesture provide key channels signaling affective and motivational states. Posture has been used as an affective feature in multiple systems, but interpretation of postural movements is very complex (D'Mello, Dale, & Graesser, 2012; Kleinsmith & Bianchi-Berthouze, 2012). Early work focused on postural movement as a signal; for example, pressure-sensitive chairs have long been used for fine-grained measurement of posture (Kapoor et al., 2007; Woolf et al., 2009). Early studies of posture have indicated that the signal is involved in multiple cognitive-affective states, such as boredom, focus, and frustration (Kapoor et al., 2007; Woolf et al., 2009). Over the years, a replicated result in analyses of postural movement has arisen: increases in postural movement are linked with negative affect or disengagement (D'Mello et al., 2012; Rodrigo & Baker, 2011; Sanghvi et al., 2011; Woolf et al., 2009). There have also been recent developments in techniques for tracking

postural movement. Posture has been tracked in webcam video using computer vision (D'Mello et al., 2012; Sanghvi et al., 2011). These computer vision-based approaches have the advantage of directly identifying postural components such as body lean angle and slouch factor (Sanghvi et al., 2011) that were indirectly measured in the signals from pressure-sensitive chairs.

2.3. Gesture and Learning-Centered Affect

There is abundant cultural and anecdotal evidence for the importance of gestures (McNeill, 2005), yet empirical research results on the cognitive-affective states underlying gesture are sparse. A system trained on acted expressions of cognitive-affective states relied on combinations of facial expression and gesture features (el Kaliouby & Robinson, 2005), with meaning ascribed by human judges. Gestures have also been tangentially reported on in the intelligent tutoring systems community (Rodrigo & Baker, 2011; Woolf et al., 2009), but other phenomena were the primary focus of those studies. A recent study investigated different categories of hand-over-face gestures, with the researchers providing possible interpretations ranging over cognitive-affective states such as thinking, frustration, or boredom (Mahmoud & Robinson, 2011).

2.4. Computer-Mediated Communication and Affect

Tutoring may be conducted through computer-mediated communication, as is the case in this dissertation proposal. Thus, careful consideration of the communication medium is warranted. Most studies of computer-mediated communication have focused on the explicit act of communication itself (Derks, Fischer, & Bos, 2008; Walther, 1992). An example of this line of investigation is a system that automatically detects emotional expression from computer-mediated textual dialogue (Neviarouskaya, Prendinger, & Ishizuka, 2011). This system analyzes affective content of messages at the level of both

words and statements, and interprets emoticons and common expressions used in internet-based communication.

While studying textual communication of affect is useful, there also appears to be an implicit affective channel in computer-mediated communication. It may be necessary to investigate nonverbal affective phenomena in order to understand the human processes behind implicit affective interpretation. However, research into the relationship between nonverbal behavior and implicit interpretation of affect is scarce. A recent study examined instant messaging interactions while also recording the nonverbal behaviors unseen by the participants (Marcoccia & Atifi, 2008). The participants exhibited nonverbal behaviors indicative of cognitive and affective states (postural leaning, facial expressions, gestures), even though these bodily movements were not transmitted to the recipient of the textual dialogue. A limitation of that study is that it did not use surveys or self-reports to gauge affective experience of either participant.

From a theoretical perspective, the functions of nonverbal expression in computer-mediated textual dialogue differ significantly from those in face-to-face interaction. Nonverbal signals in general may express *affective/attitudinal states* (what a person feels), *manipulators* (interaction with objects in the environment, including self or others), *emblems* (culture-specific signals), *illustrators* (accompanying or depicting spoken concepts) or *regulators* (signals to control flow of conversation) (Pantic, Pentland, Nijholt, & Huang, 2006). In textual dialogue, the bodily expressions of *emblems*, *illustrators*, and *regulators* are rare or absent (Derks et al., 2008; Kiesler, Siegel, & McGuire, 1984; Marcoccia & Atifi, 2008; Walther, 1992). However, textual substitutes for these bodily expressions may be present (e.g., emoticons) (Neviarouskaya, Prendinger, & Ishizuka, 2007). The rarity of *emblems*, *illustrators* and *regulators* aside, nonverbal behavior of participants in computer-mediated textual dialogue contains expressions of *affective/attitudinal states* and *manipulators*

(Marcoccia & Atifi, 2008). In the present line of research, bodily expressions of posture display *affective/attitudinal states* and hand-to-face gestures are *manipulators*.

2.5. Sequential Analysis

Sequential analysis techniques aim to identify salient patterns in events that occur in series. While there are similarities across sequential analysis techniques, we may consider them within the following rough categories: statistical approaches, model-based approaches, and sequential pattern mining.

Statistical techniques for sequential analysis have existed for decades (Bakeman & Gottman, 1997). An example is lag sequential analysis, which may be used to identify frequent sequences of events that have some degrees of separation. For example, this technique was used to identify patterns of behavior and dialogue during marital discussions (Gottman, Markman, & Notarius, 1977). Due to sensitivity to random noise, lag sequential analyses were set aside for more recent log-linear approaches, which use chi-square statistics to better analyze the data (Bakeman & Quera, 2011). These techniques are especially effective when combined with reliable annotations of behavior and are often applied to dyadic interaction (Kenny, Kashy, & Cook, 2006).

Model-based techniques have an underlying foundation in viewing observed events as outputs of a system. One example is hidden Markov modeling, which uses an *initial probability distribution* across hidden states, *transition probabilities* among hidden states, and *emission probabilities* for each hidden state and observation symbol pair (Rabiner, 1989). Since the probabilities of emitting observations varies upon transition to a new hidden state, this technique provides a richer representation than the statistical techniques described above. However, a complication in this method is choosing parameters of the model, such as the number of hidden states.

The final category, sequential pattern mining, encompasses algorithmic approaches that are targeted toward generic sequential domains. These techniques have the advantage of robustly handling a wide variety of data that may have high

dimensionality (Fournier-Viger, Gomariz, Soltani, Lam, & Gueniche, 2014). An example of this is the Fournier-Viger algorithm, which accommodates constraints on intervals between events in a sequence (lag) and sequence constraints, such as requiring events to occur consecutively (with or without lag) and specifying a range of pattern lengths (Fournier-Viger, Nkambou, & Nguifo, 2008).

2.6. Sequential Affect Modeling in Tutoring

In intelligent tutoring research, sequential analyses have been used to address the issue of analyzing learning-centered affective states over time. The approaches used have evolved from direct analyses of conditional probabilities to more complex statistical and algorithmic techniques.

Initial work focused on identifying which learning-centered affective states were most likely to occur next given a particular affective state. This has been termed the *L* metric, which provides a value of how likely a transition from one affective state to another is, while correcting for the base probability of transitioning to that state. This was originally used to identify which student self-reports of affect were most likely to occur in interactions with AutoTutor (D'Mello & Graesser, 2012; D'Mello, Taylor, & Graesser, 2007). In follow-up work, Baker and colleagues applied the *L* metric to classroom observations of student affect (Baker et al., 2010; Baker, Rodrigo, & Xolocotzin, 2007). Additionally, the *L* metric has been used to identify frequent student self-reports in interactions with narrative-centered learning environments (McQuiggan, Robison, & Lester, 2010). While this metric provides information about paired transition probabilities of affective states, it cannot reveal how subsequences of affect relate to temporally situated tutoring events.

Motif discovery is an algorithmic approach to identify frequent subsequences in time series (Shanabrook, Cooper, Woolf, & Arroyo, 2010). In order to use motif discovery, substrings of a chosen length are produced and overlapping events are tabulated. The most frequently overlapping pairs of event symbols are compared using

hamming distance, which measures the number of changes needed to make the symbol sequences identical. If the hamming distance for a pair of substrings is within the selected distance threshold, they are added as a motif (up to a selected number of motifs). Thus, motif discovery allows researchers to identify frequently recurring sequences across a corpus. This has been applied to interactions with the Wayang Outpost intelligent tutoring system, revealing subsequences associated with frustration, on-task behavior, and gaming the system.

Differential sequence mining is an approach that has been used to identify repeated sequences in tutoring event time series (Kinnebrew, Loretz, & Biswas, 2013). This approach considers time series across students, identifying the most frequent sequences across the corpus and for each individual student. This focus on student sequences allows for comparison between two groups of students. Statistical comparison is used to identify the most significant differences in the subsequences across the groups of students. This approach has identified subsequences of tutorial interaction related to productive or unproductive work in the Betty's Brain tutoring system, but it was not applied to analysis of affect.

2.7. Engagement and Frustration

Engagement and frustration are both important cognitive-affective phenomena that may have impacts on learning (Pekrun, 2006). They have often been considered to be at opposite sides of the affective spectrum, with engagement positive and frustration negative. However, the story is not that straightforward, as there are specific theoretical considerations that should be discussed for both.

Engagement may be colloquially discussed as a monolithic phenomenon, but there are multiple ways in which it manifests (Pekrun & Linnenbrink-Garcia, 2012). There are low-level processes of attention and memory that may be considered *cognitive engagement*. From another standpoint, we may observe someone involved in an activity (without regard to internal processes), which is *behavioral engagement*.

When both activity and internal processes are considered, we may be interested in an individual's strategy and goals (i.e., the higher-level processes) as *cognitive-behavioral engagement*. In interactions between individuals, we may want to focus on the exchange of information and quality of collaboration as *social-behavioral engagement*. Finally, if we are interested in an individual's ability to persist toward a goal or degree of interest in an activity, then we may consider *motivational engagement*. In the analyses conducted in the course of this dissertation, the focus is on cognitive-behavioral engagement and motivational engagement. Students who were engaged in computer-mediated tutoring tended to display interest and focus during the task, but this did not appear to correspond with greater social interaction (dialogue). Additionally, all students were working on the task (behavioral engagement), so the differences we consider may be more aligned along the dimensions of cognitive-behavioral and motivational engagement.

Frustration, on the other hand, is generally regarded to be on the opposite end of the cognitive-affective spectrum (Baker et al., 2010; D'Mello, Lehman, Pekrun, & Graesser, 2014). This negative state may arise when encountering problems that are above one's current capabilities. In the case of a student, this may result in a "vicious cycle" of negative disposition toward the task. However, there has also been the recent proposal of *pleasurable frustration*, wherein exposure to a difficult task may actually result in joy and a sense of accomplishment (Gee, 2004). These seemingly opposite forms of frustration are not actually at odds with each other, as there is a significant division in the context and timing of their occurrence. As described in the theory of cognitive disequilibrium, a student may encounter new information and be confused by it (Graesser & Olde, 2003). If this confusion is not resolved, the student may transition to frustration with the learning task. In the case of a great video game, a player may persist through an extremely difficult level and prevail in the end (Gee, 2004). The differing outcomes of these cases are related to the level of difficulty of the task and the capability of the individual. In one, the learning content was beyond the student's

current level of capability and knowledge, while in the other case, the challenge was difficult, but surmountable. This discrepancy is the underlying intuition for the concept of the *zone of proximal development* (Vygotsky, 1978). The difficulty of the task ought to be suited to the capabilities of the learner and the teacher, as there is an optimal zone within which one can learn.

Chapter 3. Corpus Collection and Annotation

This chapter describes the JavaTutor Study I and JavaTutor Study II corpora and manual annotations of the webcam recordings and dialogue in the JavaTutor Study I corpus. The facial videos were annotated manually using the Facial Action Coding System (FACS), which enumerates the possible movements of the face through a set of facial *action units* (Ekman, Friesen, & Hager, 2002a). The dialogues were annotated using a symmetric student and tutor task-oriented dialogue act protocol (Boyer et al., 2010). Sections 3.1 and 3.2 describe the JavaTutor Study I corpus and JavaTutor Study II corpus, respectively. Section 3.3 describes annotation of fourteen tutoring sessions with one facial action unit that corresponds to a brow lowering movement, while Section 3.4 presents an extended annotation for upper and lower face movements across seven tutoring sessions.

3.1. JavaTutor Study I: First-year Java programmers

The JavaTutor Study I corpus was collected during a tutorial dialogue study (Boyer et al., 2009). Students solved an introductory computer programming problem and engaged in computer-mediated textual dialogue with a human tutor. The corpus consists of 48 dialogues annotated with dialogue acts (detailed in Section 3.3). Annotations also include information about student progress on the programming task (Boyer et al., 2010). The tutoring sessions ranged in duration from thirty minutes to over an hour.

3.2. JavaTutor Study II: Novice programmers

The JavaTutor Study II corpus consists of computer-mediated tutorial dialogue for computational concepts. Students (N=67) and tutors interacted through a web-based interface that provided learning tasks, an interface for computer programming, and textual dialogue. The participants were university students in the United States, with average age of 18.5 years (*stdev*=1.5). The students voluntarily participated for course

credit in an introductory engineering course, but no prior computer science knowledge was assumed or required. Substantial self-reported prior programming experience was an exclusion criterion. Each student was paired with a tutor for a total of six sessions on different days, each session limited to forty minutes. Recordings of the sessions included database logs, webcam video, skin conductance, and Kinect depth video. The JavaTutor interface is shown in Figure 3.1.

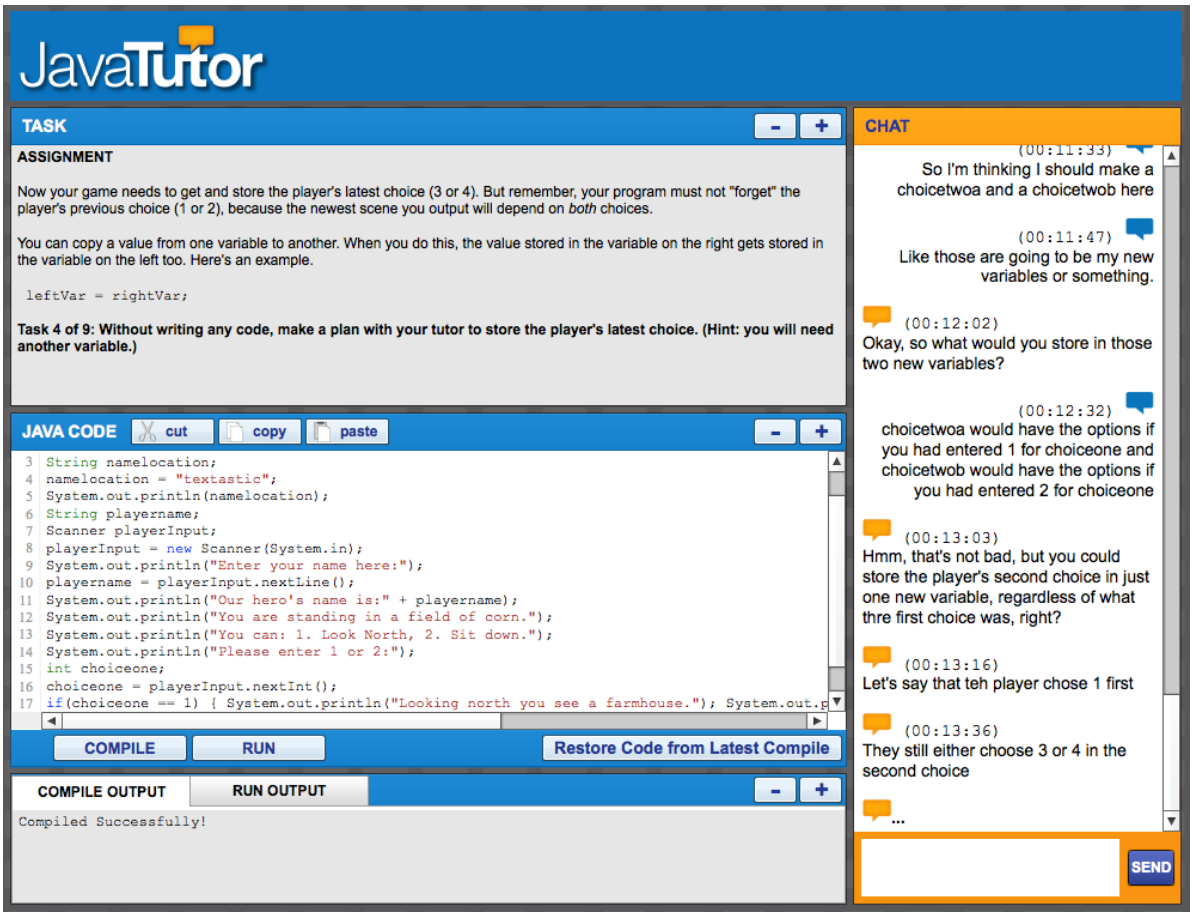


Figure 3.1. The JavaTutor interface

The recording configuration consisted of a webcam, Kinect depth sensor, and skin conductance bracelet (Figure 3.2).

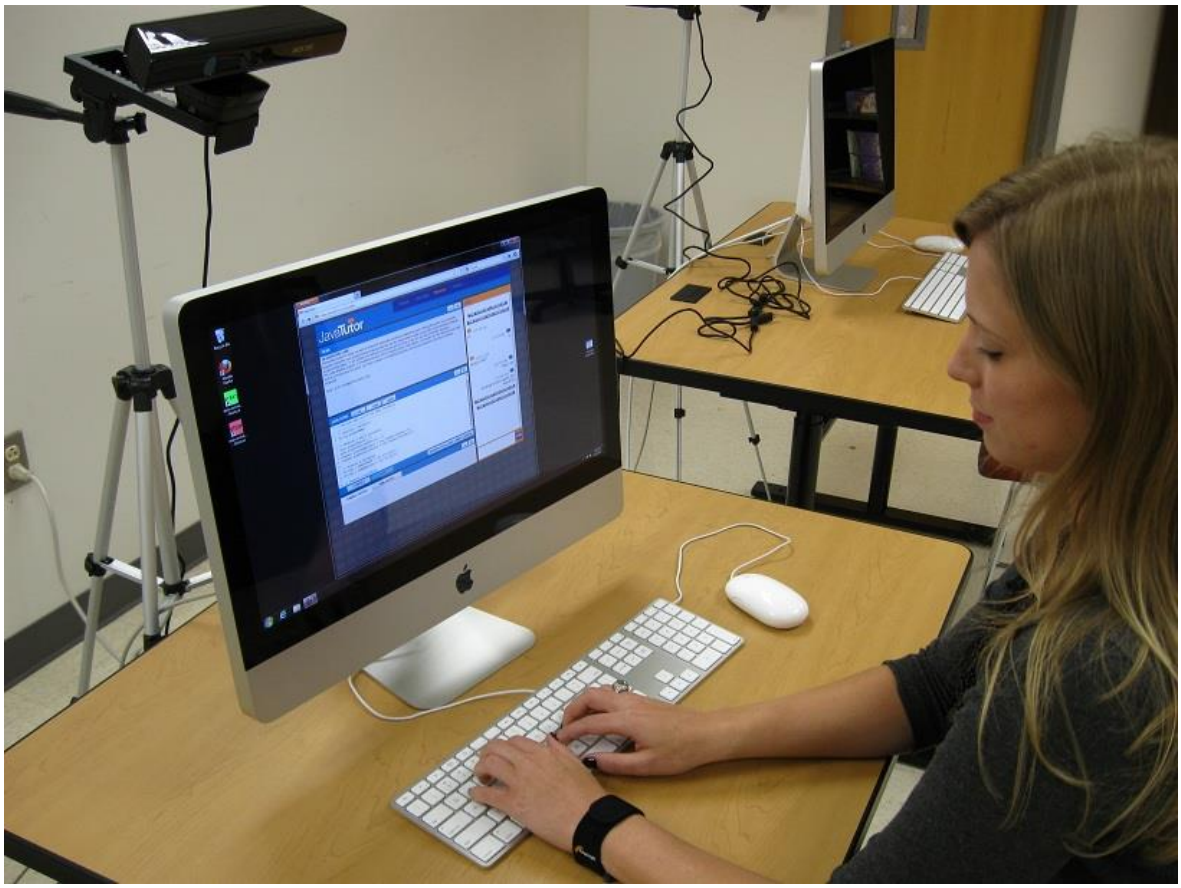


Figure 3.2. Student workstation with Kinect depth sensor, skin conductance bracelet, and computer with webcam

The study coordinators started the recordings at the beginning of each tutoring session. Thus, the students were aware of the recordings. However, once started, the recording windows were automatically hidden, so the students did not see themselves during the tutoring sessions. Additionally, review of the video recordings further confirmed that students did not attend to the recording devices (webcam and Kinect),

which indicates that the recordings were unobtrusive. The tutoring video corpus is comprised of approximately four million webcam video frames totaling thirty-seven hours across the first tutoring lesson. Two session recordings were missing due to human error (N=65). The recordings were taken at 640x480 pixel resolution and thirty frames per second.

Before each session, students completed a content-based pretest. After each session, students answered a post-session survey and posttest (identical to the pretest). The post-session survey items were designed to measure several aspects of engagement and cognitive load. The survey consisted of a User Engagement Survey (UES) (O'Brien & Toms, 2010) with Focused Attention, Endurability, and Involvement subscales, and the NASA-TLX workload scale (Hart & Staveland, 1988), which consisted of response items for Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration Level. The UES subscales were each comprised of multiple Likert-style items, while each NASA-TLX item was self-reported on a scale from zero to one hundred. As noted by one of the authors of NASA-TLX in a retrospective survey of studies using the scale, individual response items can be used separately to identify specific dimensions of task workload (Hart, 2006). Thus, the analyses performed on the JavaTutor II corpus consider the response items, including the item for Frustration Level, independent of each other.

3.3. Brow Lowering: Action Unit 4

Facial recordings of students were collected using built-in webcams during JavaTutor Study I. The tutors were not shown the student facial videos. Video quality was ranked based on how completely each student's face was visible within the frame, and the fourteen highest quality videos were selected for analysis. They have a total running time of eleven hours 55 minutes and include dialogues with three female subjects and eleven male subjects.

The seven selected facial videos were manually annotated using the Facial Action Coding System (FACS) (Ekman et al., 2002a). The FACS coders viewed videos from start to finish, pausing at observed instances of action unit 4, which is a brow lowering movement (Figure 3.3). Facial movements were encoded as events with a start frame and an end frame. One certified FACS coder annotated all fourteen videos from start to finish, pausing at all observed instances of AU4. A second certified FACS coder annotated a subset of six videos. After the tagging was complete, the sessions were discretized into one-second intervals. Cohen's kappa for inter-coder agreement on AU4 across all one-second intervals was κ=0.86, which indicates very good reliability. Displays of AU4 were noted during a total of 53 minutes of the approximately 12 hours of video, with high variance across individual students (*min*=0 seconds; *max*=33 minutes).

Figure 3.3. Student AU4 facial expressions

The JavaTutor Study I corpus consists of 48 dialogues, which were annotated with dialogue acts. Table 3.1 shows the dialogue act frequencies in the subset of fourteen tutoring sessions annotated with brow lowering (AU4). Annotations also include information about student progress on the programming task (Boyer et al., 2010).

Table 3.1. Dialogue act tags and frequency across the fourteen sessions (*S* = student, *T* = tutor)

Act	Description	*S*	*T*
ASSESSING QUESTION	Task-specific query or feedback request	44	83
EXTRA DOMAIN	Unrelated to task	37	42
GROUNDING	Acknowledgement, thanks, greetings, etc.	57	38
LUKEWARM CONTENT FDBK	Partly positive/negative elaborated feedback	2	23
LUKEWARM FEEDBACK	Partly positive/negative task feedback	3	21
NEGATIVE CONTENT FDBK	Negative elaborated feedback	5	77
NEGATIVE FEEDBACK	Negative task feedback	10	10
POSITIVE CONTENT FDBK	Positive elaborated feedback	10	21
POSITIVE FEEDBACK	Positive task feedback	23	119
QUESTION	Conceptual or other query	31	24
STATEMENT	Declaration of factual information	55	320

Annotated excerpts from the corpus are shown in Figure 3.4. Excerpt 1 shows an exchange where the tutor corrects the student's understanding of Java programming. The student displayed AU4 while correcting the program and while considering the tutor's question. Excerpt 2 has another instance of the tutor providing feedback and improving the student's knowledge. In this case, the student presents AU4 facial expressions while correcting multiple problems in the Java program.

Excerpt 1		
13:16:03	Tutor:	no, it's easier than that, you just have to make the middle if into an "else if" [NEGATIVE CONTENT FEEDBACK]
	Student:	CORRECT TASK ACTION AU4
13:16:31	Tutor:	does that make sense? [ASSESSING QUESTION AU4]
13:16:41	Tutor:	that way it only checks the 2nd conditional if the first one failed [STATEMENT]
13:17:20	Student:	it makes sense now that you explained it […] [POSITIVE CONTENT FEEDBACK]
Excerpt 2		
14:52:18	Tutor:	no, before we start sorting [NEGATIVE CONTENT FEEDBACK AU4]
	Student:	CORRECT TASK ACTION AU4
14:52:27	Tutor:	so, before the first loop you can use i for this loop counter if you want to [STATEMENT AU4]
	Student:	MIXED PROGRESS TASK ACTION AU4
14:53:52	Student:	i try to keep them different so i don't confuse myself [STATEMENT]

Figure 3.4. Excerpts from the AU4-annotated JavaTutor Study I corpus (typographical errors appear verbatim from the corpus)

Figure 3.5 shows the frequencies of AU4 corresponding to tutor and student dialogue acts. For tutor dialogue acts, an instance of AU4 is considered "corresponding" if it occurs within ten seconds after the tutor move; for student acts, ten seconds before the student move. These durations were empirically determined to account for student preparation of an utterance and reception of a tutor utterance. Of student utterances, LUKEWARM CONTENT FEEDBACK corresponds to the highest probability of student AU4. In this dialogue move students articulate partially correct knowledge. Of tutor utterances, the most likely to correspond to student AU4 is NEGATIVE FEEDBACK, in which the tutor states that the student has made a mistake but does not provide an explanation. Student ASSESSING QUESTION, which constitutes a direct request for task-based feedback, also had a relatively high probability of AU4. Students generally make these requests when their confidence in a recent task action is low.

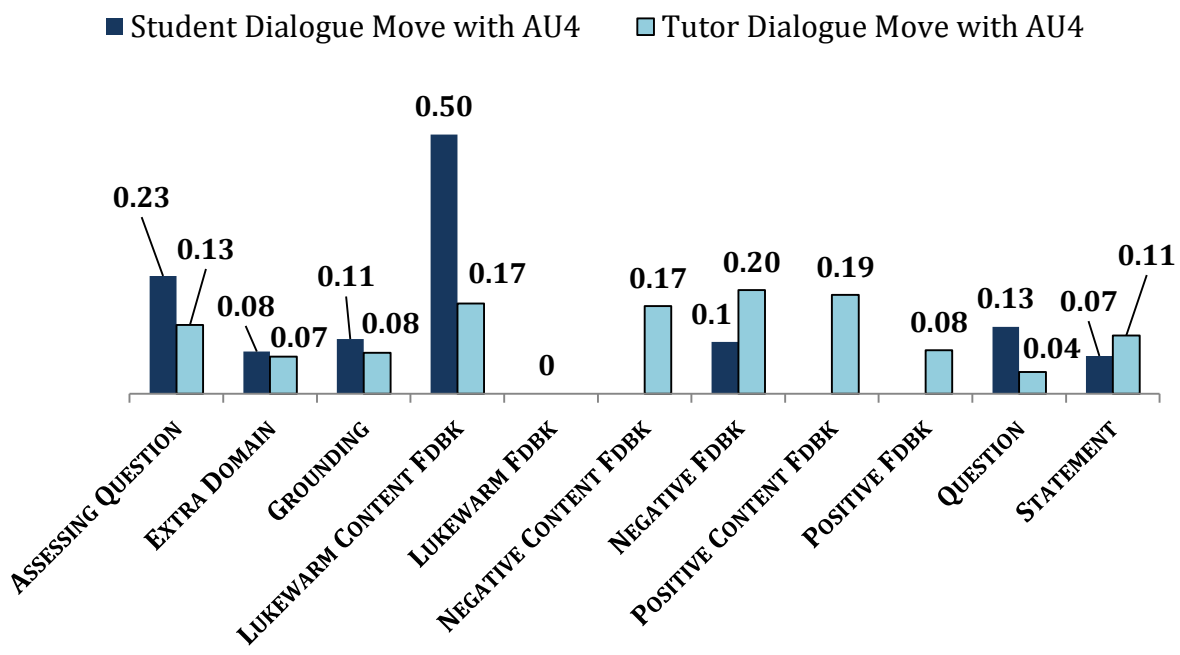


Figure 3.5. Relative frequency of student and tutor dialogue moves with AU4

Task actions were labeled based on progress toward a correct solution to the programming problem at hand, at a between-dialogue-moves granularity. Each task action cluster was characterized as CORRECT, INCOMPLETE, INCORRECT, or MIXED PROGRESS (a mixture of correct, incomplete, and/or incorrect task actions). As shown in Figure 3.6, students most frequently displayed AU4 during episodes of MIXED PROGRESS (37% of the time). Students were less likely (24%) to display AU4 during episodes of INCORRECT task action. As novices, these students were likely unaware of their mistakes when undertaking a completely incorrect task action. On the other hand, partially correct and partially incorrect task episodes indicate the student had sufficient knowledge to recognize that errors were present, and may have been experiencing constructive confusion toward reaching increased understanding.

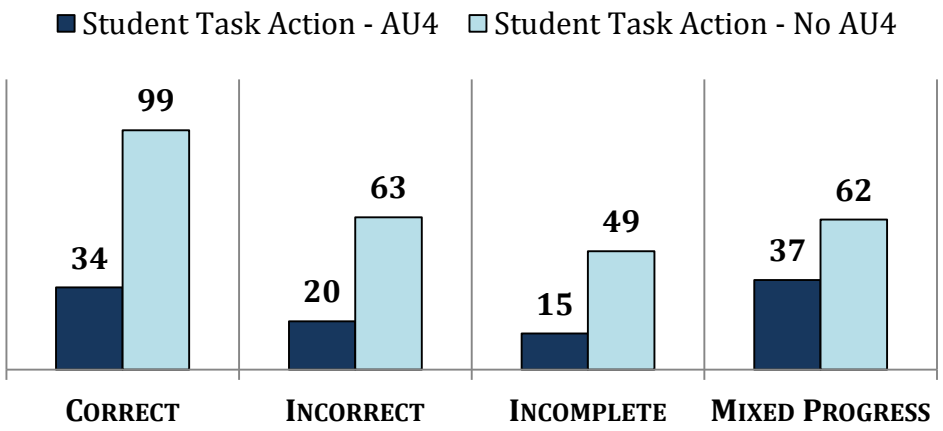


Figure 3.6. Frequency of student task actions with AU4 present or absent

3.4. Upper and Lower Face: 16 Action Units

In a follow-up round of manual annotation, the seven highest quality facial videos were selected for the extent to which the student's entire face was visible during a recording, and for near-even split across genders and tutors. These videos were annotated with sixteen facial action units that include common upper and lower face movements (selected examples are shown in Figure 3.7). These facial expressions may also be interpreted based on prior literature: AU1+2 or "surprise" (left), AU14+17 or "doubt" (center), and AU4+12 or "confusion and frustration" (right). Arrows indicate facial movements.

Figure 3.7. Examples of facial action units across the upper and lower face

The seven selected facial videos were manually annotated using the Facial Action Coding System (FACS). Two certified FACS coders viewed entire videos, encoding facial events of one or more AUs with a start and end frame. Some FACS AUs were excluded due to excessive burden in manual FACS coding (e.g., mouth opening, blinking) or anticipated rarity (e.g., lip pucker, lip funneler). Sixteen were selected for coding: AUs 1, 2, 4-7, 9, 10, 12, 14-17, 20, 23, 24, and 31.

In the first phase of the condensed FACS protocol, the two certified FACS coders independently annotated occurrences of AUs. The coders met in a second phase to produce a combined set of facial event instances without discussing specific AUs, during which event instances were merged or eliminated. By the end of the second phase, the coders agreed completely upon the start and end time of facial events (without discussing specific AUs). In the third phase, one of the coders reviewed where the facial events occurred and decided on precisely which AUs occurred. Finally, the second coder annotated 9.3% of the facial events independently, establishing an agreement average of Cohen's κ=0.67, comparable with similar studies (D'Mello, Lehman, & Person, 2010). Table 3.2 shows the dialogue act frequencies across the seven sessions.

Table 3.2. Dialogue act tags and frequency across the seven sessions (*S* = student, *T* = tutor)

Act	Description	*S*	*T*
ASSESSING QUESTION	Task-specific query or feedback request	16	29
EXTRA DOMAIN	Unrelated to task	20	26
GROUNDING	Acknowledgement, thanks, greetings, etc.	26	16
LUKEWARM FEEDBACK	Partly positive/negative task feedback	2	12
LUKEWARM CONTENT FDBK	Partly pos./neg. elaborated feedback	1	9
NEGATIVE FEEDBACK	Negative task feedback	5	5
NEGATIVE CONTENT FDBK	Negative elaborated feedback	1	34
POSITIVE FEEDBACK	Positive task feedback	10	76
POSITIVE CONTENT FDBK	Positive elaborated feedback	2	5
QUESTION	Conceptual or other query	13	9
STATEMENT	Declaration of factual information	18	143

Chapter 4. Facial Expression Recognition

While manual annotation of facial expression provides high quality data (as described in Chapter 3), it does not scale to large datasets. Thus, automated facial expression recognition is leveraged to analyze facial expressions in the JavaTutor Study II corpus. Section 4.1 describes the facial expression recognition toolbox used, while Section 4.2 considers the problem of comparing facial expression tracking output across individuals. Additionally, Section 4.3 presents a validation analysis using the manual facial expression annotations from the JavaTutor Study I corpus and automated facial expression tracking output.

4.1. Computer Expression Recognition Toolbox (CERT)

The Computer Expression Recognition Toolbox (CERT) (Littlewort, Whitehill, et al., 2011) was used in this research because it allows frame-by-frame tracking of a wide variety of facial *action units* (AUs). CERT finds faces in a video frame, identifies facial features for the nearest face using Gabor filters, and outputs weights for each tracked facial action unit using support vector machines (Wu et al., 2012). CERT has been previously used with both adults and children (Littlewort, Bartlett, et al., 2011; Littlewort, Whitehill, et al., 2011; Wu et al., 2012). A screenshot of CERT processing is shown in Figure 4.1.

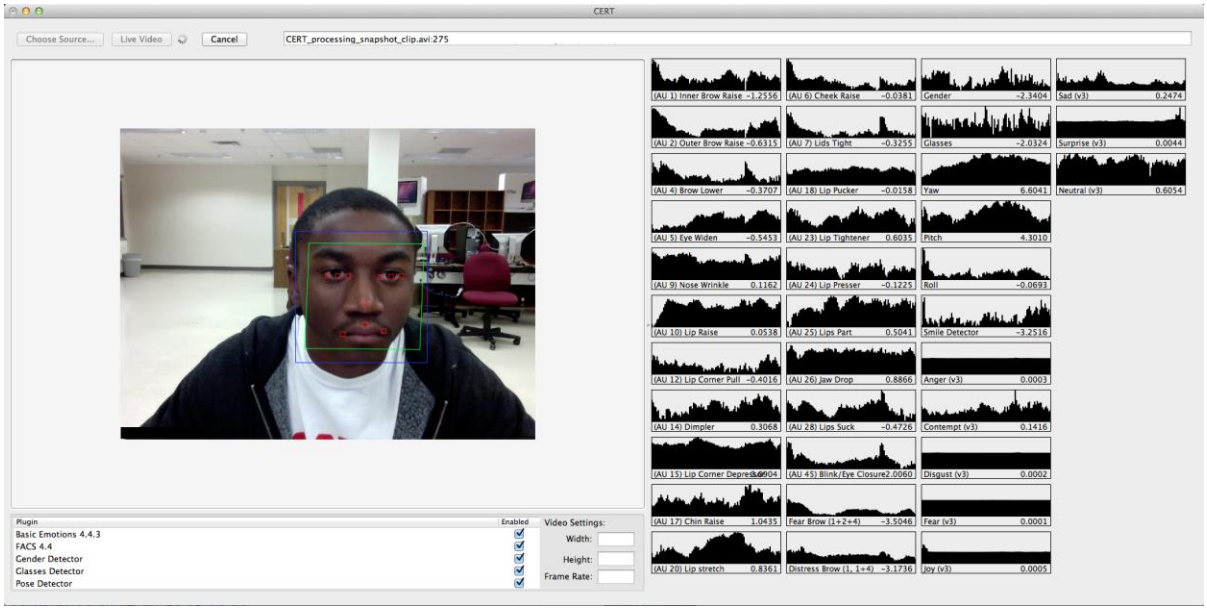


Figure 4.1. Screenshot of CERT video processing

Based on observations from manual annotation (Section 3.4), a subset of the 30 facial action units that CERT detects was selected as the focus of the present analyses. This set of facial action units was informed by a subset of the JavaTutor Study I corpus, used as a validation set (see Section 3.3), consisting of approximately 650,000 FACS-annotated video frames across seven tutoring sessions. In the validation set, sixteen facial action units were annotated. The five most frequently occurring action units each occurred in over 10% of the facial expression events. The remaining facial action units occurred substantially less frequently. The five frequently occurring action units were selected for further analyses on the JavaTutor Study II corpus.

Table 4.1 shows the relative frequency of each action unit's participation in discrete facial expression events and the number of frames annotated with each action unit from the validation corpus.

Table 4.1. The five most frequent facial action units in the validation corpus from JavaTutor Study I

Facial action unit	Frames	Event Freq.
AU1: Inner Brow Raiser	12,257	15.5%
AU2: Outer Brow Raiser	15,183	21.7%
AU4: Brow Lowerer	127,510	18.6%
AU7: Lid Tightener	9,474	13.2%
AU14: Dimpler	14,462	24.2%

The JavaTutor Study II corpus is comprised of approximately four million video frames totaling thirty-seven hours across the first tutoring session. Two session recordings were missing due to human error (*N*=65). CERT successfully tracked faces across a great majority of the tutoring video corpus (*mean*=83% of frames tracked, *median*=94%, *stdev*=23%).

4.2. Comparison of Facial Expression Tracking Across Individuals

In the course of processing videos with CERT, the range of output values noticeably varied between individuals due to their hair, complexion, or wearing eyeglasses or hats. This has also been noted by the creators of CERT (Wu et al., 2012). In order to better capture instances of facial expression displays in light of these individual tracking differences, output values were adjusted using the following procedure. First, the average output value for each student was computed for each facial *action unit* (AU).

These values correspond to individual baselines of facial expression. The average output value per session was subtracted for each action unit, resulting in individually adjusted CERT output. While any positive output value indicates that CERT recognizes an action unit, a threshold of 0.25 was empirically determined to reduce the potential for false positives. This threshold was based on observations of CERT output in which action unit instances that were more than slightly visible corresponded with output values above 0.25. CERT successfully tracked faces across a large majority of the validation corpus (*mean*=76% of frames tracked, *median*=87%, *stdev*=23%).

In order to interpret CERT output to indicate presence or absence of a facial action unit, a detection threshold of 0.25 was empirically determined. For instance, the average adjusted CERT output value for AU7 present in the validation corpus was 0.29, while the average for AU7 absent was -0.01. In comparison, the average "raw" CERT output (i.e., CERT output without adjustment) for AU7 present was 0.25, while the average for AU7 absent was 0.19. Thus, facial action unit displays may be compared across individuals using a combination of baseline adjustment of CERT output and an empirically-determined detection threshold.

Figure 4.2 provides detailed comparison of raw CERT output versus baseline-adjusted CERT output of AU7. Note that the raw CERT output indicates that AU14 (mouth dimpling) is present, though it is apparent to a certified FACS coder viewing the video that the action unit is not present. The adjusted CERT output correctly indicates that AU7 is present and AU14 is absent (based on the 0.25 threshold).

Raw CERT Output

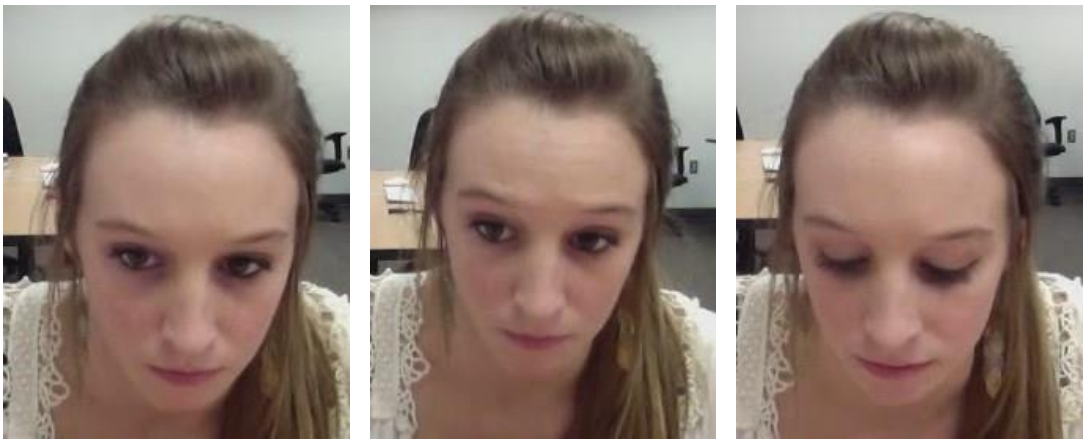
Action Unit	Value
AU1	-1.36
AU2	-0.67
AU4	0.05
AU7	**0.52**
AU14	**1.39**

Adjusted CERT Output

Action Unit	Value
AU1	-0.22
AU2	-0.27
AU4	0.11
AU7	**0.26**
AU14	-0.04

Figure 4.2. A comparison of AU7 with raw and baseline-adjusted CERT output

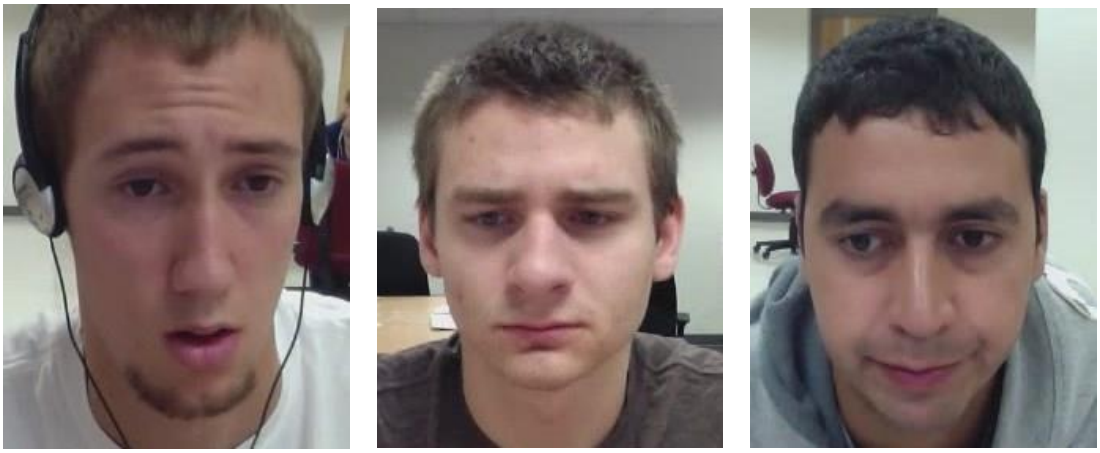
Figure 4.3 shows adjusted CERT output from an example of AU2, at three moments of the facial expression event: just before onset, apex (most intense video frame), and just after offset.



AU2 Onset: Outer brow raiser	**AU2 Apex:** Outer brow raiser	**AU2 Offset:** Outer brow raiser
AU1(-0.76) AU2(-0.21) AU4(-0.09) AU7(-0.24) AU14(0.13)	AU1(0.17) **AU2(0.27)** AU4(0.08) AU7(-0.09) AU14(-0.53)	AU1(-0.94) AU2(-0.21) AU4(0.12) AU7(0.09) AU14(0.00)

Figure 4.3. The onset, apex, and offset of an AU2 event with baseline-adjusted CERT output

Figure 4.4 shows adjusted CERT output for AUs 1, 4, and 14. The AU1 image (left) also includes AU4. In this case, AU1 brings the inner eyebrows upward, while AU4 pulls them inward. The other two examples, AU4 and AU14, each have a single facial movement detected through adjusted CERT output.



AU1 and AU4 Apex: Inner brow raiser and brow lowerer	**AU4 Apex:** Brow lowerer	**AU14 Apex:** Dimpler
AU1(0.80) AU2(-0.12) **AU4(0.25)** AU7(-0.23) AU14(-0.06)	AU1(-0.02) AU2(0.07) **AU4(0.47)** AU7(0.08) AU14(-0.85)	AU1(-0.02) AU2(0.00) AU4(-0.04) AU7(0.00) **AU14(0.46)**

Figure 4.4. Examples of AU1+AU4, AU4, and AU14 with baseline-adjusted CERT output

4.3. Validation Analysis

CERT was developed using thousands of posed and spontaneous facial expression examples of adults outside of the tutoring domain. However, naturalistic tutoring data often has special considerations, such as a diverse demographic, background noise within a classroom or school setting, no controls for participant clothing or hair, and facial occlusion from a wide array of hand-to-face gesture movements. Therefore, this

portion of the dissertation work aims to validate CERT's performance within the naturalistic tutoring domain. CERT's adjusted output was compared to manual annotations from the validation corpus.

The creators of CERT have applied the tool to the problem of understanding children's facial expressions during learning. To validate CERT's output, they compared it with manual FACS annotations across 200 video frames (Littlewort, Bartlett, et al., 2011). However, the goal in this analysis is to validate CERT's performance across a validation corpus of approximately 650,000 video frames. It is important to know whether average CERT output values for video frames with a specific facial *action unit* (AU) are different from those without that action unit. If the values are differentiable, then CERT may be an appropriate tool for general use at a large scale. If the values cannot be distinguished, then CERT is likely to provide many false positives and false negatives. Thus, this novel validation analysis provides needed insight into how well CERT performs across an entire corpus. The design of the validation analysis is shown in Figure 4.5.

Adjust CERT Output: Adjusting CERT output values with baselines for each individual allows for comparison across the entire corpus

↓

Binary Split on AU: Output values are divided between AU-present and AU-not-present and then averaged; these values serve as the input variable for logistic regression

↓

Build Predictive Model: A logistic regression model is built from all students' average values for AU-present and AU-not-present, producing predicted categories of AU/Not-AU

↓

Compare Predictions: Compute Cohen's kappa and accuracy from the logistic regression model predictions and manual tags to evaluate how well AU/Not-AU was predicted

Figure 4.5. Design of the validation analysis

Adjusted CERT output was computed for each video frame as described in Section 4.2. The CERT output values were then averaged within five binary splits, one for each facial action unit under consideration. Each binary split was comprised of frames with a specific facial action present and frames without that particular action unit, as labeled in the validation corpus. For example, to evaluate performance on brow lowering (AU4), video frames were divided between presence or absence of AU4 via the manual annotations. Once the binary split was performed, the frames were further subdivided by student. Thus, each student has an average value for frames with a specific action unit present and an average value for frames without that action unit. Logistic regression models were constructed using the average value as the sole parameter. One logistic regression model was built per action unit, for a total of five. The binary response variable categories (action unit present/action unit absent) were

produced from each regression model. The predicted categories were compared to the categories from manual annotation, yielding Cohen's κ and percent accuracy.

The validation results show that CERT output has an excellent capability to distinguish facial expression events from baseline across the validation corpus, yielding an average κ across the five action units of 0.82. Naturalistic data is challenging for computer vision techniques, so the validation analysis confirms the accuracy of CERT facial expression recognition. Table 4.2 displays the validation results.

Table 4.2. Comparison of agreement on validation corpus from JavaTutor Study I Manual FACS vs. logistic regression of CERT output

FACS Coder	AU1	AU2	AU4	AU7	AU14
Manual κ^*	0.88	0.82	0.79	0.78	0.73
CERT κ^*	0.86	0.86	0.68	1	0.71
CERT Accuracy*	93%	93%	85%	100%	86%

*Manual κ on face events; CERT evaluated on avg. output

In order to explore the effectiveness of the correction for individual differences, the validation analysis was performed again, this time without baseline-adjusted output values. With "raw" CERT output, the logistic regression models could not distinguish between the average values for AU-present versus AU-not-present (Table 4.3). Thus, agreement with the manual annotations was poor. The validation analyses illustrate that CERT output should be corrected with average values if a comparison across individuals is desired. This correction is straightforward to apply in post-processing. In a real-time application of such a tool, a running average could be computed at each video frame.

Table 4.3. Secondary validation analysis on raw CERT output

	AU1	**AU2**	**AU4**	**AU7**	**AU14**
CERT κ	0.14	0.29	0.05	0.29	0.29
CERT Accuracy	57%	64%	54%	64%	64%

A difficulty that remains for facial expression recognition is face occlusion, where the face is covered by an object, hand, etc. One source of face occlusions is hand-to-face gestures (Grafsgaard, Fulton, Boyer, Wiebe, & Lester, 2012), where one or two hands touch the lower face. These gestures are particularly prominent in the JavaTutor video corpora, as students often place a hand to their face while thinking or cradle their head in both hands while apparently tired or bored. These gestures can result in loss of face tracking or incorrect output. Accordingly, the analyses for this portion of the dissertation only considered video frames where face tracking and registration were successful (i.e., where CERT produced facial action unit output). Examples of both types of occlusion errors are shown in Figure 4.6. The CERT adjusted output values for the mostly occluded face frame (in the left image) are [AU1 = 1.34, AU2 = 0.65, AU4 = 0.62, AU7 = 0.30, AU14 = -1.17]. If these values are interpreted with the 0.25 threshold, then they represent presence of multiple action units, but that is clearly not the case when viewing the video. CERT was unable to find the student's face in the partially occluded frame (in the right image), though the presence of brow lowering is apparent. While hand-to-face gestures present a significant complication in naturalistic tutoring data, there has been preliminary progress toward automatically detecting these gestures (Section 6.2), so their effect may be mitigated in future facial expression tracking research.

Figure 4.6. Facial recognition errors due to gestures: mostly occluded (left) and partially occluded (right)

Chapter 5. Automated Facial Expression Analysis

Automated facial expression recognition enables fine-grained analyses of facial movements across an entire video corpus. With such tracking, there is potential to discover previously unidentified ways in which both frequency (D'Mello et al., 2009) and intensity (Littlewort, Bartlett, et al., 2011) of facial expressions inform diagnosis of student affective states. A first step toward this possibility is to quantify facial expressions as they occurred throughout tutoring and compare these with tutorial outcomes. Therefore, predictive models of both affective and learning outcomes were built leveraging both the average intensity and frequency of facial movements, as described in Section 5.1. Additionally, Section 5.2 presents a focused investigation of affect through the lens of learning gain.

5.1. Predicting Tutoring Outcomes from Facial Expression

Predictive models were constructed using minimum Bayesian Information Criterion (BIC) in forward stepwise linear regression, using JMP statistical software. These models are conservative in how they select predictive features because the explanatory value of added parameters must offset the BIC penalty for model complexity. Tutoring outcomes (affective and learning) were the dependent variables. Therefore, a model was constructed to predict each of the post-session survey scales and normalized learning gain (ten in total).

The predictive models of facial expression consider two features for each facial action unit: average intensity (average magnitude of CERT output values that were above the detection threshold) and relative frequency (percent of tracked frames that were above the detection threshold) of each facial action unit. These features were calculated across each tutoring session, resulting in ten feature values per student. The models for which facial action unit features were significantly explanatory are described below.

5.1.1 Predicting Affective Outcomes

Endurability was the student's self-report of whether he or she found the tutoring session to be worthwhile and whether he or she would recommend JavaTutor tutoring to others. Endurability was predicted by inner brow raising (AU1) intensity and brow lowering (AU4) intensity. AU1 was a positive predictor, while AU4 was negative. After adjusting for degrees of freedom (i.e., the number of model parameters), the model effect size was $r = 0.37$. The model is shown in Table 5.1.

Table 5.1. Stepwise linear regression model of Endurability

Endurability =	**Partial R^2**	**Model R^2**	***p***
-10.58 * *AU4_Intensity*	0.088	0.088	0.004
6.60 * *AU1_Intensity*	0.075	0.162	0.023
16.61 (intercept)			<0.001
RMSE = 10.01% of range in Endurability scale			

Temporal demand captures the student's self-report of whether he or she felt rushed or hurried during the session. Temporal demand was negatively predicted by outer brow raising (AU2) frequency; that is, students with higher frequency of this action unit reported feeling more rushed during the session. The adjusted model effect size was $r = 0.23$. The model is shown in Table 5.2.

Table 5.2. Stepwise linear regression model of Temporal Demand

Temporal Demand =	**Partial R^2**	**Model R^2**	***p***
-103.15 * *AU2_Freq*	0.068	0.068	0.037
34.90 (intercept)			<0.001
RMSE = 19.69% of range in Temporal Demand scale			

Performance was the student's self-report of how successful he or she felt in accomplishing the task. Performance was positively predicted by frequency of mouth dimpling (AU14), so students who displayed AU14 more frequently reported a higher sense of performance. The adjusted model effect size was $r = 0.26$. The model is shown in Table 5.3.

Table 5.3. Stepwise linear regression model of Performance

Performance =	**Partial R^2**	**Model R^2**	***p***
64.65 * *AU14_Freq*	0.081	0.081	0.022
72.74 (intercept)			<0.001
RMSE = 8.50% of range in Performance scale			

Frustration was the student's self-report of how insecure, agitated or upset he or she was during the tutoring session. Frustration was positively predicted by intensity of brow lowering (AU4); that is, students who displayed more intense AU4 reported feeling more insecure, agitated, or upset. The adjusted model effect size was $r = 0.29$. The model is shown in Table 5.4.

Table 5.4. Stepwise linear regression model of Frustration

Frustration =	**Partial R^2**	**Model R^2**	***p***
77.27 * *AU4_Intensity*	0.098	0.098	0.011
-15.34 (intercept)			0.165
RMSE = 17.05% of range in Frustration scale			

5.1.2 Predicting Learning Gain

An additional predictive model examined how facial movements predicted learning gains. Normalized learning gain was computed using the following formula if posttest score was greater than pretest score:

$$NLG = \frac{Posttest - Pretest}{1 - Pretest}$$

Otherwise, normalized learning gain was computed as follows:

$$NLG = \frac{Posttest - Pretest}{Pretest}$$

Thus, negative learning gains were possible, although 61 out of the 65 students had positive learning gain (*min*=-0.29, *max*=1.00). Normalized learning gain was predicted by outer brow raising (AU2) intensity and mouth dimpling (AU14) frequency. AU2 was a negative predictor and AU14 was a positive predictor; that is, lower AU2 intensity corresponded to lower learning gain, while greater AU14 frequency corresponded to higher learning gain. The adjusted model effect size was $r = 0.43$. The model is shown in Table 5.5.

Table 5.5. Stepwise linear regression model of Normalized Learning Gain

Norm. Learn Gain =	**Partial R^2**	**Model R^2**	***p***
-2.29 * *AU2_Intensity*	0.145	0.145	<0.001
2.13 * *AU14_Freq*	0.064	0.208	0.031
0.73 (intercept)			0.053
RMSE = 29.49% of range in Normalized Learning Gain			

5.1.3 Interpretation of Prediction Results

The results highlight that specific facial movements predict tutoring outcomes of engagement, frustration, and learning. Particular patterns emerged for almost all of the facial action units analyzed. Each of the results is discussed in turn along with the insight they provide into mechanisms of engagement, frustration, and learning as predicted by facial expression.

Average intensity of brow lowering (AU4) was associated with negative outcomes, such as increased frustration and reduced desire to attend future tutoring sessions. Brow lowering (AU4) has been correlated with confusion in prior research (D'Mello et al., 2009, 2014) and interpreted as a thoughtful state in other research (Grafsgaard et al., 2011; Littlewort, Bartlett, et al., 2011). Here, the average intensity of brow lowering is found to be a positive predictor of student frustration and a negative predictor of students finding the tutoring session worthwhile. It may be that the tutor and student were unable to overcome student confusion, resulting in frustration instead of deep learning (D'Mello et al., 2014). This interpretation is compatible with the theory of cognitive disequilibrium, which maps possible transitions from confusion to deep learning when a new concept is successfully acquired or to frustration when the concept cannot be reconciled with the student's present understanding. It is also possible that in some cases, AU4 displays represent an angry or agitated affective state.

AU4 is a key component of the prototypical display of anger (Ekman, Friesen, & Hager, 2002b). Further study that accounts for student progress through the programming task may reveal whether there is a significant cognitive aspect to this result.

Average intensity of inner brow raising (AU1) was positively associated with students finding the tutoring session worthwhile. At first glance, this finding seems to be in marked contrast to prior research that implicated both inner and outer brow raising as components of frustration displays (D'Mello et al., 2009). However, intensity of the facial expressions was not considered in the prior work. AU1 is also a component of prototypical expressions of surprise or sadness (Ekman et al., 2002b). From among these possible affective states—frustration, sadness, and surprise—surprise may be most likely to explain higher ratings of endurability. Students may have found the tutoring session to be surprising because it was a first exposure to computer programming. Surprise displays were observed while processing the videos through CERT and such displays were numerous in the validation corpus. However, further study is required to disambiguate this result.

Lower frequency of outer brow raising (AU2) predicted a lesser sense of being hurried or rushed; in contrast, greater intensity of displays of AU2 predicted reduced learning gains. Outer brow raising (AU2) has been associated with frustration in prior research (D'Mello et al., 2009). As frustrated students may not achieve high learning gains, the intensity of AU2 may be indicative of frustration. However, AU2 was not predictive of students' self-reported frustration levels, so this may be capturing a subtly different phenomenon. An alternative interpretation comes from research into facial expressions of anxiety, in which "fear brow" facial movements were found to occur more often during anxiety (Harrigan & O'Connell, 1996). The prototypical "fear brow" includes AU1, AU2, and AU4 present in combination (Ekman et al., 2002b). An example of this facial expression is shown as AU2 in Figure 4.3. Greater anxiety during tutoring may result in feeling rushed or hurried and may also negatively impact learning. Thus, anxiety is consistent with the results for AU2. However, the other action units expected

in facial expressions of anxiety, AU1 and AU4, did not have the same results. This is likely due to the conflicting nature of brow raising and brow lowering, as the CERT values for AU1 and AU4 may be reduced during their combined movement in the "fear brow" (see Figure 4.3). Further analyses of combined facial movements would provide insight into this complication of automated facial expression recognition.

Frequency of mouth dimpling (AU14) predicted increased student self-reports of task success, as well as increased learning gains. There have not been conclusive associations of mouth dimpling (AU14) and learning-centered emotions. However, this action unit has been implicated as being involved in expressions of frustration (D'Mello et al., 2009) and concentration (Littlewort, Bartlett, et al., 2011). In this study, frequency of AU14 was positively predictive of both self-reported performance and normalized learning gains. While the effect appears to be fairly subtle (effect size below 0.3 for both), it appears to be a display of concentration. This leads to the interesting question of whether AU4 or AU14 better represents a thoughtful, contemplative state. Further research in this vein may resolve the question.

While eyelid tightening (AU7) was not added to any of the predictive models, there appear to be reasons for this. Observation of CERT processing and the results of the validation analysis indicate a way to adjust CERT's output of AU7, enabling refined study of the action unit. AU7 is an important facial movement to include, as it has been correlated with confusion (D'Mello et al., 2009). A possible method for correcting AU7 output comes from the observation that CERT tends to confuse AU7 with blinking or eyelid closing. In prior manual annotation efforts, AU7 was labeled only when eyelid movements tightened the orbital region of the eye (as in the FACS manual). Thus, manual annotation seems more effective due to this complication of eye movements. However, note that CERT's AU7 output perfectly agreed with manual annotations in the validation analysis (Section 4.3). Thus, CERT clearly tracks eyelid movements well. The problem may be that CERT's AU7 output is overly sensitive to other eyelid movements. One way to mitigate this problem may be to subtract other eye-related movements

from instances of AU7. For instance, if AU7 is detected, but CERT also recognizes that the eyelids are closed, the detected AU7 event could be discarded.

The results demonstrated predictive value not only for frequency of facial movements, but also intensity. The relationship between facial expression intensity and learning-centered affect is unknown, but perhaps action unit intensity is indicative of higher-arousal internal affective states. Additionally, it is possible that intensity will inform disambiguation between learning-centered affective states that may involve similar action units (e.g., confusion/frustration and anxiety/frustration). Lastly, intensity of facial movements may be able to aid diagnosis of low arousal affective states. For instance, a model of low intensity facial movements may be predictive of boredom, which current facial expression models have difficulty identifying.

5.2. Facial Expression, Affect, and Learning

The analyses in this section build upon the predictive modeling results of Section 5.1 with a focused investigation of the JavaTutor Study II corpus through the lens of learning gain. First, frustration was the only post-session affect self-report that correlated with learning gain. Second, particular facial expressions that occurred throughout the sessions correlated with frustration and learning gain (Section 5.2.1). Third, facial expressions at the beginning and end of the sessions yielded distinct correlations with frustration and learning gain (Section 5.2.2).

Correlational analyses of the post-session affective survey scales and learning gains were conducted to identify potential relationships between affect and learning. Frustration was the only affect self-report to correlate with learning gain, with higher frustration corresponding to lower learning gain (Table 5.6).

Table 5.6. Correlations of Affective Post-Survey Scales and Normalized Learning Gain

Affect Survey Scale	*r*	*p*
ENGAGEMENT	0.14	0.288
MENTAL DEMAND	0.02	0.871
PHYSICAL DEMAND	-0.06	0.610
TEMPORAL DEMAND	0.03	0.830
PERFORMANCE	0.18	0.160
EFFORT	0.04	0.732
FRUSTRATION LEVEL	**-0.30**	**0.015**

5.2.1 Correlation of Facial Expression with Frustration and Learning

Because of the important relationship between frustration and learning, these analyses of facial expression are focused on frustration and learning gain. The analyses of facial expression consider two features for each facial action unit: average intensity (average magnitude of CERT output values that were above the detection threshold) and relative frequency (percent of tracked frames that were above the detection threshold) of each facial action unit. These features were calculated across each tutoring session, resulting in ten feature values per student. These correlations were first performed across entire tutoring sessions, and then across the first five minutes and last five minutes of the sessions. The beginning of a session may inform early prediction, while the end may more closely reflect self-reports due to temporal proximity. A statistical correction for multiple tests was applied, arriving at a Bonferroni *p*-value threshold of 0.0025 for each set of analyses of facial movements: entire session, beginning of session, and end of session. This more stringent threshold controls for the risk of false positives, and is intended to increase the generalizability of the findings. Results that were significant to this threshold are displayed in bold, and all other correlations with $p<0.05$ are shown.

The initial analyses correlated facial action units throughout the tutoring sessions with normalized learning gains and post-session self-reports of frustration (Table 5.7). Intensity of AU4 was positively correlated with frustration. Thus, greater intensity of AU4 corresponded with higher self-report of frustration. Both intensity and frequency of AU2 were negatively correlated with normalized learning gain, with only AU2 intensity significant after Bonferroni correction.

Table 5.7. Correlations of Frustration, Learning and facial Action Units throughout the tutoring session

Action Unit Variable	Tutoring Outcome	*r*	*p*
AU 4 Avg. Intensity	Frustration	0.31	0.011
AU 2 Avg. Intensity	**Norm. Learn. Gain**	**-0.38**	**0.002**
AU 2 Relative Freq.	Norm. Learn. Gain	-0.27	0.029

Analyses were conducted to examine whether facial expressions in the first five minutes of tutorial interaction were correlated with frustration and normalized learning gain. Four results emerged (Table 5.8): intensity of AU14 was positively correlated with frustration and frequency of AU2 was negatively correlated with normalized learning gain. Additionally, intensity of AU4 was positively correlated with frustration, and intensity of AU2 was negatively correlated with normalized learning gain. The correlations involving AU2 intensity and AU4 intensity retained their statistical significance after Bonferroni correction.

Table 5.8. Correlations of Frustration, Learning and facial Action Units in the first five minutes of tutoring

Action Unit Variable	Tutoring Outcome	*r*	*p*
AU 4 Avg. Intensity	**Frustration**	**0.41**	**0.001**
AU 14 Avg. Intensity	Frustration	0.32	0.010
AU 2 Avg. Intensity	**Norm. Learn. Gain**	**-0.38**	**0.002**
AU 2 Relative Freq.	Norm. Learn. Gain	-0.28	0.023

There was one significant result in the correlational analysis of frustration, learning, and facial action units in the last five minutes of tutoring. Relative frequency of AU14 was found to positively correlate with normalized learning gain (r=0.52, p<0.001). This result was significant after Bonferroni correction.

5.2.2 Early Prediction of Frustration and Learning Outcomes

Linear regression models were constructed to predict frustration and normalized learning gain from facial expressions in the first five minutes of tutoring. These models are intended to inform the use of facial expression features for early prediction of frustration and learning.

Both models were constructed using the significantly correlated facial action unit features from Table 5.8. The early prediction model of frustration is shown in Table 5.9. The model R^2 value corresponds to r=0.49, and the model effect is greater than either feature alone. The root mean squared error (RMSE) value indicates the overall magnitude of error. The RMSE of this model shows that it would not distinguish well between similar frustration levels, but would perform well at distinguishing between very high or low levels of frustration.

The model for early prediction of normalized learning gain is shown in Table 5.10. The model R^2 value corresponds to r=0.40. Thus, the model effect is similar to the

most explanatory feature, AU2 intensity. The significance values for the features also show that AU2 frequency did not significantly explain variance beyond AU2 intensity. The RMSE of this model is similar to that of the early prediction model of frustration. Very high or low values may be accurately distinguished, but it is likely to misidentify similar values.

Table 5.9. Early prediction model of Frustration

Frustration Level =	***p***
81.72 * AU4 Intensity	.002
38.62 * AU14 Intensity	.022
Intercept = -41.69	.002
RMSE = 21% of range in self-reports **Model R^2** = 0.24	

Table 5.10. Early prediction model of Normalized Learning Gain

Normalized Learning Gain =	***p***
-1.45 * AU2 Intensity	0.020
-0.46 * AU2 Relative Frequency	0.273
Intercept = 1.12	< 0.0001
RMSE = 24% of range in outcomes **Model R^2** = 0.16	

The early prediction results illustrate that facial expression at the beginning of tutoring sessions may provide a useful set of features for early diagnosis of affective states related to post-session outcomes. The models presented here are descriptive and are not designed to be used for intervention, but richer models may be constructed using machine learning techniques. Further models may incorporate timing of facial expression and learning task context to increase predictive accuracy.

5.2.3 Interpretation of Learning-Centric Analyses of Facial Expression

This dissertation presents results that demonstrate important relationships among frustration, learning and facial expression within a corpus of computer-mediated human-human tutoring. Two notable characteristics of the corpus facilitate interpretation of these findings. First, the corpus reflects few social effects on nonverbal behavior due to remote dialogue because the students and tutors did not see one another. Nonverbal behaviors that are common in face-to-face communication, such as *emblems* (e.g., thumbs-up gesture), *illustrators* (e.g., gesticulating to illustrate an idea during speech), and *regulators* (e.g., gesturing for a conversational participant to speak) were not displayed (Pantic et al., 2006). Second, the video recording of students was accomplished discreetly (e.g., not making noise or displaying a red light during recording). If the act of recording were obtrusive, students would likely become distracted and self-conscious, perhaps resulting in inhibition of facial expression. However, students did not attend to the recording devices during the tutoring sessions.

The analyses highlight ways in which intensity and frequency of facial action unit displays are associated with normalized learning gains and summative self-reports of frustration. Each facial action unit has been explored in prior research. Thus, the key results have a number of theoretical implications in light of past findings based on each facial action unit.

Both intensity and frequency of outer brow raising (AU2) were negatively correlated with normalized learning gain, based on AU2 displays at the beginning of tutoring sessions and throughout tutoring sessions. Based on prior literature, AU2 may be associated with frustration, surprise, or anxiety. AU2 has been identified as a component of frustration (along with AU1) in prior intelligent tutoring systems literature, with correlations of AU1 and AU2 displays and students' emote-aloud self-reports of frustration (D'Mello et al., 2009). Similarly to frustration, AU1 and AU2 together are components of the prototypical expression of surprise (Ekman et al.,

2002b). However, frustration and surprise do not seem consistent with the results of the correlational analyses because AU1 was absent from the significant correlations.

Anxiety has been linked to prototypical displays of fear (AU1+AU2+AU4+AU5+AU25; AU5 is eyelid opening, AU25 is mouth opening) (Harrigan & O'Connell, 1996). While this combination of AUs seems similar to those of frustration and surprise, the presence of AU4 introduces a conflicting movement of the brow that may impact detection of the expression. Figure 4.3 shows an example of AU1+AU2+AU4 (also shown are the moments before and after the facial expression—just before onset and after offset). The CERT output values from the apex of the facial expression show an interaction between AU1 and AU4. AU1 raises the inner eyebrows, while AU4 lowers the inner brow. The result of AU1+AU2+AU4 is tensing of the inner brow with creasing across the forehead, as is apparent in Figure 4.3 to a FACS coder. This conflict of facial movements at the inner brow may result in reduced CERT output values for both AU1 and AU4. This complication of CERT output may explain how only AU2 was negatively correlated with normalized learning gain. It also indicates that anxiety may be the most consistent interpretation of AU2.

Brow lowering (AU4) intensity at the beginning of sessions and throughout sessions was positively correlated with summative self-reports of frustration. AU4 has long been noted as an indicator of mental effort, notably mentioned by Darwin (Littlewort, Bartlett, et al., 2011). AU4 has also been correlated with self-reports and judgments of confusion in intelligent tutoring systems research (D'Mello et al., 2009). In this interpretation, AU4 at the beginning of sessions may have indicated effortful thinking and confusion. It is possible that such confusion may have gone unresolved, resulting in frustration.

Mouth dimpling (AU14) intensity at the beginning of sessions positively correlated with frustration, while AU14 frequency at the end of sessions positively correlated with normalized learning gain. Unilateral AU14 is a component of a prototypical expression of contempt (Matsumoto & Ekman, 2004). However, students

were observed to frequently display bilateral AU14 in the JavaTutor video corpora, as in Figure 4.4. Prior literature provides slight evidence of correlation between AU14 and frustration, with a statistical trend that AU14 occurred during student emote-aloud self-reports of frustration (D'Mello et al., 2009). In the same study, expert judges did not identify AU14 as an indicator of frustration. However, AU14 appeared as a frequent 'mouth fidgeting' movement in both JavaTutor corpora. Thus, AU14 may be easily overlooked by judges of emotion as noise, since it occurs frequently over time, similar to blinking or brow lowering. As an affective feature, AU14 does seem to be a facial indicator that repeatedly occurs over time and coincides with a thoughtful state, as suggested by a prior study (Littlewort, Bartlett, et al., 2011). The correlation between intensity of AU14 displays at the beginning of sessions and frustration may parallel the discussion of AU4 above. It is possible that effortful thought or confusion transitioned to frustration (Baker et al., 2010; D'Mello et al., 2009). Additionally, the correlation of AU14 at the end of the session with normalized learning gain suggests that students who were concentrating more at the end of the session tended to have increased learning gain.

The early prediction results illustrate that facial expression at the beginning of tutoring sessions may provide a useful set of features for early diagnosis of affective states related to post-session outcomes. While the modality of facial expression has yielded significant insights into frustration and learning, there are additional nonverbal behavior channels that may be considered. The next two chapters consider bodily expressions of affect, such as gesture and posture.

Chapter 6. Posture Tracking and Gesture Detection

Posture tracking and gesture detection algorithms were developed as part of this dissertation to enable automated analysis of the recorded depth images. These algorithms leverage regularities in the depth recordings in the JavaTutor Study II corpus (e.g., student in center, frontal view). This chapter describes the algorithms, their output, and evaluation. Section 6.1 presents the posture tracking algorithm, while Section 6.2 describes the hand-to-face gesture detection algorithm.

6.1. Posture Tracking Algorithm

A posture estimation algorithm was designed to compute posture for a given frame as a triple, (*headDepth*, *midTorsoDepth*, *lowerTorsoDepth*), as shown in Figure 6.1. Prior to applying the algorithm, extraneous background pixels were discarded using a distance threshold. An overview of the posture estimation algorithm is given in Algorithm 1 below. The algorithm computes bounding regions for head, mid torso and lower torso based on the height of the top depth pixel. Then, a single point is selected from each bounding region to estimate posture. For the head, the nearest pixel is selected. For the torso points, the farthest pixel in the bounding regions is selected, as the torso was often behind the desk and arms. Distances for each posture estimation point were normalized using standard deviations from the median position for each student workstation in order to account for different camera angles. This overall approach is robust to seated postures that are occluded by a desk, a distinct advantage over the alternative of Kinect skeletal tracking.

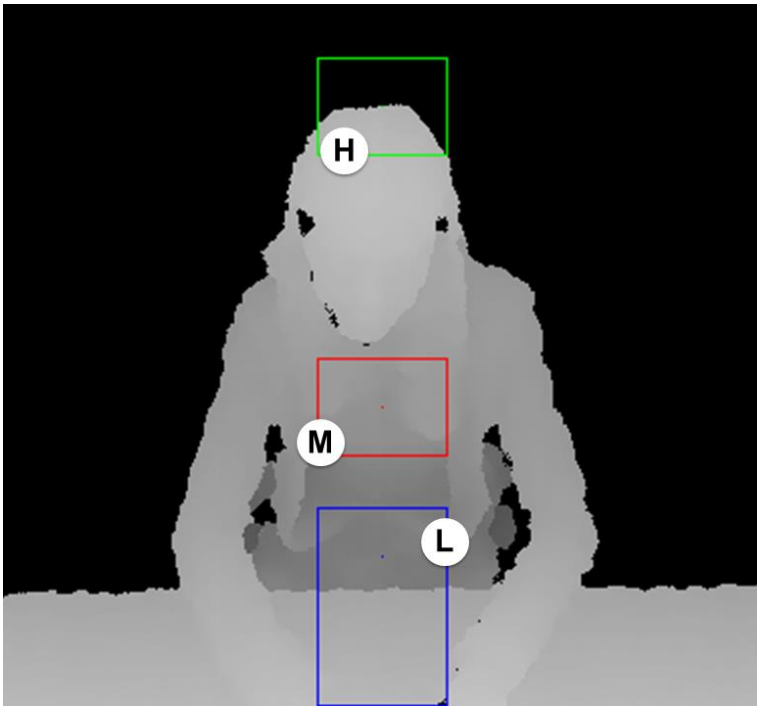


Figure 6.1. Automatically tracked posture points

Algorithm 1: POSTUREESTIMATION(*I*)

input : a depth image *I*
output : a triple of posture estimation points

1 *width* ← width of depth image *I*;
2 *height* ← height of depth image *I*;
3 *bottomRow* ← *height* – 1;
4 *center* ← *width* / 2;
5 *headRow* ← row of first depth pixel in center column;
6 *midRow* ← (*bottomRow* + *headRow*) / 2;
7 *lowRow* ← *midRow* + (*bottomRow* – *headRow*) / 2;
8 *sideBound* ← columns at ± (5% of *width*) from *center*;
9 *headBound* ← rows at ± (5% of *height*) from *headRow*;
10 *midBound* ← rows at ± (5% of *height*) from *midRow*;
11 *lowBottom* ← *lowRow* + (*bottomRow* – *headRow*) / 4;
12 *lowTop* ← *lowRow* – (5% of *height*);
13 *headDepth* ← closest pixel in [*sideBound*, *headBound*];
14 *midTorsoDepth* ← farthest pixel in [*sideBound*, *midBound*];
15 *lowerTorsoDepth* ← farthest pixel in [*sideBound*, *lowTop,* and *lowBottom*];
16 return (*headDepth*, *midTorsoDepth*, *lowerTorsoDepth*);

The output of the posture estimation algorithm was evaluated manually. The performance metric was the percent of frames in which the detected points (*headDepth*, *midTorsoDepth*, *lowerTorsoDepth*) coincided with the head, mid torso, and lower torso/waist. Two human judges individually examined images corresponding to one frame per minute of recorded video. The judges had moderate agreement on error instances with Cohen's κ=0.57. To provide a conservative measure of accuracy, the algorithm output was classified as erroneous if either judge found that any of the posture tracking points did not coincide with the target region (i.e., union of errors). Thus, the resulting accuracy was 92.4% over 1,175 depth images. Error conditions occurred primarily when students shifted their head or torso out of frame or covered their torso or waist with their arms and hands.

6.2. Gesture Detection Algorithm

A second algorithm was developed to detect hand-to-face gestures, which have been shown to co-occur with cognitive-affective states (Mahmoud & Robinson, 2011). The algorithm uses surface propagation to avoid issues of occlusion and hand deformation that pose problems for standard hand tracking techniques. Two variants of hand-to-face gestures were detected: one hand to the student's face and two hands to the student's face. Examples of detected hand-to-face gestures are shown in Figure 6.2.

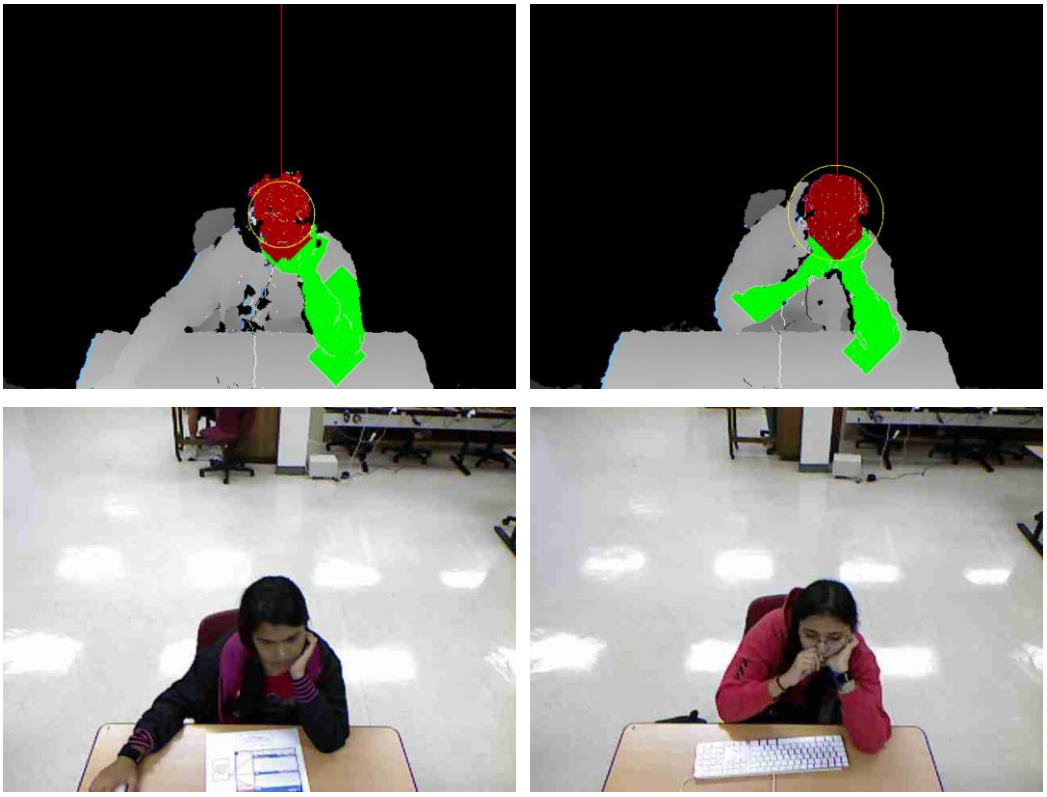


Figure 6.2. Detected hand-to-face gestures: one-hand-to-face (top left) and two-hands-to-face (top right). Color images are also shown (bottom row).

An overview of the hand-to-face gesture detection algorithm is shown in Algorithm 2. The breadth-first surface propagation mentioned on line 5 of Algorithm 2 adds pixels to the set of "surface pixels" through a between-neighbors comparison of an empirically-determined surface gradient threshold. Thus, the "head surface" propagates outward from the *headPixel* (the pixel from which propagation began). If a hand-to-face gesture was detected during surface propagation (as determined by difference between mean and median distances of surface pixels from *headPixel* on line 10), then the later-propagated surface pixels were considered "hand pixels."

Algorithm 2: HANDTOFACEGESTUREDETECTOR(*I*)

input : a depth image *I*
output : a value indicating gesture presence or absence

headCenter ← median center column selected from top 10% of rows containing non-zero depth values;
headRow ← lowest row in top 10% of rows containing non-zero depth values;
headPixel ← pixel location at (*headCenter*, *headRow*);
gestureDetected ← **false**;
while performing breadth-first surface propagation **do**
    medianXDistance ← median of horizontal distances of surface pixels from *headPixel*;
    meanXDistance ← mean of horizontal distances of surface pixels from *headPixel*;
    medianYDistance ← median of vertical distances of surface pixels from *headPixel*;
    meanYDistance ← mean of vertical distances of surface pixels from *headPixel*;
if |mean – median| ≥ 2.5% of mean or median **do**
        gestureDetected ← **true**;
if *gestureDetected* **do**
    handPixels ← surface pixels propagated after hand-to-face gesture was detected;
    if ≥33% of *handPixels* to upper right of *headPixel* **do**
        return NOGESTURE;
    else if ≥33% of *handPixels* to upper left of *headPixel* **do**
        return NOGESTURE;
    else if ≥33% of *handPixels* to lower left of *headPixel* **and** ≥33% of *handPixels* to lower right of *headPixel* **do**
        return TWOHANDSTOFACE;
    else if ≥33% of *handPixels* to lower left of *headPixel* **or** ≥33% of *handPixels* to lower right of *headPixel* **do**
        return ONEHANDTOFACE;
    else return NOGESTURE;
else return NOGESTURE;

To evaluate the algorithm, two human judges individually examined images corresponding to one-minute snapshots of the interactions and identified one-hand-to-face and two-hands-to-face gestures. The algorithm output was compared against all

instances where the judges agreed (Cohen's κ=0.96 for one-hand-to-face and κ=0.87 for two-hands-to-face). For those agreed-on instances, the accuracy of the algorithm was 92.6% across 1,170 depth images. Error cases typically involved such things as the surface propagation algorithm misidentifying clothing or hair as a hand.

Chapter 7. Automated Posture and Gesture Analysis

Automated tracking of posture and gesture enables analysis of bodily expressions of affect. These embodiments of affect, such as posture shifting or hand-to-face gestures, occur throughout tutoring interactions and have intuitive emotional meaning from our everyday lives. Thus, analyses of how these nonverbal behaviors coincide within tutorial dialogue may provide empirical evidence of their associations with affect. Sections 7.1 and 7.2 consider the less obvious connection between a tutor's impression of a student and student bodily expressions of affect. A student may implicitly communicate their affective state through indirect aspects of computer-mediated communication, such as their rate of typing, moments of pause, or time spent reading a task description. Together, these indirect signals may form an *implicit affective channel* that the tutor perceives. Section 7.3 considers affective phenomena involving direct communication, investigating the relationship between student embodiments of affect and tutorial dialogue moves.

7.1. Implicit Affect in Computer-Mediated Communication

An underlying notion of the implicit affective channel hypothesis is that tutors may have been able to identify student cognitive-affective states to some extent even when bodily movements associated with those affective states were not communicated to the tutor. To examine this, tutor perceptions of student affect and cognition were collected after each tutoring session (see Appendix A: Tutor Post-Session Survey). The tutor reports were compared against student self-reports using correlational analyses in a two-step process designed to mitigate the potential for false positives (Type I error). In the first step, significant correlations were identified on a test data set drawn from the second of six tutoring sessions in which each tutor/student pair engaged (N=42). Tutor perceptions of student affect were compared against learning outcomes (posttest minus pretest) and student affect self-reports. Student affect self-reports were in turn compared against tutor reports of cognitive variables, tutor reports of student affect,

and learning outcomes. Forty-three significant correlations were identified in this first step, and in the second step analyses were conducted to identify which of these significant correlations also held in a different data set, the first tutoring session, which is the main focus of the present study (*N*=42). The two-step analysis identified eight statistically reliable correlations, shown in Table 7.1.

Table 7.1. Significant correlations of survey variables (T=tutor report; S=student report)

First Variable	Second Variable	*r*	*p*
Focused Attention[S]	Helped Speed[T] (Figure 4, item 3)	-0.42	0.019
	Helped Mastery[T] (Figure 4, item 4)	-0.48	<0.01
	Student Confusion[T] (Figure 4, item 15)	-0.39	0.029
Physical Demand[S]	Student Frustration[T] (Figure 4, item 18)	0.44	0.014
	Tutor Frustration[T] (Figure 4, item 25)	0.42	0.019
Frustration Level[S]	Student Confusion[T]	0.53	<0.01
Student Confusion[T]	Student Frustration[T]	0.59	<0.01
Posttest Score	Student Confusion[T]	-0.38	0.038

The significant correlations highlight three student cognitive-affective states: focused attention, physical demand, and frustration. Students' reports of focused attention correlated with tutors' beliefs that they were less helpful and with tutors' beliefs that students were less confused. Tutors may have perceived that students who focused on the programming tasks did not need as much help to complete the tasks within the time allotted or to understand the related concepts. This result is compatible with the theory of optimal experience (Csikszentmihalyi, 1990), which posits an

optimally productive state of *flow* in which a student is learning well and it is often desirable not to interrupt his or her progress. Additionally, tutors may have perceived focused students as having less confusion throughout the session.

In addition to negatively correlating with students' reports of focused attention, tutor reports of student confusion were positively correlated with student self-reports of frustration. The relationship between frustration and confusion during learning may be a complex one. Specifically, frustration is typically considered to be a negative affective state, with persistent frustration referred to as a *state of stuck* (Kapoor et al., 2007), in which performance on the task at hand is negatively impacted. However, confusion has been theorized to be a cognitive-affective state with either positive or negative outcomes, depending on its resolution. In the theory of *cognitive disequilibrium* (Graesser & Olde, 2003), confusion occurs with partial understanding of new knowledge, which may lead to learning when new knowledge is understood. However, it may lead to frustration when the confusion is not resolved. In the current study, positive resolutions of confusion may not have been as memorable for tutors, which could leave the tutor to report lingering confusion that may have led to student frustration. The negative correlation between tutor reports of student confusion and posttest scores supports this interpretation, as frustration is known to negatively impact learning (Kort, Reilly, & Picard, 2001).

Tutor reports of their own frustration, and of student frustration, correlated with student-reported physical demand. That is, the more physically demanding the student felt the task was, the more frustrated the tutor felt and believed the student felt as well. The student rating of physical demand was measured with an item phrased, "How physically demanding was the task?" At first glance, it is unclear whether the students were rating discomfort related to movement/sitting or physiological stress, since (as will be described in Section 7.2) student report of physical demand did not correlate with measures of posture and gesture. The significant correlation with this

report of physical demand may indicate a co-occurring pattern of negative interaction in which the tutor was frustrated and the student was stressed.

It is also worth exploring the cognitive dimensions of the tutors' reporting. The tutors' reports of helping the student complete the task more swiftly and helping the student better master the subject material were both negatively correlated with focused attention, as described above. However, none of the other tutor cognitive reports correlated across the two-phase analysis. Additionally, student performance, as measured by test performance and learning gains, yielded a single correlation between posttest score and tutor report of student confusion. This may indicate that the phenomena evidenced here are related more to implicit perception of cognitive-affective phenomena than to purely cognitive or task-related phenomena. However, the interplay of cognition and affect in task-oriented domains merits further study.

7.2. Multimodal Analysis of Implicit Affect

The results in the previous section suggest that cognitive-affective states are implicitly communicated in textual dialogue. Examining bodily expressions such as posture and gesture may reveal aspects of affective ground truth related to the implicit affective channel.

Two aspects of posture were used as features in the models reported here. First, variance of the tracked posture points was used as a measure of quantity of motion. Second, the average postural position across a session was used to capture the predominant body position of the student. In addition to these, gesture features include relative frequencies of one-hand-to-face and two-hands-to-face gestures. The set of posture and gesture features is shown in Table 7.2. The *H*, *M*, and *L* prefixes correspond to the three posture estimation points (Figure 6.1), while the *All* prefix indicates the sum of all three points. After removing sessions with error-prone depth recordings, thirty-one sessions were included in the multimodal analyses.

Table 7.2. Posture and gesture features used in multimodal analyses

Feature Set	Feature Names
Averages of posture points	*HAvg*, *MAvg*, *LAvg*, *AllAvg*
Variances of posture points	*HVar*, *MVar*, *LVar*, *AllVar*
Relative frequencies of hand-to-face gestures	*NoGestRFreq*, *OneHandRFreq*, *TwoHandsRFreq*

7.2.1 Correlation of Survey Variables with Posture and Gesture

The first analysis goal was to determine whether the cognitive-affective dimensions investigated earlier also correspond to bodily movements. To accomplish this, correlational analyses were performed between posture/gesture and variables that were involved in the significant correlations identified in Table 7.1. The posture and gesture features in Table 7.2 were paired with the survey variables in Table 7.1, and the resulting statistically significant correlations are shown in Table 7.3 (N=31).

Posture and gesture primarily correlate with survey variables for three cognitive-affective phenomena: student self-report of focused attention, tutor report of student confusion, and tutor report of both student and tutor frustration. Students' report of increased focused attention corresponded to less movement in the lower torso (*LVar*), and to a lower frequency of two-hands-to-face gestures (*TwoHandsRFreq*). These correlations may highlight instances of students leaning forward onto both hands while also moving about the lower torso. Additionally, *LVar* negatively correlated with posttest score, which may illustrate a trend related to lower focused attention.

Table 7.3. Posture and gesture correlations with survey variables (T=tutor report; S=student report)

First Variable	Second Variable	*r*	*p*
Focused Attention[S]	*LVar*	-0.37	0.040
	TwoHandsRFreq	-0.39	0.031
Student Confusion[T] (Figure 4, item 15)	*HAvg*	-0.44	0.012
	MAvg	-0.47	<0.01
	LAvg	-0.40	0.026
	AllAvg	-0.48	<0.01
Student Frustration[T] (Figure 4, item 18)	*MAvg*	-0.38	0.035
	LVar	-0.37	0.043
	OneHandRFreq	-0.43	0.017
Tutor Frustration[T] (Figure 4, item 25)	*HVar*	0.44	0.013
	NoGestRFreq	0.37	0.043
	OneHandRFreq	-0.39	0.029
Posttest Score	*LVar*	-0.40	0.026

Tutor reports of student confusion negatively correlated with average student postural distance. Thus, higher tutor reports of student confusion co-occurred with a more forward student body position. Conversely, farther postural distances co-occurred with lesser tutor reports of student confusion. Similarly, farther mid torso distances co-occurred with tutor reports of student frustration. Taken as a whole, these correlations appear to suggest that forward-leaning postures occur with negative cognitive-affective experience (as perceived by the tutor). Conversely, average postural configurations closer to a straight sitting posture co-occurred with more positive cognitive-affective experience.

Tutor reports of student and tutor frustration both negatively correlated with relative frequency of one-hand-to-face gestures. This is in contrast with the two-hands-to-face gesture, which co-occurred with lower student focused attention. It may be that one-hand-to-face gestures tend to express a positive or thoughtful state, as noted in related literature (Mahmoud & Robinson, 2011).

7.2.2 Predicting Affective Outcomes with Posture and Gesture

The correlations presented in the previous section suggest ways in which posture and gesture are associated with student and tutor perceptions of cognitive-affective experience. To further elucidate these relationships, multivariate regression models were built with the significantly correlated variables as predictors. The three survey variables for student focused attention, confusion, and frustration were modeled as outcome variables. Each stepwise linear regression used a conservative 0.05 significance threshold for addition of features.

The regression model for focused attention, shown in Table 7.4, incorporates two-hands-to-face gestures with variance and average postural position of the lower torso. The model R^2 shows that a moderate amount of the variance in focused attention is explained. Both two-hands-to-face gestures and lower torso variance were negative predictors of focused attention. However, lower torso average distance explains further variance of focused attention as a positive predictor. This model augments the results of the correlational analyses by showing that posture and gesture together combine to predict students' focused attention.

Table 7.4. Stepwise linear regression model for student-reported Focused Attention. Partial R^2 shows the contribution of each feature, while model R^2 shows cumulative model effect.

Focused Attention =	Partial R^2	Model R^2	*p*
-40.90 * *TwoHandsRF*	0.150	0.150	0.031
-2.90 * *LVar*	0.143	0.293	0.025
0.99 * *LAvg*	0.103	0.396	0.041
23.66 (intercept)	**RMSE** = 10% of variable's range		

The stepwise linear regression model for tutor-reported student confusion, displayed in Table 7.5, contains a single posture feature that explains a small amount of variance. The absence of additional features shows that the other posture features correlated with student confusion in Table 7.3 were redundant.

Table 7.5. Stepwise linear regression model for tutor-reported student confusion.

Confusion =	Partial R^2	Model R^2	*p*
-0.16 * *AllAvg*	0.231	0.231	<0.01
4.24 (intercept)	**RMSE** = 20.4% of variable's range		

The stepwise linear regression model for tutor-reported student frustration, shown in Table 7.6, includes relative frequency of one-hand-to-face gestures and lower torso variance as negative predictors. Contrary to the correlational analyses in Table 7.3, head variance and absence of hand-to-face gestures did not meet the threshold of significance. This model underscores the interplay of posture and gesture, as the addition of lower torso nearly doubles the explained variance.

The regression analyses revealed cumulative effects when posture and gesture were integrated into linear regression models. The model for focused attention

incorporated two-hands-to-face gestures and lower torso variance and average distance. Either posture or gesture alone would have explained a small amount of variance, so this demonstrates that the combination of multimodal features such as posture and gesture can improve a predictive model.

Table 7.6. Stepwise linear regression model for tutor-reported student frustration.

Frustration =	**Partial R^2**	**Model R^2**	***p***
-2.96 * *OneHandRF*	0.182	0.182	0.017
-0.16 * *LVar*	0.155	0.337	0.016
2.86 (intercept)	**RMSE** = 16.2% of variable's range		

The regression model for student confusion did not include gesture, with a small amount of variance explained. However, the regression model built for student frustration included one-hand-to-face gestures and lower torso variance features, resulting in greater explained variance. The root mean squared error of the model for student frustration was less than that of the model for student confusion, as would be expected of the model that explains more variance.

7.2.3 Interpretation of the Implicit Affective Channel

Although textual communication has limited bandwidth, interactions through the medium still retain high cognitive and affective complexity. If the textual content itself is emotionally sparse, the participants may rely on implicit interpretation of affect. The information relevant to this interpretation may be conceived of as being transmitted through an implicit affective channel. Understanding this implicit affective channel may hold great benefit for systems that aim to interact in naturalistic ways with humans. Toward this end, this section of the dissertation work has presented an empirical study to investigate two aspects of implicit affective communication in textual dialogue: 1) the

extent to which the participants converge on shared perceptions of affect, and 2) the ways in which affective ground truth, as captured by depth recordings of gesture and posture, correlates with those affective perceptions even when the bodily movements were not transmitted to the other participant.

The automatically recognized posture and gesture features were explored within models that indicate *focused attention*, *physical demand*, and *frustration* were perceived through the hypothesized implicit affective channel accompanying textual dialogue. These results support the notion that an implicit affective channel is at work within computer-mediated textual communication, and that bodily displays of posture and gesture co-occurred with implicit affective communication.

7.3. Embodied Affect in Task-Oriented Tutorial Dialogue

This section presents an analysis of posture and gesture within computer-mediated textual tutorial dialogue in the JavaTutor Study II corpus. Utilizing automated algorithms that measure postural quantity of motion, one-hand-to-face gestures, and two-hands-to-face gestures, interdependencies between dialogue acts and student posture and gesture were examined in order to identify ways in which the nonverbal behaviors may influence or be influenced by dialogue. Additionally, an analysis of group-wise differences in nonverbal behavior displays showed that students with lower self-efficacy tended to produce more two-hands-to-face gestures. These findings comprise a step toward understanding how embodied affect intertwines with tutorial dialogue.

Dialogue acts were annotated using a parallel coding scheme that was applied to both tutor and student utterances. The coding scheme used here is an update to a prior task-oriented dialogue annotation scheme (Boyer et al., 2010). Three annotators tagged a subset of the corpus (N=36). Fourteen percent of these annotated sessions were doubly annotated, with a resulting average agreement across dialogue acts of Cohen's κ=0.73. The dialogue act tags and frequencies in the corpus are shown in Table 7.7.

Table 7.7. Dialogue act tags ordered by frequency in the corpus (*S*=student, *T*=tutor)

Act	Example Tutor Utterances	*S*	*T*
STATEMENT	*"java does things in the order you say."*	282	1255
QUESTION	*"Any questions so far?"*	213	630
POSITIVE FEEDBACK	*"great debugging!"*	2	539
DIRECTIVE	*"change that in all three places"*	-	252
HINT	*"it is missing a semicolon."*	-	223
ANSWER	*"yes, now line 1 is a comment."*	547	162
ACKNOWLEDGMENT	*"alright" "okay" "Yes"*	323	68
LUKEWARM FEEDBACK	*"Right, nearly there"*	-	32
NEGATIVE FEEDBACK	*"no" "nope"*	-	19
CORRECTION	Repairing a prior utterance: *"*can use"*	11	15
REQUEST CONFIRMATION	*"Make sense?" "okay?"*	6	14
OTHER	*"LOL"*	11	6
REQUEST FOR FEEDBACK	*"How does that look?"*	11	1

The posture tracking and gesture detection algorithms described in Chapter 6 were run on all sessions, but four sessions had no Kinect recordings due to human error (*N*=38). The combined corpus of dialogue acts and nonverbal tracking data contains 32 sessions. Sample output of posture and gesture tracking is shown in Figure 7.1.

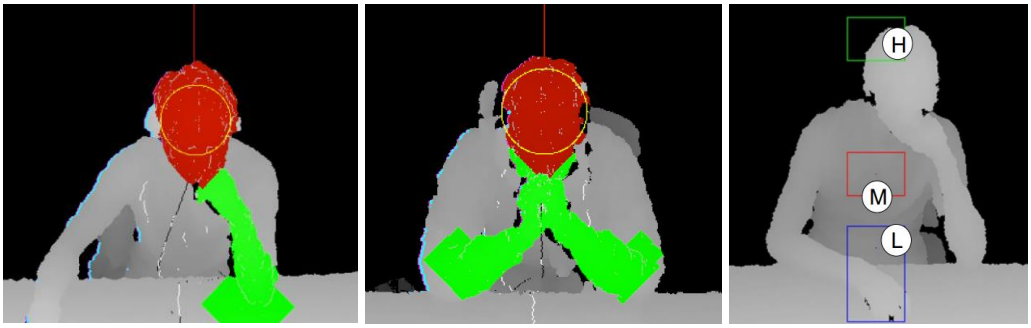


Figure 7.1. Tracked gestures (one-hand-to-face, two-hands-to-face) and posture

The posture tracking values were converted into a "postural shift" feature, a discrete representation of *quantity of motion* (Sanghvi et al., 2011). Postural shifts were identified through tracked head distances as follows. The median head distance of students at each workstation was selected as the "center" postural position. Distances at one standard deviation (or more) closer or farther than "center" were labeled as "near" or "far," respectively. Postural shifts were labeled when a student moved from one positional category to another (e.g., from "near" to "center"). Both postural shift and gesture events were smoothed by removing those with duration of less than one second. This smoothing mitigated the problem of jitter at decision boundaries (e.g., slight movements at the boundary between "center" and "far" postural positions that cause rapid swapping of both labels). The nonverbal behaviors will hereafter be referenced with the labels ONEHAND, TWOHANDS, and PSHIFT.

7.3.1 Co-Occurrence of Dialogue and Nonverbal Behavior

Tutorial dialogue and nonverbal behavior have both been extensively examined separately from each other, but there are few investigations of their interactions (Ha, Grafsgaard, Mitchell, Boyer, & Lester, 2012). This dissertation includes a series of analyses to identify co-dependencies between tutorial dialogue and nonverbal behavior. First, overall dialogue act frequencies and dialogue act frequencies

conditioned on presence of nonverbal displays were compared. Then, a series of group-wise analyses identified whether differences existed between students based on gender, prior knowledge, and domain-specific self-efficacy. Statistically significant results are shown in bold.

The first analyses consider the frequency of dialogue acts given that a nonverbal behavior occurred either before or after a dialogue act. An empirically determined fifteen-second interval was used to tabulate occurrence of nonverbal behavior events both before and after dialogue acts. The frequencies were normalized for individuals and averaged across the corpus. Thus, the values shown in the analyses below are average relative frequencies. Dialogue acts with overall average relative frequency below 1% were excluded from the analyses.

The analyses of student dialogue acts consider two situations for each nonverbal behavior. The first examines student dialogue acts given that a nonverbal behavior occurred prior to a dialogue act. This may show how student dialogue moves are affected by the nonverbal behaviors. The second situation considers student dialogue acts given that a nonverbal behavior followed. This represents differences in how a student proceeded following their own dialogue act. In both situations, the nonverbal context may provide insight into the dialogue.

The analyses of student dialogue acts conditioned on prior ONEHAND events revealed a statistically significantly lower frequency of student QUESTIONS following ONEHAND gestures. There was also a trend of more student answers following ONEHAND gestures (Table 7.8).

Table 7.8. Analyses of student dialogue acts preceded by ONEHAND gesture

Student Dialogue Act	Relative Freq. of Stud. Act (*stdev*)	Rel. Freq. of Stud. Act with ONEHAND Prior (*stdev*)	*p*-value (paired *t*-test, two-tailed, *N*=30)
ANSWER	0.42 (*0.16*)	0.50 (*0.27*)	0.114
ACKNOWLEDGMENT	0.22 (*0.08*)	0.22 (*0.23*)	0.878
QUESTION	**0.14 (*0.09*)**	**0.08 (*0.16*)**	**0.048**
STATEMENT	0.18 (*0.09*)	0.18 (*0.22*)	0.896

The analyses of student dialogue acts followed by PSHIFT events showed a statistically significant lower frequency of student questions followed by PSHIFT (Table 7.9).

Table 7.9. Analyses of student dialogue acts followed by PSHIFT postural event

Student Dialogue Act	Relative Freq. of Stud. Act (*stdev*)	Rel. Freq. of Stud. Act Followed by PSHIFT (*stdev*)	*p*-value (paired *t*-test, two-tailed, *N*=24)
ANSWER	0.40 (*0.13*)	0.43 (*0.33*)	0.649
ACKNOWLEDGMENT	0.23 (*0.09*)	0.29 (*0.29*)	0.296
QUESTION	**0.15 (*0.09*)**	**0.08 (*0.12*)**	**0.019**
STATEMENT	0.20 (*0.11*)	0.16 (*0.20*)	0.246

The analyses of tutor dialogue acts are conditioned on student nonverbal behaviors present after a tutor move, which may show how students reacted to tutor moves. The analyses of tutor dialogue acts followed by posture identified statistically significant lower frequencies of tutor DIRECTIVES and tutor POSITIVE FEEDBACK followed by PSHIFT (Table 7.10). The analyses of tutor dialogue acts followed by TWOHANDS revealed

statistically significant lower frequencies of tutor ANSWERS and tutor DIRECTIVES followed by TWOHANDS (Table 7.11). Additionally, there was a trend of greater frequency of questions followed by TWOHANDS.

Table 7.10. Analyses of tutor dialogue acts followed by PSHIFT postural event

Tutor Dialogue Act	Relative Freq. of Tutor Act (*stdev*)	Rel. Freq. of Tutor Act Followed by PSHIFT (*stdev*)	*p*-value (paired *t*-test, two-tailed, *N*=24)
ANSWER	0.04 (*0.03*)	0.04 (*0.07*)	0.722
ACKNOWLEDGMENT	0.03 (*0.03*)	0.06 (*0.13*)	0.162
DIRECTIVE	**0.08 (*0.04*)**	**0.05 (*0.06*)**	**0.012**
HINT	0.07 (*0.05*)	0.11 (*0.20*)	0.350
POSITIVE FEEDBACK	**0.18 (*0.05*)**	**0.13 (*0.10*)**	**0.033**
QUESTION	0.21 (*0.07*)	0.26 (*0.24*)	0.359
STATEMENT	0.36 (*0.10*)	0.32 (*0.23*)	0.419

Table 7.11. Analyses of tutor dialogue acts followed by TWOHANDS gesture

Tutor Dialogue Act	Relative Freq. of Tutor Act (*stdev*)	Rel. Freq. of Tutor Act Followed by TWOHANDS (*stdev*)	*p*-value (paired *t*-test, two-tailed, *N*=23)
ANSWER	**0.05 (*0.03*)**	**0.01 (*0.03*)**	**<0.001**
ACKNOWLEDGMENT	0.03 (*0.03*)	0.01 (*0.04*)	0.258
DIRECTIVE	**0.08 (*0.03*)**	**0.03 (*0.05*)**	**<0.001**
HINT	0.06 (*0.05*)	0.04 (*0.11*)	0.382
POSITIVE FDBK	0.18 (*0.05*)	0.21 (*0.18*)	0.524
QUESTION	0.19 (*0.07*)	0.26 (*0.25*)	0.135
STATEMENT	0.39 (*0.09*)	0.39 (*0.30*)	0.977

The primary focus of the above analyses was to investigate the relationships between tutorial dialogue and student nonverbal behaviors. However, the broader nature of nonverbal behavior in tutoring can be explored through analyses conditioned upon student characteristics. For this purpose, three group-wise analyses were conducted to examine gender and domain-specific self-efficacy. First, students were grouped into categories of male (*N*=28) and female (*N*=10). Comparisons of PSHIFT, ONEHAND, and TWOHANDS yielded no significant differences (*t*-tests with unequal variance, two-tailed). Second, students were grouped through a median split on pretest score, with high prior knowledge (*N*=19) and low prior knowledge (*N*=19). Comparisons of PSHIFT, ONEHAND, and TWOHANDS yielded no significant differences (*t*-tests with unequal variance, two-tailed). Finally, a median split on domain-specific self-efficacy was performed to create groups of high self-efficacy (*N*=19) and low self-efficacy (*N*=19). No differences were found in ONEHAND or PSHIFT across the groups (*t*-tests with unequal variance, two-tailed). However, students who reported low self-efficacy were found to display more TWOHANDS gestures (*t*-test with unequal variance, two-tailed). Students in the low self-efficacy group had an average of 0.53 TWOHANDS displays per minute (*N*=19, *stdev*=0.52), while the high self-efficacy group had an average of 0.20 TWOHANDS displays per minute (*N*=19, *stdev*=0.34). This result was statistically significant with p=0.029.

7.3.2 Interpretation of Posture and Gesture in Tutoring

The hand-to-face gestures examined here are in a class different from those involved in social conversation and face-to-face tutoring. In face-to-face interaction, social communication guides the nonverbal interaction (McNeill, 2005). Objects in the surrounding environment and spoken concepts form a common substrate that is referenced in conversational gestures. In the case of computer-mediated tutoring, social displays are greatly reduced (Grafsgaard, Fulton, et al., 2012). Thus, hand-to-face

gestures may be more representative of the cognitive-affective states that accompany them compared to communicative or social gestures.

One-hand-to-face gestures are often thought of as embodiments of a thoughtful state.[1] Here, student questions were found to be less frequent following a one-hand-to-face gesture. It may be that students who presented one-hand-to-face gestures had fewer questions to ask. Only fifteen percent of one-hand-to-face gestures occurred before student utterances. Additionally, one-hand-to-face gestures most frequently occurred before student answers. Students are likely to think before providing an answer and in work on task outside of the dialogue. The occurrence of one-hand-to-face gestures coincides with both of these thought-provoking events. Thus, this tutoring corpus supports interpretation of one-hand-to-face gestures as a nonverbal behavior with an underlying thoughtful state.

The group-wise self-efficacy analysis presented here showed that students with lower self-efficacy tend to produce more two-hands-to-face gestures. Coupled with a prior result (Section 7.2) that found two-hands-to-face gestures to be negatively correlated with focus, a picture emerges of this gesture as an embodiment of reduced focus and lower confidence. Here, tutor answers and tutor directives were less likely to be followed by two-hands-to-face displays. This appears to indicate that students were more focused after these tutor moves. Both tutor answers and directives provide responsive instruction to the student. In the case of answers, the student would have asked a question, and thus would be attentively waiting for the tutor's answer. With directives, the tutor is supplying the student with direct task solution steps that the student must then act upon. The interface did not allow tutors to edit students' computer programming code, so tutor directives imply subsequent student work.

Postural shifts have been linked with disengagement or negative affect. Studies in different contexts agree: whether it is a child playing a game with a robot (Sanghvi et al., 2011) or a student interacting with a tutoring system (D'Mello et al., 2012; Rodrigo

[1] One such gesture has even been cast in bronze as a timeless exemplar, "The Thinker."

& Baker, 2011; Woolf et al., 2009), postural shifting has repeatedly been shown to co-occur with disengaged or negative cognitive-affective states. Thus, the postural shifts examined in these analyses most likely indicate a disengaged affective state. In this case, less disengagement followed student questions, tutor answers, and tutor positive feedback. Each of these dialogue acts is directly related to collaborative tutorial interaction in which the student is more likely to be engaged. In the case of student questions and tutor answers, the student has posed the question and subsequently received a response. The student clearly plays an active role in this pattern, so it is not surprising that their body reflects this. With tutor positive feedback, the tutor has praised the student for completing a sub-task. The student was actively engaged in the computer programming task, so this result shows that both the student's body and tutor praise reflect the student's engagement.

As noted in (Mahmoud & Robinson, 2011), there are many variants of hand-to-face and hand-over-face gestures. The hand-to-face gestures tracked here consider contact between hands and the lower face, without more detail as to how the hand is touching the face (e.g., the difference between holding one's chin and leaning on the palm of a hand). Additionally, temporal characteristics of these gestures may be important. An individual may stroke his or her chin, as opposed to resting on a hand. Thus, the present analyses aggregate an array of more specific gestures into categories of one-hand-to-face or two-hands-to-face. Further development efforts are needed to provide tracking algorithms that distinguish between the spatiotemporal subtleties of hand and face (Kleinsmith & Bianchi-Berthouze, 2012).

The results presented in this section indicate that posture and hand-to-face gestures are significantly associated with student questions, tutor answers, tutor directives and tutor positive feedback. Additionally, two-hands-to-face gestures occurred significantly more frequently among students with low self-efficacy. The results shed light on the cognitive-affective mechanisms that underlie these nonverbal behaviors. Collectively, the findings provide novel insight into the interdependencies

among tutorial dialogue, posture, and gesture, revealing a new avenue for automated tracking of embodied affect during learning.

Chapter 8. Hidden Markov Models

This chapter considers sequential analysis of affect. The manual annotations of the JavaTutor Study I corpus formed sequences of dialogue, task progress, and nonverbal behavior. These sequences are the basis for identifying multimodal subsequences related to affect. In this approach, hidden Markov models were used to predict nonverbal behavior (Section 8.1) and also analyzed to identify and interpret recurring subsequences as indicators of affective tutorial interaction (Sections 8.1.1 and 8.2).

A hidden Markov model (HMM) is defined by an *initial probability distribution* across hidden states, *transition probabilities* among hidden states, and *emission probabilities* for each hidden state and observation symbol pair (Rabiner, 1989). The hidden states represent the underlying probabilistic system that generates a given sequence of observed events. The initial state probability gives the possibility of beginning in any hidden state. Transition probabilities encode the likelihood of entering one hidden state from another. Emission probabilities encode the likelihood of producing a given observation from a particular hidden state. HMMs learn statistical dependencies between hidden states and the corresponding observations. The hidden state structure can then be analyzed to identify underlying trends. Using HMMs, it is possible to uncover a rich interplay between learner affect, tutorial dialogue and task context.

8.1. Sequential Modeling of Brow Lowering

The observation sequences consist of annotated observations from the corpus, including dialogue moves by tutor or student, or student task action segments. Each of these observations also includes a tag for whether student AU4 was associated with that event. For example, the observation symbol sequence that corresponds to Excerpt 1 in Figure 8.1 is [STUDENT NEGATIVE CONTENT FEEDBACK NOAU4, CORRECT TASK ACTION AU4, TUTOR ASSESSING QUESTION AU4, TUTOR STATEMENT NOAU4, STUDENT POSITIVE CONTENT FEEDBACK NOAU4].

Excerpt 1			
13:16:03	Tutor:	no, it's easier than that, you just have to make the middle if into an "else if" [NEGATIVE CONTENT FEEDBACK]	STATE 10
	Student:	CORRECT TASK ACTION AU4	STATE 6
13:16:31	Tutor:	does that make sense? [ASSESSING QUESTION AU4]	STATE 10
13:16:41	Tutor:	that way it only checks the 2nd conditional if the first one failed [STATEMENT]	STATE 8
13:17:20	Student:	it makes sense now that you explained it […] [POSITIVE CONTENT FEEDBACK]	STATE 4
Excerpt 2			
14:52:18	Tutor:	no, before we start sorting [NEGATIVE CONTENT FEEDBACK AU4]	STATE 10
	Student:	CORRECT TASK ACTION AU4	STATE 6
14:52:27	Tutor:	so, before the first loop you can use i for this loop counter if you want to [STATEMENT AU4]	STATE 10
	Student:	MIXED PROGRESS TASK ACTION AU4	STATE 6
14:53:52	Student:	i try to keep them different so i don't confuse myself [STATEMENT]	STATE 7

Figure 8.1. Excerpts from annotated tutoring session corpus, with most probable sequences of HMM hidden states

The HMMs were learned within a leave-one-out framework. Within each fold, five random restarts of model parameters were performed to reduce the potential of model convergence at a local optimum. An additional outer training loop, ranging from two to twenty, was performed to identify the optimal number of hidden states. The best-fit model had eighteen hidden states.

The leave-one-out design resulted in fourteen training/testing folds, one for each tutoring session. Four of these sessions contained an observation symbol (combination of dialogue move and AU4 presence/absence) that occurred nowhere else in the data, so the learned model was not used to predict on these sessions. Predictive findings from the remaining ten test sessions are presented here, though an online predictive model

used during tutoring could address this by learning across all possible symbols in the state space, regardless of absence in a particular session. The predictive accuracies of the HMMs are compared against a majority class baseline as well as a first-order observed Markov model (OMM) (Table 8.1). Accuracies that are statistically significantly better than baseline are in bold (paired *t*-test, $p < 0.005$).

Table 8.1. Comparison of predictive accuracy of classifiers

Classifier	Accuracy (across sessions)	Std. Dev. of Accuracy
HMM Train	**0.868**	0.021
HMM Test	**0.907**	0.059
OMM Train	0.186	0.013
OMM Test	0.557	0.284
Baseline	0.845	-

OMMs do not include hidden states, and thus condition the present state purely on the transition probability distribution from the previous state. The training set predictive accuracies indicate that HMMs fit the training data better than the other models. The predictive accuracy of the learned HMMs on the test set was higher on average than predictions on the training set, but not surprisingly, the standard deviation was also greater. Both the training and test predictions greatly outperformed the predictive accuracy of the OMMs, which were below baseline. This below-baseline performance of OMMs indicates that the presence or absence of AU4 at time *t* is not highly predictive of the presence or absence of AU4 at time *t+1*, which is an interesting

discovery in this corpus. However, the additional stochastic structure provided by the HMM is able to predict AU4 significantly above baseline.

8.1.1 Hidden Markov Model Interpretation

The predictive accuracies of the HMMs suggest that these models hold great promise for learner affect prediction. On unseen test data, the HMMs predicted significantly better than an OMM and a (very high) majority class baseline. To gain more insight into how the HMM structure facilitates prediction of student AU4, the structure of the learned (best-fit) HMM is examined.

HMMs' predictive power is gained in part by the way these models can learn higher-order structure (in the form of hidden states) based on observation sequences. The model structure shown in Figure 8.2 illustrates this, with emission probability distributions displayed as bar graphs and transitions as edges. A subset of four hidden states is shown, including transition (arrow) and emission (bar chart) probabilities greater than 0.05. To facilitate discussion, the states were named after model learning through qualitative analysis. STATE 6, *Student Work with Confusion,* is dominated by student task actions with AU4 present. STATE 10, *Tutor Help*, emits a combination of tutor dialogue moves with AU4 present, and tutor feedback with no AU4. STATE 4, *Overcoming Confusion*, is dominated by tutor statements with AU4 present and student positive feedback. This state corresponds to tutor statements that are not consistent with students' prior knowledge. Interestingly, the state also generates student positive feedback, which may indicate that students moved past cognitive disequilibrium and into a state of understanding. STATE 16, *Conversational Grounding*, primarily encompasses non-task-oriented student and tutor dialogue moves, but also generates with small probability student negative feedback without AU4.

These emission probability distributions indicate ways in which the HMM abstracts from observation sequences to meaningful higher-order structure. Of equal importance are the transition probabilities between the hidden states. STATE 6 and

STATE 10 are more likely to transition to each other than any other hidden states. This transition is illustrated in one sequence of events (Figure 8.1 Excerpt 1) in which the tutor provides negative content feedback followed by a student correct task action with confusion present (as evidenced by AU4). The tutor then asks an assessing question to gauge the student's understanding, which also coincides with a moment of confusion. The tutor further explains the computer programming concept with an instructional statement clarifying prior feedback. The student then takes a moment to reflect on the material and informs the tutor that the explanation was helpful. This example demonstrates the strong connection between *Student Work with Confusion*, STATE 6, and *Tutor Help*, STATE 10. Meaningful tutor feedback and instruction that induce confusion are produced in STATE 10, while STATE 6 corresponds with student tasks actions accompanied by confusion. Both states are highly relevant to learning.

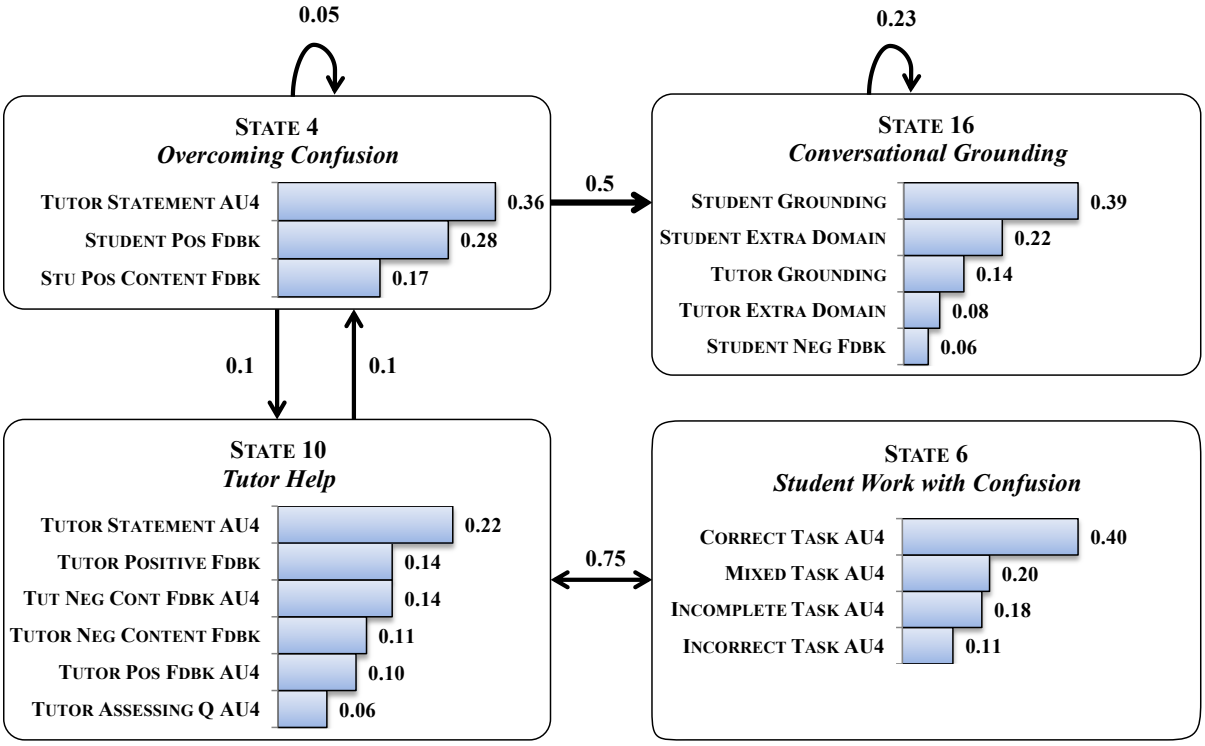


Figure 8.2. A subset of four hidden states in the best-fit HMM

The second example (Figure 8.1 Excerpt 2) further characterizes the interplay of STATE 6 and STATE 10, with the student progressing on the programming task (with AU4 present) while receiving tutor feedback and instruction. The excerpt begins with tutor negative content feedback on the student's current task progress, with student confusion indicated by AU4. The student then immediately completes the subtask, still showing AU4. The tutor continues instructing the student with a comment on a relevant programming concept (AU4 still present). The student then continues programming, with mixed progress (AU4 continues). After approximately a minute of working on the task, the student responds to the tutor statement with an explanation of the work performed. Thus, the student displays confusion until after the tutor completes instruction. Both excerpts seem to show effortful learning, with a combination of instruction during *Tutor Help* and task progress in *Student Work with Confusion*.

Therefore, HMMs represent a promising approach to automatically learn semantically meaningful affect-rich models of tutorial interaction.

8.2. Sequential Modeling of Upper and Lower Face Movements

The facial expression and dialogue data described in Section 3.4 were merged into sequences of observations needed to build a descriptive HMM. Each observation consisted of a facial expression (denoted as facial action units (AUs) (Ekman et al., 2002a)), dialogue act or both. The Baum-Welch algorithm with log-likelihood measure was used for model training. Ten random initializations were performed to reduce convergence to local maxima. A hyperparameter optimization outer loop produced candidate HMMs across a range from three to twenty-two hidden states. Average log likelihood was computed across candidate HMMs for each number of hidden states. The models with best average log-likelihood had ten hidden states, and the best-fit model had the highest log-likelihood among these.

With the model in hand, the Viterbi algorithm was applied to map the most probable hidden state to each observation. Exhaustive search to length five across each session's hidden state sequences revealed five frequently recurring sequences (or "patterns") of affective tutorial interaction. Each pattern occurred at a relative frequency greater than 0.05 across multiple sessions. Seven (of ten) hidden states comprised the patterns. The hidden states participating in the patterns are shown in Figure 8.3, with the most frequent transmissions and emissions (with probabilities greater than 0.5 or 0.05, respectively).

In order to examine the persistence of the five frequently-occurring patterns of affective tutorial interaction, average sequence lengths were calculated for each session (shown in the bottom right of Figure 8.3). There are subtle differences between relative frequency as a measure of prevalence and average sequence length as a measure of persistence. When the measures agreed (as was often the case), they showed prevalence and persistence of specific patterns of affective tutorial interaction within a

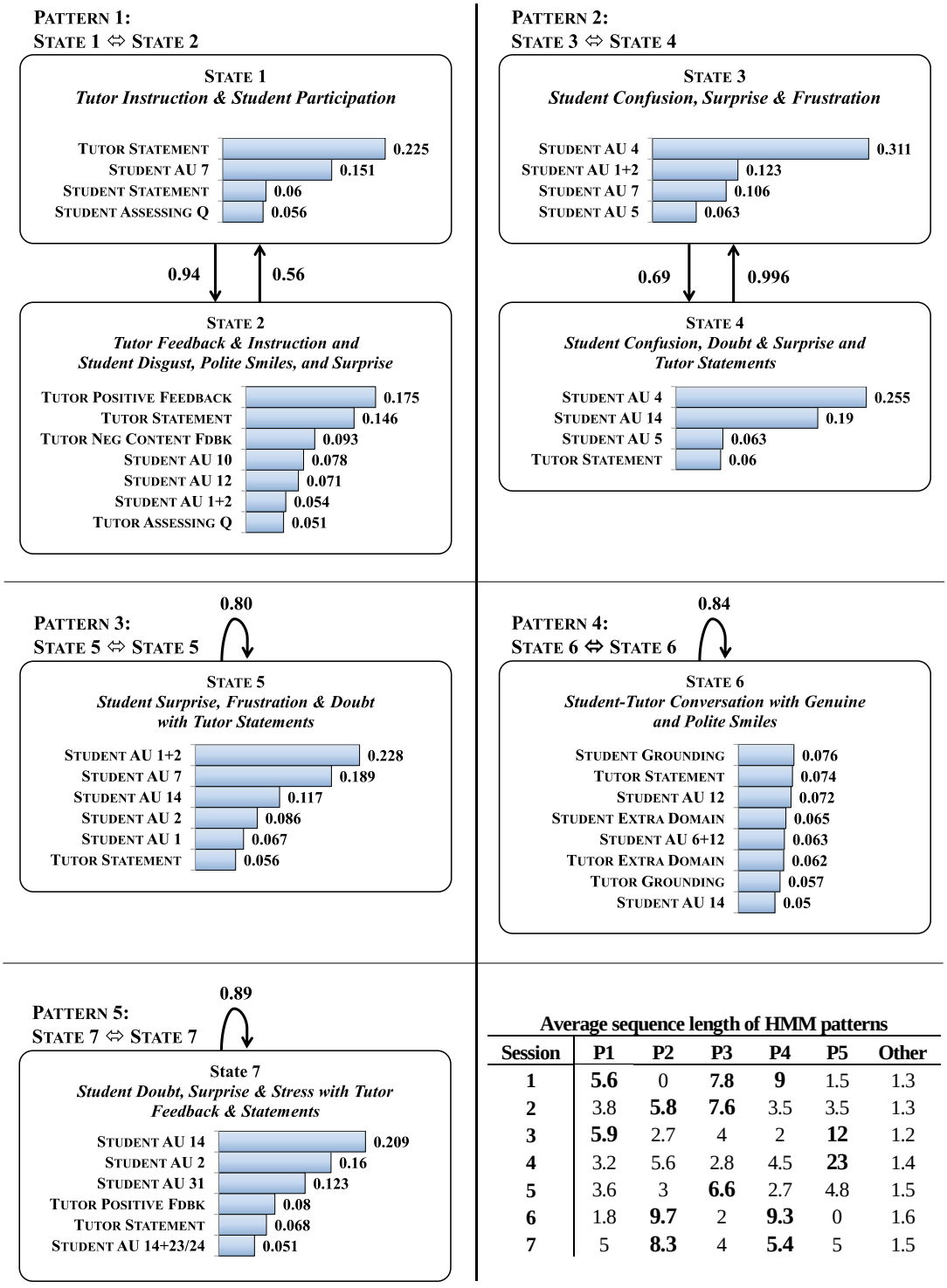
particular session. When the measures differed, a persistent pattern recurred in long, but rare, sub-sequences or a prevalent pattern recurred in short sub-sequences.

The average sequence lengths shown in Figure 8.3 indicate notable differences in affective tutorial interaction within sessions. Thus, it may be possible to group sessions that have similar quantitative profiles. For instance, sessions 6 and 7 both have persistent sequences of PATTERN 2 and PATTERN 4, indicative of persistent student confusion with tutor statements and conversational dialogue during those sessions. Likewise, PATTERN 1 models tutor lecturing and instruction with occasional student participation and student affective states, PATTERN 3 is dominated by student facial displays (mostly surprise and frustration), and PATTERN 5 is largely composed of doubt, surprise, and stress with occasional tutor feedback and statements. In this way, quantitative application of HMMs provides insight into profiles of affective tutorial interaction across tutoring sessions.

These results highlight a potential approach for further analysis of multimodal sequences in task-oriented domains. As presented in Chapter 5 through Chapter 7, automated approaches have been successfully applied to identify facial expression, gesture, and posture. These modalities produce rich sequences of nonverbal behavior observations that have been associated with affect and learning. Constructing HMMs from such multimodal sequences may reveal repeated sequences that are indicative of positive or negative affect.

However, when this approach was applied to the multimodal data streams in the JavaTutor Study II corpus, the large number of observation symbols (facial expression, posture, gesture, dialogue, task) resulted in a multiplicative expansion of the symbol space. For example, when considering a greater number of facial movements coupled with posture and gesture, the result is the Cartesian product of the symbols, rather than a simple sum. In this vast symbol space, the resulting HMMs were sparse and did not model coherent patterns in the data. Despite this problem, there may be ways to improve this approach, including the possibility of combining low-level observations to

form a smaller set of high-level observations. Subsequently, further explorations of the multimodal data streams contained in the JavaTutor Study II corpus were conducted, as described in the next chapters.



Average sequence length of HMM patterns

Session	P1	P2	P3	P4	P5	Other
1	**5.6**	0	**7.8**	**9**	1.5	1.3
2	3.8	**5.8**	**7.6**	3.5	3.5	1.3
3	**5.9**	2.7	4	2	**12**	1.2
4	3.2	5.6	2.8	4.5	**23**	1.4
5	3.6	3	**6.6**	2.7	4.8	1.5
6	1.8	**9.7**	2	**9.3**	0	1.6
7	5	**8.3**	4	**5.4**	5	1.5

Figure 8.3. Five patterns (i.e. frequently recurring sequences of hidden states)

Chapter 9. Predictive Models from Multimodal Data

This chapter presents models predicting affective and learning outcomes from moment-to-moment nonverbal behavior and task performance in the JavaTutor Study II corpus. This line of investigation seeks to identify nonverbal behavioral correlates of both affect and learning. The present results indicate that facial expression, gesture, and posture may have differing affective interpretations based on the tutoring context in which they occur. The nonverbal features were found to be more predictive than incoming student self-efficacy and pretest score. Additionally, the nonverbal features were largely contingent upon student work on the programming task, illustrating that these moments of student task activity may be most salient to affect.

9.1. Multimodal Features

The tutoring session database logs were combined with automated facial action unit tracking on webcam videos (Chapter 4) and gesture and posture tracking across Kinect depth image frames (Chapter 6). Thus, the multimodal data streams consist of student task actions, facial expression, gesture, and posture.

The posture tracking distances were processed to extract features capturing postural positions and amount of postural movement. The median head distance of students at each workstation was selected as the "mid" postural position. Distances at one standard deviation (or more) closer or farther than "center" were labeled as "near" or "far," respectively. Additionally, postural movements were identified based on acceleration of the head tracking point. The absolute sum of frame-to-frame acceleration was accumulated in a rolling one-second window at each frame. The average amount of acceleration in a one-second interval was computed across all students. If acceleration in the present interval was above average, it was marked as a postural movement (PosMove). Average frequencies of gesture and posture features are shown in Table 9.1. Students tended to spend more time in a MID postural position and most frequently did not display a hand-to-face gesture. Additionally, students moved

less than average during each interval, indicating that there were short moments of high movement that raised the average.

Table 9.1. Average frequency of gesture and posture features

Feature	Avg. Freq.	Feature	Avg. Freq.
NEAR	15%	ONEHAND	16%
MID	68%	TWOHANDS	5%
FAR	17%	NOGESTURE	79%
POSMOVE	29%		
NOMOVE	71%		

The automatically recognized nonverbal behaviors were combined with task-related features in order to form the multimodal tutoring corpus. As students worked on programming tasks, the database logged dialogue messages, typing, and task progress. Tutorial dialogue occurred at any time during the sessions, with student and tutor messages sent asynchronously (STUDENTMSG and TUTORMSG, respectively). As a student completed the programming task, he or she would also press a compile button to convert the Java program code into a format that is ready to run. These compile attempts may be successful (COMPILESUCCESS) or fail due to an error in the program code (COMPILEERROR). The student would also run his or her program (RUNPROGRAM) in order to test the output and interact with it. In parallel with the task events described above, the database logged whether the student was typing at any given moment. The student may not be typing anything (NOTTYPING), working on the program code (CODING), or typing a message to the tutor (TYPINGMSG) at each moment. Additionally, the student was considered to have paused on the task if he or she had made changes to the program and then stopped. This sort of break may be due to the student having resolved the current task, taking a moment to think, or going off-task; therefore, it was

introduced as a task event (TASKPAUSE). The average frequency of each task event and typing status is shown in Table 9.2. The majority of time intervals occurred after tutor messages and when students were not typing. These majority events represent moments when the student may have been reading the task description or reflecting on tutor messages. Tutors were also more active in the dialogue than students, resulting in more time following tutor messages.

Table 9.2. Average frequency of task events and typing status

Task Event	Avg. Freq.	Typing Status	Avg. Freq.
COMPILEERROR	1.7%	CODING	15%
COMPILESUCCESS	2.1%	TYPINGMSG	12%
RUNPROGRAM	7.9%	NOTTYPING	73%
STUDENTMSG	26.4%		
TUTORMSG	53.1%		
TASKPAUSE	8.8%		

Task events and typing statuses were combined with nonverbal behaviors at one-second intervals across each tutoring session. The most recent event of a given type (nonverbal, task, typing) was counted as the current value at each interval. For instance, if a student had been typing but stopped after half a second into the current interval, the typing status would be assigned to NOTTYPING.

A tutoring session excerpt is shown in Figure 9.1. The excerpt shows a rich set of nonverbal behaviors occurring around student work on the programming task. This student produced a variety of facial expressions, particularly when examining and testing the Java program. Additionally, the student performed a one-hand-to-face gesture prior to compiling the program. The corresponding multimodal features for a

segment of the excerpt are shown in Figure 9.2. The multimodal feature vectors cover a twelve-second segment from the excerpt.

26:54	Tutor:	ready?
26:59	Student:	yes! [*Student starts coding*]
28:02	Student:	TASKPAUSE [*Student stops coding*]
28:03	Student:	GESTURE: ONEHANDTOFACE; FACE: AU2 & AU14
28:12	Student:	TASK: COMPILESUCCESS; FACE: AU2
28:14	Student:	FACE: AU14
28:17	Student:	TASK: RUNPROGRAM; FACE: AU1
28:19	Student:	FACE: AU7
28:21	Tutor:	excellent

Figure 9.1. Tutoring session excerpt

	28:08	28:09	28:10	28:11	28:12	28:13	28:14	28:15	28:16	28:17	28:18	28:19
AU1										AU1		
AU2	AU2				AU2							
AU4												
AU7												AU7
AU14	AU14						AU14					
ONEHAND	ONEHAND											
TWOHAND												
POSTURE	FARPOSTURE											
POSMOVE												
TASK	TASKPAUSE				COMPILESUCCESS					RUNPROGRAM		
TYPING												

Figure 9.2. Multimodal feature vectors for a twelve-second segment of tutoring: gray shading indicates presence of a nonverbal behavior, task event, or typing. Time is shown in minutes and seconds from the start of the tutoring session.

Relative frequencies of nonverbal behavior were calculated separately for task events and typing status. For instance, at each one-second time interval, AU1 was marked as present or absent. Each interval was associated with a task event, with frequency counts tabulated across all task events. The relative frequency of AU1 presence and absence was computed across these task-contingent counts. Thus, the percentages of time intervals occurring with specific task events and particular values of AU1 presence or absence sum to one hundred percent. For instance, one student may have AU1 after RUNPROGRAM 2.12% of the time and NOAU1 after RUNPROGRAM 3.24% of the time. These relative frequencies sum to one hundred percent when combined with the remainder of task-contingent relative frequencies of AU1. Relative frequencies were similarly computed across typing statuses for each nonverbal behavior. Thus, the relative frequencies account for the percent of time in which a student displayed a nonverbal behavior after a specific task event or during a particular typing status (i.e., a student with a 5% relative frequency of ONEHAND after TUTORMSG in a thirty minute session would have displayed a one-hand-to-face gesture for a total of ninety seconds after tutor messages). This resulted in a set of one hundred and sixty-two nonverbal features contingent upon task events and typing statuses. The distribution of these multimodal features across nonverbal behaviors, task events, and typing statuses is shown in Table 9.3.

Table 9.3. Counts of multimodal features across nonverbal behaviors, task events, and typing statuses

	Task Event	Typing Status
AU1	12	6
AU2	12	6
AU4	12	6
AU7	12	6
AU14	12	6
GESTURE	18	9
POSTURE	18	9
POSMOVE	12	6

9.2. Predictive Models

Fine-grained analyses of multimodal affective expressions are enabled by automated tracking of nonverbal behavior. Such analyses have the potential to reveal previously undiscovered ways in which affective displays relate to task performance, learning, and affective outcomes within a tutoring context. For instance, the same affective expression may have different causes depending on the tutoring context. As a first step toward examining the fine-grained tutoring context of learner affective displays, predictive models of affective and learning outcomes were constructed using the multimodal tutoring corpus, in which facial expression, gesture, and posture are combined with task actions.

Initial feature selection was performed using model averaging in JMP statistical software, which created regression models for all possible combinations of predictive variables (Symonds & Moussalli, 2010). Model averaging was used to identify and remove weakly predictive variables across all models. Specifically, the twenty most predictive variables were selected using the average coefficient estimate from models with one, two, or three predictive variables. The predictive models were then constructed using minimum Bayesian Information Criterion (BIC) in forward stepwise

linear regression. These models are conservative in how they select predictive features because the explanatory value of added parameters must offset the BIC penalty for model complexity. Thus, model averaging was used to identify the most generally predictive variables, while minimum BIC was used to constrain model complexity. Tutoring outcomes (*engagement*, *frustration* and *learning*) were the dependent variables. All variables were standardized (i.e., centered on the mean and scaled to unit standard deviation) to enable comparison. The predictive models shown in the following sections have been constructed using the entire corpus, with associated regression coefficients and R^2 values. Additionally, leave-one-out cross-validated R^2 values were computed using the same predictive variables (but different coefficients in each fold) to examine generalizability of the predictive models.

9.2.1 Predicting Engagement

Each student's Engagement score was the sum of the Focused Attention, Felt Involvement, and Endurability sub-scales in the User Engagement Survey (O'Brien & Toms, 2010) administered following the tutoring session. This model only uses self-reports of engagement from students who fully completed the User Engagement Survey (N=61). The predictive model of *engagement* was composed of three features, including students' incoming computer science self-efficacy, one-hand-to-face gestures after successful compile, and brow lowering (AU4) after sending a student dialogue message. Each of the nonverbal features explains more variance than the trait-based feature of computer science self-efficacy. This seems to indicate that state-based nonverbal features are more indicative of *engagement*. The cross-validated model effect size was $r = 0.39$. The model is shown in Table 9.4.

Table 9.4. Stepwise linear regression model for Engagement

Engagement =	Partial R^2	Model R^2	p
0.31 * ONEHAND after COMPILESUCCESS	0.10	0.10	0.009
-0.31 * AU4 after STUDENTMSG	0.09	0.19	0.008
0.27 * Computer Science Self-Efficacy	0.07	0.26	0.020
~0 (intercept)			0.959
RMSE = 0.88 standard deviations in Engagement			
Leave-One-Out Cross-Validated R^2 = 0.15			

9.2.2 Predicting Frustration

The Frustration Level scale from NASA-TLX (Hart & Staveland, 1988) was the student's retrospective self-report of how insecure, agitated or upset he or she was during the tutoring session. The predictive model of *frustration* included students' incoming general self-efficacy and two features that accounted for the absence of nonverbal behavior. The sole feature predictive of higher *frustration* corresponded with compile errors, which intuitively may be frustrating. The absence of brow lowering (AU4) after running the Java program reinforces a prior result that indicated AU4 as a marker of *frustration* (Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013b). Also, students with higher general self-efficacy tended to have less *frustration*, as represented in the model. The cross-validated model effect size was $r = 0.41$. The model is shown in Table 9.5.

Table 9.5. Stepwise linear regression model for Frustration

Frustration =	Partial R^2	Model R^2	p
-0.42 * General Self-Efficacy	0.14	0.14	0.004
-0.56 * NoAU4 after RunProgram	0.08	0.22	0.004
0.42 * NoGesture after CompileError	0.08	0.30	0.011
~0 (intercept)			1.000
RMSE = 0.85 standard deviation in Frustration Level			
Leave-One-Out Cross-Validated R^2 = 0.17			

9.2.3 Predicting Learning Gain

Normalized learning gain measures how much a student learned relative to what he or she could have learned (Marx & Cummings, 2007). This accounts for relative differences in learning between students who scored high or low on the pretest. Normalized learning gain was computed using the following formula if posttest score was greater than pretest score:

$$NLG = \frac{Posttest - Pretest}{1 - Pretest}$$

Otherwise, normalized learning gain was computed as follows:

$$NLG = \frac{Posttest - Pretest}{Pretest}$$

The predictive model of normalized learning gain is the only one of the three to include postural features. These features indicate that MID and FAR postural positions are predictive of learning, though whether they are positive or negative predictors is dependent upon the tutoring context. Mouth dimpling (AU14) after running the Java program was predictive of learning. This supports a prior result that AU14 is positively associated with learning (Grafsgaard et al., 2013b). Finally, general self-efficacy predicted higher learning gains. The cross-validated model effect size was $r = 0.62$. The model is shown in Table 9.6.

Table 9.6. Stepwise linear regression model for Normalized Learning Gain

Norm. Learning Gain =	Partial R^2	Model R^2	p
0.10 * AU14 after RUNPROGRAM	0.11	0.11	0.004
0.10 * General Self-Efficacy	0.08	0.19	0.002
-0.12 * MIDPOSTURE after COMPILEERROR	0.08	0.27	<0.001
-0.21 * FARPOSTURE during CODING	0.04	0.31	<0.001
0.20 * FARPOSTURE after COMPILESUCCESS	0.18	0.49	<0.001
0.43 (intercept)			<0.001
RMSE = 0.24 std. dev. in Normalized Learning Gain			
Leave-One-Out Cross-Validated R^2 = 0.38			

9.3. Discussion

The results demonstrate that nonverbal behaviors at specific moments in the tutoring session are predictive of *engagement*, *frustration*, and *learning*. The combination of task events, typing, and nonverbal behaviors in multimodal features is predictive beyond incoming student characteristics, such as pretest score and self-efficacy. Additionally, the affective valence (positive or negative) of the nonverbal behaviors depended upon the tutoring context in which they occurred.

The predictive model of *engagement* was composed of three features, including students' incoming computer science self-efficacy, one-hand-to-face gestures after successful compile, and brow lowering (AU4) after sending a student dialogue message. One-hand-to-face gestures may have different affective valence depending on the physical position of the hand. A student may rest his/her head on the hand as a sign of boredom (Baker et al., 2010), or touch his/her chin in a moment of contemplation (Mahmoud & Robinson, 2011). Here, one-hand-to-face gestures after compile success were predictive of higher post-session self-report of *engagement*. This may coincide with student focus on the programming task. In the moments after updating the program code and compiling it, the student is no longer typing and may then reflect on current progress. Brow lowering (AU4) after the student sends a dialogue message, on

the other hand, was a predictor of lower *engagement*. This may indicate that a student is having difficulty with the subject matter, most likely responding to a tutor message (in this corpus, tutor messages were predominant and students rarely took initiative in the dialogue). Both of the nonverbal features were more predictive than computer science domain-specific self-efficacy, which was associated with greater *engagement*.

Frustration was significantly predicted by general self-efficacy. Higher levels of general self-efficacy coincided with lower post-session reports of *frustration*. Students with higher general self-efficacy are more confident in their ability to complete difficult tasks and therefore may be less intimidated by a novel learning task. However, inclusion of two nonverbal features doubled the explanatory power of the model. Each of the nonverbal features captured absence of nonverbal behaviors after specific task events. Absence of brow lowering (AU4) after running the Java program was predictive of lower *frustration*. At this point, the student is testing the program to see whether it matches his/her expectation. A prior result on this tutoring corpus found that AU4 was an indicator of *frustration*. Therefore, the present result supports that finding, but also provides a specific tutoring context (running the program) that is particularly meaningful for *frustration*. The sole feature predictive of higher *frustration* corresponded with compile errors (which occur when the program is incorrect). This correspondence between compiling the program and frustration is similar to results of prior analyses of student emotions during computer programming (Bosch, D'Mello, & Mills, 2013; Lee, Rodrigo, Baker, Sugay, & Coronel, 2011). Not all students had compile errors, so this feature represents those students who may have found the task to be more difficult. The absence of gestures after compile errors may be due to swift tutor interventions to remediate problems with the program. In this case, a student may feel frustrated due to overly active tutoring strategies.

Normalized learning gain was predicted by a combination of students' incoming general self-efficacy, mouth dimpling (AU14) after running the program, and three posture-related features. The model shows that students with more general confidence

in their ability to complete novel and difficult tasks tended to learn more than their peers. Displays of AU14 after running the program also were predictive of higher learning gain. Two aspects of AU14 discovered in prior results may shed light on this. First, occurrence of AU14 in general was associated with greater learning gain (Grafsgaard, Wiggins, Boyer, Wiebe, & Lester, 2013a). Second, AU14 in the first five minutes of tutoring was correlated with higher *frustration*, while AU14 in the last five minutes of tutoring was correlated with greater learning gain (Grafsgaard et al., 2013b). Running the program occurs most frequently during the later portion of the session. So, AU14 displays after running the program may also occur toward the end. With this timing-related interpretation, it may be that continued mental effort throughout the tutoring session is reflected in displays of AU14. Further study of AU14 may confirm whether it is generally an indicator of mental effort.

The posture-related features included both MID and FAR distances. These postural positions may encode information beyond whether a student is sitting at a certain distance from the computer. For instance, when a student is sitting at MID distance, the shoulders may be hunched or the student may have a straight back. FAR postural position was both predictive of higher learning gain (when occurring after compile success) and lower learning gain (when present during coding). It may be that bored students slouched in a FAR position during coding, while relaxed (but active) students were similarly farther back. New tracking methods may be developed to disambiguate these subtleties of posture. Interestingly, postural position was predictive of learning, but moment-to-moment postural movement was not. Discretization across one-second intervals may not have adequately captured brief postural movements.

The predictive models largely include nonverbal features that occur around moments of student work on the programming task. These may be pivotal moments on a student's path to learning, as students are actively working on the task and confirming whether the program works as intended. Prior results in analysis of skin conductance on this tutoring corpus showed that students' physiological responses to compile

attempts and failures were associated with *learning* and *frustration* (Hardy, Wiebe, Grafsgaard, Boyer, & Lester, 2013). The predictive models presented in this chapter further underscore the importance of tutoring context in interpretation of nonverbal behavior.

Chapter 10. Comparison of Multimodal Feature Sets

This chapter presents the first analysis comparing how well multimodal feature sets predict whole-session retrospective self-reports of affect and learning gain within human-human tutoring. Multimodal feature sets were constructed from input streams of dialogue, nonverbal behavior, and task actions in the JavaTutor Study II corpus. Thus, the nonverbal behavior input streams included automatically tracked facial expression, hand-to-face gestures, and posture. Unimodal, bimodal, and trimodal feature sets (i.e., consisting of combined features from one, two, or three modalities) were used to predict retrospective self-reports of engagement and frustration during the tutoring session and learning gain. The complete trimodal feature set was most predictive of each of the three tutoring outcomes. Among bimodal feature sets, dialogue-based feature sets were most predictive of each tutoring outcome. Importantly, the findings demonstrate that the role of nonverbal behavior may depend on the dialogue and task context in which it occurs. These results provide a promising direction for investigating multimodal feature sets in affective tutorial interaction.

10.1. Multimodal Feature Sets

The multimodal data streams in the JavaTutor Study II corpus were used in these analyses. As students worked on programming tasks, the database logged dialogue messages, typing, and task progress. Tutorial dialogue occurred at any time during the sessions, with student and tutor messages sent asynchronously (StudentMessage and TutorMessage, respectively). Each of these student and tutor messages has an associated dialogue act label (Appendix F: Fine-Grained Dialogue Acts) given by a dialogue act classifier constructed on this corpus (Vail & Boyer, 2014a). These dialogue events comprise the Dialogue data stream.

The Task data stream consists of student task actions. As a student completed the programming task, he or she would press a compile button to convert the Java

program code into a format that is ready to run. These compile attempts may be successful (COMPILESUCCESS) or fail due to an error in the program code (COMPILEERROR). The student would also run his or her program (RUNPROGRAM) in order to test the output and interact with it. The student may also be working on the program code (CODING), or have stopped coding (STOPCODING) at each moment.

The NONVERBAL data stream consists of student facial expression, hand-to-face gestures, and posture. Each of the nonverbal behaviors were tracked at all times, so they were combined in parallel (i.e., each interval records the presence or absence of facial expression, gesture, and posture).

These data streams were discretized using one-second intervals. The most recent event of a given type (DIALOGUE, NONVERBAL, TASK) was used as the current value at each interval. For instance, if a student had been coding but stopped after half a second into the current interval, the task action would be assigned to STOPCODING. These one-second time intervals were used to calculate relative duration following a specific dialogue event or task action, or during a particular nonverbal behavioral display. Thus, each possible feature has a single numerical value (relative duration) for each student session. The simplest feature sets constructed in this way (the *unimodal* sets) consist of a single data stream. The average relative durations of each feature in the unimodal TASK feature set are shown in Table 10.1.

Table 10.1. Average relative duration in the TASK feature set

	Avg. Relative Duration (%)
CODING	18
STOPCODING	25
COMPILESUCCESS	9
COMPILEERROR	4
RUNPROGRAM	44

The *bimodal* feature sets each consist of the Cartesian product of two unimodal feature sets. This resulted in three bimodal feature sets: DIALOGUE × NONVERBAL, DIALOGUE × TASK, and NONVERBAL × TASK. Similarly, the complete *trimodal* feature set consists of the Cartesian product of all three unimodal feature sets, DIALOGUE × NONVERBAL × TASK. A final feature set combined all three bimodal feature sets through set union. This BIMODAL UNION feature set allows for comparison of the combined bimodal feature sets versus the complete trimodal feature set.

10.2. Multimodal Feature Analysis

This section presents two levels of analysis: 1) comparison of feature set performance in predicting tutorial outcomes; and 2) comparison of the most predictive features across feature sets. These were performed for each of the three tutoring outcomes: engagement, frustration, and learning gain.

10.2.1 Model Construction

Model averaging was used to identify the most generalizable features in models with bimodal or trimodal feature sets and to reduce the feature space (Symonds & Moussalli, 2010). This approach produces average coefficient estimates and standard error across a wide range of models. Ratios of absolute value of the coefficient estimate versus standard error were also computed. These ratios provided a tradeoff between predictive weight and numerical stability, as estimates with lower standard error varied less across models. The features were then sorted using the ratios and the top twenty were selected for use in model building. (In the case of unimodal feature sets, all features were used and model averaging was not performed.) Predictive models were built using forward stepwise linear regression. Features were selected to optimize the leave-one-out cross-validated R^2 value of each model.

10.2.2 Engagement

The predictive models of engagement across feature sets, shown in Table 10.2, display a clear advantage of the trimodal feature set combining dialogue, nonverbal behavior, and task actions (D×N×T). The next best model uses the bimodal feature set combining dialogue and task actions (D×T). The union of bimodal feature sets (BIMODAL UNION) performs worse than D×T, but it uses fewer parameters. The trimodal feature set explains 50% more variance in retrospective self-reported engagement compared to the next best model (D×T) and uses fewer parameters.

Table 10.2. Engagement feature sets

Feature Set	R^2	# parameters
Unimodal Feature Sets		
DIALOGUE	0.048	3
NONVERBAL	-0.030	2
TASK	0.006	4
Bimodal Feature Sets		
DIALOGUE × NONVERBAL	0.146	3
DIALOGUE × TASK	0.187	9
NONVERBAL × TASK	0.112	5
Trimodal Feature Sets		
BIMODAL UNION	0.157	7
DIALOGUE × NONVERBAL × TASK	0.282	6

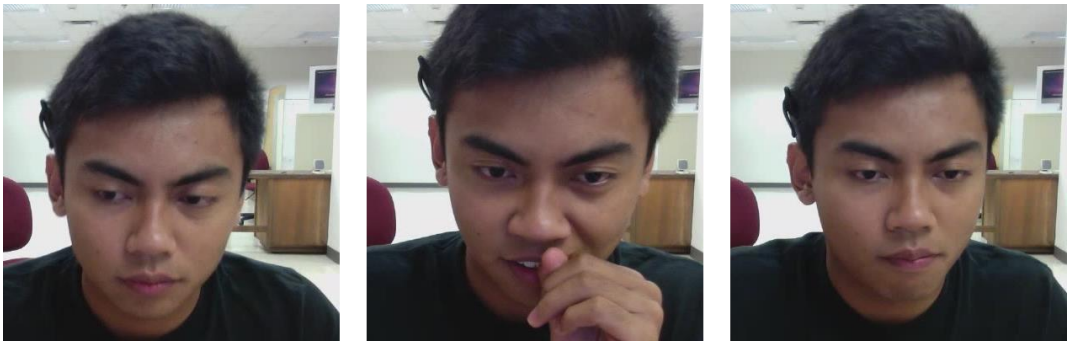
The two predictive models of retrospective self-reported engagement incorporating all three modalities of dialogue, nonverbal behavior, and task actions are shown in Table 10.3. There is slight overlap between the models (features that occur in both models are in bold type). The combination of student observational statements (O-STUDENT) and stopping coding was a negative predictor of engagement in both models. This may indicate moments when the student stopped working on the task in order to

make a comment, which is consistent with loss of focus on the task. Mouth dimpling (AU14) was also involved in negatively predictive features in both models. This facial action unit has been associated with frustration and mental effort in prior analyses (Grafsgaard et al., 2013a; Littlewort, Bartlett, et al., 2011). Open questions from the tutor (QO-TUTOR) with no student brow lowering (NoAU4) were positively predictive of engagement in the model with the combined bimodal feature set (Bimodal Union). Brow lowering has been previously associated with confusion, frustration, and mental effort (D'Mello et al., 2014; Grafsgaard et al., 2013a, 2013b; Littlewort, Bartlett, et al., 2011). Thus, this may highlight moments when the student was not perplexed by a tutor question. In the model with the trimodal feature set, student informational questions (QI-STUDENT) with outer brow raising (AU2) and running the program were positively predictive of engagement. This may indicate some interest in how the program operates, with the student actively engaging in discussion with the tutor.

Table 10.3. Engagement feature comparison

Feature	β	p
BIMODAL UNION, R^2 = 0.157		
O-STUDENT, **STOPCODING**	-0.310	0.002
QEX-TUTOR, **AU14**	-0.283	0.005
GRE-STUDENT, COMPILESUCCESS	-0.275	0.007
FO-TUTOR, **AU2**	-0.247	0.061
FO-TUTOR, AU7	-0.195	0.137
QO-TUTOR, NOAU4	0.269	0.008
INTERCEPT	0.003	1
DIALOGUE × NONVERBAL × TASK, R^2 = 0.282		
AEX-STUDENT, **AU14**, RUNPROGRAM	-0.333	0.003
O-STUDENT, AU1, **STOPCODING**	-0.201	0.082
TYPINGMESSAGE, AU4, CODING	-0.197	0.110
AYN-TUTOR, POSNEAR, **STOPCODING**	0.256	0.014
QI-STUDENT, **AU2**, RUNPROGRAM	0.707	<0.001
INTERCEPT	0.076	1

A sequence of tutoring events related to engagement is shown in Figure 10.1. The left image shows the student coding the program. The middle shows when the student had stopped coding and read a yes/no answer from the tutor (AYN). In the right image, the student has returned to coding the program.



CODING | AYN-TUTOR, POSNEAR, STOPCODING | CODING

Figure 10.1. Sequence related to engagement

10.2.3 Frustration

A comparison of predictive models of retrospective self-reported frustration across unimodal, bimodal, and trimodal feature sets is shown in Table 10.4. As with the models of engagement, there is a clear advantage of the trimodal feature sets. The trimodal feature sets greatly improve upon the bimodal feature sets, with the complete trimodal feature set (DIALOGUE × NONVERBAL × TASK) explaining more than three times the variance of the best bimodal feature set (DIALOGUE × TASK). The union of bimodal features (BIMODAL UNION) also explains over two times the variance of the individual bimodal feature sets, demonstrating an additive effect across the combined feature sets.

Table 10.4. Frustration feature sets

Feature Set	R^2	# parameters
Unimodal Feature Sets		
DIALOGUE	-0.033	1
NONVERBAL	-0.010	2
TASK	-0.033	1
Bimodal Feature Sets		
DIALOGUE × NONVERBAL	0.019	2
DIALOGUE × TASK	0.137	2
NONVERBAL × TASK	0.134	5
Trimodal Feature Sets		
BIMODAL UNION	0.347	7
DIALOGUE × NONVERBAL × TASK	0.520	5

Both predictive models of retrospective self-reported frustration that include all three modalities of dialogue, nonverbal behavior, and task actions are shown in Table 10.5. Interestingly, both models provide only positive predictors of frustration. There is also a fair degree of overlap between the models (shared features are indicated in bold). In both models, higher frustration is predicted by students providing feedback on their current understanding (FU) when working on the program. Students often did this after receiving tutor input on how the program works. So, this type of dialogue act may confirm that they understood the new information given by the tutor. However, the underlying trend may be that students had just received help from the tutor in order to remove a misconception. Thus, the students may feel frustration partially due to the lingering misconception.

Table 10.5. Frustration feature comparison

Feature	β	p
BIMODAL UNION, R^2 = 0.347		
FNU-STUDENT, **POSNEAR**	0.141	0.210
FNU-STUDENT, AU14	0.202	0.042
QEX-TUTOR, **STOPCODING**	0.260	0.013
STUDENTMESSAGE, RUNPROGRAM	0.260	0.008
OEX-TUTOR, POSMOVE	0.323	0.003
FU-STUDENT, **CODING**	0.456	<0.001
INTERCEPT	0	1
DIALOGUE × NONVERBAL × TASK, R^2 = 0.520		
R-TUTOR, NOAU1, **STOPCODING**	0.091	0.549
FU-STUDENT, NOAU2, **CODING**	0.336	<0.001
OEX-TUTOR, AU7, **STOPCODING**	0.429	0.006
FU-STUDENT, **POSNEAR**, **CODING**	0.451	<0.001
INTERCEPT	0	1

Another prominent pattern in both models is the emphasis on moments when the student was coding the program or had stopped coding. The moments of coding aligned with student feedback on his or her current level of understanding (FU-STUDENT). On the other hand, moments when the student stopped coding were associated with tutor messages. These tutor messages were related to off-topic statements (OEX) or questions (QEX), and reassurance (R). In such instances, the tutor is focusing on off-task discussion. Therefore, these features may be related to moments when the student stopped coding and received off-task tutor messages (with causality in either direction).

A sequence of tutoring events related to frustration is shown in Figure 10.2. First, the student encounters a compile error. The tutor then directs the student on how to fix the problem. Second, the student reports her understanding (FU) of the directive and implements the solution. Third, the tutor tells the student to compile and she does so successfully.

COMPILEERROR | FU-STUDENT, POSNEAR, CODING | COMPILESUCCESS

Figure 10.2. Sequence related to frustration

10.2.4 Normalized Learning Gain

Normalized learning gain (or percent learning gain) measures how much a student learned relative to what he or she could have learned (Marx & Cummings, 2007). This accounts for relative differences in learning between students who scored high or low on the pretest. Normalized learning gain was computed with the following formula if posttest score was greater than pretest score:

$$NLG = \frac{Posttest - Pretest}{1 - Pretest}$$

Otherwise, normalized learning gain was computed as follows:

$$NLG = \frac{Posttest - Pretest}{Pretest}$$

The predictive models of normalized learning gain are shown in Table 10.6, containing all three modalities of dialogue, nonverbal behavior, and task actions. In contrast with the affective tutoring outcomes, a unimodal feature set has significant predictive power, as the DIALOGUE unimodal feature set is predictive of learning. This underscores the important role of tutorial dialogue in the process of learning. Despite

the significant predictive power of the best bimodal feature set (DIALOGUE × NONVERBAL), the complete trimodal feature set (D×N×T) still explains around sixteen percent more variance in learning gains.

Table 10.6. Normalized Learning Gain feature sets

Feature Set	R^2	# parameters
Unimodal Feature Sets		
DIALOGUE	0.370	10
NONVERBAL	0.037	3
TASK	-0.034	1
Bimodal Feature Sets		
DIALOGUE × NONVERBAL	0.465	6
DIALOGUE × TASK	0.407	13
NONVERBAL × TASK	0.243	10
Trimodal Feature Sets		
BIMODAL UNION	0.460	5
DIALOGUE × NONVERBAL × TASK	0.544	8

The trimodal feature comparison models of normalized learning gain are shown in Table 10.7. These models overlap on the mouth dimpling facial action unit (AU14), which is a negative predictor of learning in both models. In viewing the tutoring session videos, AU14 appeared to correspond with moments when students were expending mental effort or thinking about the task. In the BIMODAL UNION model, AU14 coincides with tutor extra-domain questions (QEX). In the complete trimodal model (D×N×T), AU14 co-occurs with students running the program and remarking on lack of understanding (FNU) in one trimodal feature, and with student stopping coding and tutor answers to complicated questions (AWH: not a "yes" or "no" answer) in another trimodal feature. In each case, the presence of AU14 may involve a reaction to a recent

event in the tutoring session, whether it is the behavior of the program or messages from the tutor.

Table 10.7. Normalized Learning Gain feature comparison

Feature	β	p
BIMODAL UNION, $R^2 = 0.460$		
O-STUDENT, **STOPCODING**	-0.478	<0.001
QEX-TUTOR, **AU14**	-0.341	0.001
GRE-STUDENT, **COMPILESUCCESS**	-0.314	0.002
FO-TUTOR, AU2	0.295	0.002
INTERCEPT	0.002	1
DIALOGUE × NONVERBAL × TASK, $R^2 = 0.544$		
FNU-STUDENT, **AU14**, RUNPROGRAM	-0.415	<0.001
AWH-TUTOR, **AU14**, **STOPCODING**	-0.311	0.001
E-TUTOR, POSMOVE, CODING	-0.213	0.019
FU-STUDENT, AU1, **COMPILESUCCESS**	-0.171	0.053
E-STUDENT, NOMOVE, RUNPROGRAM	0.132	0.160
ACK-TUTOR, NOAU4, CODING	0.205	0.028
QO-TUTOR, AU4, **STOPCODING**	0.231	0.011
INTERCEPT	0.002	1

A sequence of events related to learning is shown in Figure 10.3. In the first image, the student is testing his program. Further in the session, the student has begun coding more of the program. In the second image, the student has stopped coding and displayed AU4 as the tutor asked an open-ended question (QO). The third image shows the student at a farther postural distance after successfully compiling the program.

RUNPROGRAM | QO-TUTOR, AU4, STOPCODING | COMPILESUCCESS

Figure 10.3. Sequence related to normalized learning gain

10.3. Discussion

This chapter has presented an in-depth comparison of multimodal feature sets related to engagement, frustration, and learning in computer-mediated human tutoring. The results show a distinct additive effect of features across modalities of dialogue, nonverbal behavior, and task actions. Each set of models found an improvement from unimodal to bimodal features and from bimodal to trimodal features.

Prior work has demonstrated mixed results in applying multimodal feature sets to prediction of affect. Often, a particular feature set is useful for one affective state, but not another (D'Mello & Kory, 2012). However, most prior multimodal studies of tutoring did not involve a strong dialogue component (Arroyo et al., 2009; Cooper, Muldner, Arroyo, Woolf, & Burleson, 2010; Kapoor & Picard, 2005). In the present multimodal tutoring corpus, student-tutor dialogue plays a very significant role in predicting tutoring outcomes. A majority of features involve dialogue in the models built on a union of bimodal features, which had the potential to select nonverbal behavior and task features instead. Additionally, dialogue was the only unimodal feature set that was strongly predictive of a tutoring outcome—in this case, normalized

learning gain. The importance of adaptive dialogue has also been shown in studies that examined the advantages of one-to-one tutoring (VanLehn et al., 2007).

While dialogue was of primary importance in these findings, the nonverbal behavior and task modalities also provided additional explanatory power. Task actions were fairly straightforward (e.g., the student was working on the task or not), but nonverbal behaviors co-occurred with specific task contexts. For instance, both presence of brow lowering (AU4) and its absence (NoAU4) appeared as positive predictors of normalized learning gain. The task contexts associated with these predictions were tutor utterance (QO vs. ACK) and student task actions (StopCoding vs. Coding). The first feature (QO-Tutor, AU4, StopCoding) describes a moment when a student has stopped coding, has been posed an open-ended question by the tutor, and is thoughtfully reflecting on the question (as evidenced by brow lowering). The second feature (ACK-Tutor, NoAU4, Coding), in contrast, may highlight a moment when the student is focused on implementing the program after receiving an acknowledgment from the tutor. AU4 may be absent in this context because the student knows how to modify the program and is making the changes with certainty. Similarly, a near postural position (PosNear) was predictive of both higher engagement and higher frustration. The divergent contexts of PosNear were student/tutor utterance (AYN-Tutor vs. FU-Student) and student task actions (StopCoding vs. Coding). In the first case (AYN-Tutor, StopCoding), it seems that the student may sit near and stop coding while reading the tutor answer, which may reflect focused concentration. In the second case (FU-Student, Coding), the student may be responding to tutor help and continuing work on the task, in which case the student may be having difficulty with the task, in turn associated with frustration.

Some nonverbal behaviors were more consistently predictive of tutoring outcomes. Mouth dimpling (AU14) appeared in multiple predictive features as an indicator of lower engagement, higher frustration, and reduced learning gain. Additionally, postural movement (PosMove) was associated with lower learning gain

and increased frustration. Despite the alignment of these nonverbal behaviors toward negative affect, it is important to note that they were conditioned upon specific tutorial contexts. Thus, further analyses are necessary to infer generalizable situations in which these nonverbal behaviors occur.

Chapter 11. Multimodal Differential Sequence Mining

This chapter presents an approach for sequential analysis of multimodal data streams. Prior chapters detailed advances in modeling multimodal data streams that provided evidence on the links between nonverbal behavior and cognitive-affective states, such as engagement and frustration. Additionally, Chapter 8 described an initial approach to using hidden Markov models to identify patterns of affective tutorial interaction. However, this approach suffered from the problem of dimensionality, wherein the large, multimodal feature space produced sparse models that did not capture salient patterns in the data. The approach presented in this chapter extends differential sequence mining to handle multimodal data streams, resulting in discovery of a small set of patterns that highlight differences in engagement, frustration, learning gain, and self-efficacy.

11.1. Sequence Mining on the JavaTutor Study II Corpus

The JavaTutor Study II corpus includes task actions and dialogue synchronized with automatically labeled facial expression and bodily expressions of posture and gesture. Additionally, an automated dialogue act classifier (Vail & Boyer, 2014a) was applied to the data to provide fine-grained annotations of tutor and student dialogue (Appendix F: Fine-Grained Dialogue Acts). The observations of nonverbal behavior were labeled as in the multimodal features described in Chapter 10.

In order to explore these multimodal data streams within the context of the computer programming task, the data streams were combined into sequential events on student task actions, which included START CODING, STOP CODING, COMPILE SUCCESS, COMPILE ERROR, and RUN PROGRAM. Each new task action was treated as a sequential interval. For example, a student may begin writing program code, with dialogue and displays of nonverbal behavior co-occurring with this work on the programming task. The dialogue and nonverbal behaviors that occurred during this task action would be collected into a set of observation symbols, referred to in sequence mining as an

itemset. When the next task action occurs (the student stops working on the program code), the co-occurring observation symbols (for dialogue and nonverbal behavior) would be added to the next itemset. Thus, the itemset representation encodes all observation symbols that co-occurred with each student task action. The full set of observation symbols across the multimodal sequences is shown in Appendix G: Multimodal Sequence Alphabet.

Various sequence mining implementations have been developed, and the authors of the original differential sequence mining implementation (Kinnebrew et al., 2013) used PexSPAM, which was created to extract protein sequence features (Ho, Lukov, & Chawla, 2005). The present work used the Sequential Pattern Mining Framework (SPMF) (Fournier-Viger et al., 2014), a Java library that implements numerous sequential pattern mining algorithms. In particular, the Fournier-Viger algorithm (Fournier-Viger et al., 2008) was used to enable sequential pattern mining over consecutive itemsets. However, this extension of differential sequence mining required a new implementation of instance support to operate over itemsets. Thus, part of the contribution of this dissertation is this novel implementation of instance support over itemsets, which will be released publicly along with a future publication.

The student participants were divided into groups based on their retrospective self-reports of engagement and frustration, their learning gains, and their pre-tutoring responses on general and computer science self-efficacy. Since the self-report measures were taken at the end of a session, these analyses were conducted on the multimodal sequences in the first lesson. The Fournier-Viger sequence mining algorithm was used to identify sequential patterns of length three in each student group. This pattern length was selected to enhance presentation and interpretability, but larger lengths could also be used. Each of the discovered patterns consists of consecutive events collected in task action itemsets, as described above. The sequential patterns were considered "*s*-frequent" if there was at least one occurrence for seventy-five percent of the students in the group (this percentage is known as the *s*-support, or sequence support, threshold).

After identifying frequent sequential patterns, the novel implementation of instance support was run to calculate how many times each pattern occurred for each student. Then, differential comparisons (*t*-tests) were run to identify whether the frequencies of each pattern were significantly different across two sub-groups of students. Thus, the differentially significant patterns (i.e., patterns that occurred at significantly different frequencies across two groups) were identified for each student measure.

In order to maximize interpretability of the results, only patterns with $p < 0.05$ were examined further and a small subset of those were selected for discussion. For each student task action, the pattern with the least number of symbols was chosen (there was no pattern selected when the $p < 0.05$ cut-off was not met). There are a total of ten possible patterns selected for each student measure (five task actions for both the high and low group), though fewer met the selection criteria. Each pattern *s*-frequent in the high group, low group, or both. Additionally, the average instance support (*i*-support) of the high group was subtracted from the average of the low group to show which group had more instances of each pattern. A positive difference in *i*-support means that the high group had more instances of the pattern on average, while the reverse is true with a negative difference. An overview of the multimodal differential sequence mining approach is shown in Figure 11.1.

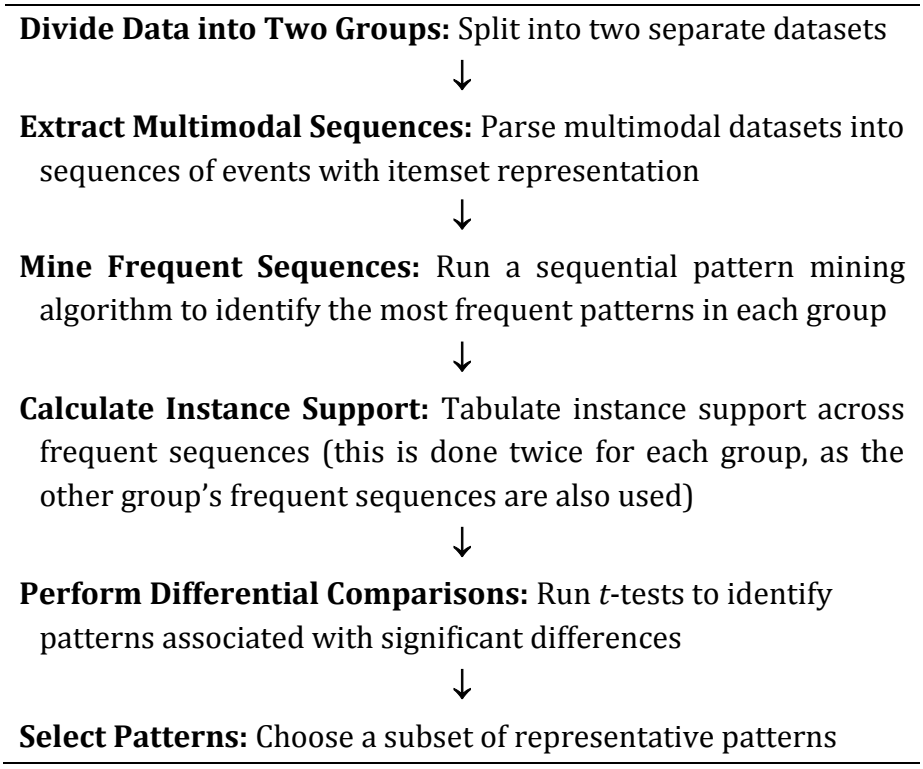


Figure 11.1. Overview of multimodal differential sequence mining

11.1.1 Sequential Patterns of Engagement

Students were divided into two groups based on the retrospective self-reports of engagement in the User Engagement Survey (see Appendix B: User Engagement Survey). A median split was performed, resulting in twenty-eight students in the high engagement group and thirty-three students in the low engagement group.

Five patterns were selected using multimodal differential sequence mining, with three occurring more frequently in the high group and two in the low group (Table 11.1). Two patterns that were more frequent with high engagement involved the one-hand-to-face gesture. Pattern E1 captured sequences where a student finished working on the program, compiled it successfully, and then displayed a one-hand-to-face gesture. Pattern E3 started with compile success, followed by a one-hand-to-face

gesture, then a postural shift. The remaining high engagement pattern, E2, began on student coding with a middle postural position and postural shift, followed by stopping coding and compile success. In contrast, the low engagement patterns (E4 and E5) involved facial expression movements, with a middle postural position and postural shift also included in E4. Overall, the high engagement patterns clearly coincided with one-hand-to-face gestures, while low engagement patterns contained facial expressions. This appears to be consistent with the interpretation of a contemplative, engaged student performing more one-hand-to-face gestures and a less engaged student performing more facial movements, which may indicate more affective reactions to the task.

Table 11.1. Differential patterns of Engagement (High N=28, Low N=33)

Engagement Pattern						s-Freq. Group	i-Sup. (Hi–Lo)	*p*
E1:	STOP CODING	→	COMPILE SUCCESS	→	ONEHANDTOFACE	Both	1.62	0.020
E2:	START CODING MID POSTURE POSTURE SHIFT	→	STOP CODING	→	COMPILE SUCCESS	Both	1.56	0.035
E3:	COMPILE SUCCESS	→	ONEHANDTOFACE	→	POSTURE SHIFT	Both	1.43	0.049
E4:	RUN PROGRAM FACE: AU14	→	FACE: AU1 FACE: AU2 FACE: AU14 MID POSTURE POSTURE SHIFT	→	FACE: AU1	Both	-0.71	0.032
E5:	STOP CODING	→	START CODING FACE: AU4	→	FACE: AU14	Both	-1.17	0.022

11.1.2 Sequential Patterns of Frustration

The post-session retrospective self-reports of frustration (Appendix C: NASA-TLX) were used to divide students into high and low frustration groups. A median split was

performed, resulting in thirty students in the high frustration group, and thirty-three students in the low frustration group.

Multimodal differential sequence mining was used to identify five patterns of frustration (Table 11.2). Four of the patterns (F1 through F4) were more frequent with high frustration. Two of those (F1 and F2) were *s*-frequent only in the high frustration group (i.e., at least seventy-five percent of high frustration students displayed the pattern). These two patterns included two-hands-to-face gestures during work on the program code. Similar to the patterns associated with low engagement in the previous section, patterns F3 and F4 involved facial movements. Additionally, F3 and F4 also included postural shifts. The sole pattern that was more frequent with low frustration started with a middle postural position and AU14 (Mouth Dimpler) after a compile success, followed by AU1 (Inner Brow Raiser) and then AU14 again. These results showed that two-hands-to-face gestures more often coincided with high frustration, and that postural shifts were associated with high frustration, while AU14 and a middle postural position were more frequent with low frustration.

Table 11.2. Differential patterns of Frustration (High N=30, Low N=33)

Frustration Pattern						s-Freq. Group	i-Sup. (Hi–Lo)	*p*
F1:	START CODING POSTURE SHIFT TWOHANDSTOFACE	→	STOP CODING	→	START CODING	High	1.68	<0.001
F2:	RUN PROGRAM	→	START CODING POSTURE SHIFT TWOHANDSTOFACE	→	STOP CODING	High	0.81	0.038
F3:	STOP CODING POSTURE SHIFT	→	FACE: AU1 FACE: AU4 FACE: AU14 POSTURE SHIFT	→	FACE: AU1 FACE: AU14	Both	1.07	0.022
F4:	START CODING POSTURE SHIFT	→	FACE: AU1 FACE: AU2	→	START CODING	Both	1.05	0.022
F5:	COMPILE SUCCESS FACE: AU14 MID POSTURE	→	FACE: AU1	→	FACE: AU14	Low	-0.98	0.040

11.1.3 Sequential Patterns of Normalized Learning Gain

Normalized learning gain accounts for how much knowledge a student gained or lost from pretest to posttest, relative to how much they could have gained (or lost). Students were divided into high and low learning groups using a median split, resulting in thirty high learning students and thirty-two low learning students.

There were six patterns selected through multimodal differential sequence mining (Table 11.3). The three patterns involved with high learning gain (L1 through L3) all involved one-hand-to-face gestures. Two of the patterns that are significantly more frequent with low learning gain (L5 and L6) included a large variety of facial movements. Middle postural positions and postural shifts were present in patterns associated with high and low learning. Thus, the most identifiable results were that one-hand-to-face gestures coincided with high learning and a large variety of facial expressions were associated with low learning.

Table 11.3. Differential patterns of Norm. Learning Gain (High N=30, Low N=32)

Normalized Learning Gain Pattern						s-Freq. Group	i-Sup. (Hi–Lo)	*p*
L1:	STOP CODING MID POSTURE	→	ONEHANDTOFACE	→	ONEHANDTOFACE	High	1.74	0.011
L2:	COMPILE SUCCESS ONEHANDTOFACE	→	POSTURE SHIFT	→	STOP CODING	High	0.75	0.008
L3:	START CODING	→	STOP CODING MID POSTURE	→	ONEHANDTOFACE	Both	1.46	0.049
L4:	RUN PROGRAM MID POSTURE	→	START CODING	→	POSTURE SHIFT	Both	-1.18	0.045
L5:	STOP CODING	→	MID POSTURE POSTURE SHIFT	→	FACE: AU1 FACE: AU7 FACE: AU14 MID POSTURE	Low	-1.16	0.030
L6:	START CODING FACE: AU1 FACE: AU2 FACE: AU4 FACE: AU14 MID POSTURE POSTURE SHIFT	→	STOP CODING FACE: AU1 FACE: AU2	→	FACE: AU2	Low	-1.01	0.047

11.1.4 Sequential Patterns of General Self-Efficacy

General self-efficacy measures how confident a student is in dealing with new and difficult tasks (Appendix D: General Self-Efficacy). This survey was given before the first tutoring session. Unlike the post-session retrospective self-reports, many students reported general self-efficacy at the median value. Therefore, students were divided into thirds with the middle excluded in order to produce high and low general self-efficacy groups. This resulted in eighteen students in the high general self-efficacy group and nineteen in the low self-efficacy group.

Eight patterns were identified through multimodal differential sequence mining (Table 11.4). Four of the patterns were associated with high general self-efficacy, while the other four were associated with low general self-efficacy. Notably, these patterns include dialogue acts, which were completely absent from the results for engagement, frustration, and normalized learning gain. Tutor positive feedback appears in three of

the patterns related to high general self-efficacy, which would presumably coincide with successful progress on the task. Additionally, a one-hand-to-face gesture appeared in pattern G3, which occurred more frequently with high general self-efficacy. In the low general self-efficacy patterns, tutor directives (telling the student which action to take) and tutor explanations indicate that the tutor felt a need to aid the student in the task. This result seems consistent with a student who is struggling and unfamiliar with the task.

Table 11.4. Differential patterns of General Self-Efficacy (High N=18, Low N=19)

General Self-Efficacy Pattern						s-Freq. Group	i-Sup. (Hi–Lo)	*p*
G1:	START CODING FACE: AU1 FACE: AU2	→	POSTURE SHIFT	→	POSTURE SHIFT TUTOR POS FDBK	High	1.53	0.036
G2:	COMPILE SUCCESS FACE: AU2	→	POSTURE SHIFT	→	POSTURE SHIFT	High	1.52	0.039
G3:	RUN PROGRAM	→	START CODING FACE: AU1 FACE: AU2 FACE: AU14 POSTURE SHIFT ONEHANDTOFACE TUTOR POS FDBK	→	STOP CODING	High	1.06	0.013
G4:	STOP CODING FACE: AU2	→	COMPILE SUCCESS	→	TUTOR POS FDBK	High	0.92	0.006
G5:	STOP CODING	→	TUTOR DIRECTIVE TYPING MESSAGE	→	FACE: AU1	Low	-0.44	0.035
G6:	RUN PROGRAM FACE: AU4	→	START CODING MID POSTURE POSTURE SHIFT TUTOR EXPLAIN	→	STOP CODING	Low	-0.54	0.007
G7:	START CODING	→	TYPING MESSAGE	→	POSTURE SHIFT	Low	-0.75	0.031
G8:	COMPILE SUCCESS MID POSTURE TUTOR EXPLAIN	→	MID POSTURE TUTOR EXPLAIN	→	TUTOR EXPLAIN	Low	-0.86	0.020

11.1.5 Sequential Patterns of Computer Science Self-Efficacy

Computer science self-efficacy measures a student's confidence in completing computer programming tasks and performing well in computer science coursework (Appendix E: Computer Science Self-Efficacy). As with general self-efficacy, the students were surveyed on computer science self-efficacy prior to the first tutoring session. Due to a large concentration of values around the median, students were divided into third, with the middle third excluded to form high and low groups. Thus, the high computer science self-efficacy group consisted of twenty students and the low computer science self-efficacy group contained nineteen students.

Multimodal differential sequence mining was used to find five patterns (Table 11.5). Two of the patterns were more frequent with high computer science self-efficacy (C1 and C2), while the other three were more frequent with low computer science self-efficacy (C3 through C5). Similar to the general self-efficacy patterns, tutor positive feedback appeared in a pattern associated with high computer science self-efficacy (C1) and tutor explanations were in all three patterns of low computer science self-efficacy. These dialogue-related results are very similar to those of general self-efficacy, perhaps further demonstrating positive feedback for good performance by confident students and more tutor aid provided to less confident students. A middle postural position appeared only in one of the high computer science self-efficacy patterns (C1), while postural shifts occurred in both high and low computer science self-efficacy patterns. This result related to a middle postural position is similar to that of frustration, in that it appeared in "positive" patterns more frequently associated with high computer science self-efficacy and low frustration.

Table 11.5. Differential patterns of C. S. Self-Efficacy (High N=20, Low N=19)

Computer Science Self-Efficacy Pattern						s-Freq. Group	i-Sup. (Hi–Lo)	*p*
C1:	COMPILE SUCCESS	→	FACE: AU1 TUTOR POS FDBK	→	MID POSTURE	High	0.95	0.023
C2:	COMPILE SUCCESS	→	START CODING POSTURE SHIFT	→	STOP CODING	Both	1.33	0.012
C3:	STOP CODING	→	TUTOR EXPLAIN	→	FACE: AU1	Both	-1.93	0.033
C4:	START CODING	→	STOP CODING	→	TUTOR EXPLAIN	Both	-2.06	0.036
C5:	RUN PROGRAM TUTOR EXPLAIN	→	POSTURE SHIFT TUTOR EXPLAIN	→	STOP CODING	Low	-1.05	0.002

11.2. Discussion

Multimodal differential sequence mining provides a new approach to analyzing multimodal data streams. This novel extension of differential sequence mining has been shown to handle numerous concurrent data streams with output of a small set of representative patterns. These patterns provide empirical evidence for how nonverbal behavior and dialogue intermingle within sequential contexts of task-oriented tutoring. Particularly, the results identified previously undiscovered sequential associations of affect and nonverbal behavior.

The results related to hand-to-face gestures were remarkably straightforward. One-hand-to-face gestures were present in patterns associated with high engagement, high learning, and high general self-efficacy. Two-hands-to-face gestures appeared in patterns related to high frustration. Both of these results provide novel empirical evidence for the importance of these hand-to-face gestures in task-oriented domains. Students who were confident, performed well, and were engaged in the task more frequently presented what may intuitively be thought of as a contemplative, one-hand-to-face gesture. On the other hand, students who experienced high frustration more frequently displayed the two-hands-to-face gesture, which may be an attempt to alleviate discomfort with the task. While these results cannot fully explain the internal

states associated with hand-to-face gestures, they do provide much needed insight into their associations with cognitive-affective phenomena.

The remaining results were less clear-cut. A greater variety of facial movements tended to appear in "negative" patterns, such as those related to low engagement, high frustration, and low normalized learning gain. The exceptions were in patterns associated with high general self-efficacy and low frustration. These facial expression results appear to highlight the importance of facial expression within the context of work on the task, as the patterns generally occurred in the context of writing program code or running the program. The facial action units that occurred in the negative patterns, but not in those of high general self-efficacy or low frustration, were AU4 (Brow Lowerer) and AU7 (Lid Tightener). Both of these involve tensing of the brow and eye region of the face that has been associated with confusion and frustration (D'Mello et al., 2009; Grafsgaard et al., 2013b). Further examination of the detailed task contexts in which these action units occur, with both presence and absence in mind, may serve to disambiguate these results.

The most difficult results to interpret were related to postural position and postural shifting. These appeared in both "positive" and "negative" patterns across each of the differential groups. Despite growing recognition of the importance of posture as a nonverbal display of affect, there is no cohesive theory of its role in affect. Further analyses may identify the fine-grained contexts needed to interpret the possibly multiple roles that posture plays.

Interestingly, dialogue was involved only in patterns related to general self-efficacy and computer science self-efficacy. These results were straightforward, with more frequent tutor positive feedback as confident students worked on the computer programming task. On the other side of the spectrum, tutors provided more explanation and direction to students with low general self-efficacy and/or low computer science self-efficacy. The lack of dialogue in the engagement, frustration, and learning gain

patterns shows that nonverbal behavior plays an important role in distinguishing affect, even (in this case) above that of dialogue.

Chapter 12. Conclusion

This dissertation addresses the research question:

How can automated approaches be used to identify multimodal sequences of positive or negative affect during tutoring?

As seen in prior chapters, a progression of techniques was applied to identify ways in which nonverbal behavior is associated with affect in tutoring. These techniques included statistical approaches, predictive models, hidden Markov models, and a novel implementation of multimodal differential sequence mining. While these efforts address this research question with significant contributions to the research literature, there is much yet to be examined and interpreted.

12.1. Hypotheses and Results

Developing and applying multimodal differential sequence mining on the JavaTutor Study II corpus produced evidence toward the hypotheses stated in Chapter 1:

H1. Multimodal sequences in the tutoring corpus are significantly associated with the positive and negative affect that students report at the end of the tutoring session, both in terms of *engagement* and *frustration*.

The results of multimodal differential sequence mining showed that there were multimodal sequences associated with engagement and frustration. One particularly interesting result was that one-hand-to-face gestures were associated with high engagement, while two-hand-to-face gestures more frequently occurred with high frustration. These results represent an advance toward understanding affect from patterns of behavior embedded in rich, multimodal data streams.

H2. Multimodal sequences differ significantly across student groupings based on incoming characteristics, specifically *computer science self-efficacy* and *general self-efficacy*.

Multimodal sequences related to general self-efficacy and computer science self-efficacy were also identified through multimodal differential sequence mining. Interestingly, the patterns associated with self-efficacy were heavily associated with tutorial dialogue. Specifically, the multimodal sequence patterns reflected the trend of tutors providing positive feedback to confident, high-performing students and giving explanations and directions to students with lower self-efficacy. These results show that multimodal differential sequence mining is a powerful, general technique that can be applied across cognitive, affective, and motivational phenomena.

12.2. Summary

Cognitive and affective processes intertwine during learning, comprising a rich layer of emotional experience. Affective states often influence progress on learning tasks, resulting in positive or negative cycles of affect that impact learning outcomes. Consequently, the influence of affective phenomena on learning has led to a recognized need to understand the occurrence, timing, and impact of cognitive-affective states during learning. Developing a clear understanding of these phenomena is critical to informing the design of affective tutorial interventions.

This dissertation addresses the fundamental problem of recognizing and understanding students' learning-centered affective states from nonverbal behavior during tutoring. Just as highly effective human tutors pay attention to more than whether a student is simply correct or incorrect, automated approaches may be used to identify and understand students' nonverbal behavior. Thus, this line of research considers posture, gesture, and facial expression as students engage in computer-mediated task-oriented dialogue with a human tutor.

In initial analyses, the Facial Action Coding System (FACS) was used to manually annotate students' facial expressions (Chapter 3). These annotations highlighted ways in which facial movements co-occurred with task progress and tutorial dialogue. For instance, brow lowering was more frequent when students had a mix of correct and incorrect programming progress, and also when students asked questions or received tutor feedback. While FACS annotation provides high quality labels, it is labor-intensive and does not scale to large datasets. Subsequent analyses leveraged automated approaches to identifying nonverbal behavior.

Subsequent larger scale analyses of facial expression were performed using the Computer Expression Recognition Toolbox (CERT), which automatically tracks numerous facial movements. In order to validate the use of CERT on naturalistic tutoring videos, the manual annotations were compared with CERT output (Chapter 4). This first-of-its-kind large-scale validation analysis showed that CERT has excellent agreement with manual annotations at an aggregate level. This approach was then applied to modeling of post-session affect self-reports and learning gains from facial expression (Chapter 5). These analyses identified associations of facial expression with learning-centered affective states, such as anxiety, engagement, and frustration.

In order to recognize a broader set of nonverbal behaviors, novel algorithms were developed to track posture and gesture from Kinect depth sensor recordings (Chapter 6). The posture tracking algorithm finds the distance of torso and head from the depth camera, while the gesture detection algorithm identifies whether one or two hands are in contact with the lower face. These tools were used to analyze aspects of embodied affect through posture and gesture (Chapter 7), such as how students shifted posture or performed gestures during the tutorial dialogue. These nonverbal behaviors were found to be indicative of engagement and frustration, with students shifting posture less when anticipating a tutor response or following up on task-related instruction.

These analyses of facial expression, gesture, and posture have highlighted ways in which learning-centered affect is presented through nonverbal behavior. Additional prior work has focused on modeling the tutorial context of nonverbal behavior. Hidden Markov models (HMMs) were used to analyze sequences of affective tutorial interaction (Chapter 8). Descriptive HMMs were machine-learned from the combined context of facial expression, tutorial dialogue, and task progress in an unsupervised approach using the Baum-Welch algorithm. These models showed how student facial expressions coincided with difficulty in the programming task or tutor dialogue modes, such as tutor-student collaboration or off-task discussion.

Subsequent work focused on multimodal representations of the task, dialogue, and nonverbal behavior data streams. Regression models were constructed using a conservative approach to identify multimodal features that represent salient moments of tutoring associated with cognitive and affective outcomes (Chapter 9). Then, multimodal feature sets were compared to identify the multimodal feature sets most predictive of tutorial outcomes (Chapter 10). The feature sets developed in this effort formed the basis for subsequent sequential analyses.

Multimodal differential sequence mining, a novel extension of differential sequence mining, was then developed to handle the large state space inherent in automatically generated multimodal data streams (Chapter 11). This technique was applied to the multimodal data streams of student task actions, dialogue, and nonverbal behavior, including facial expression, gesture, and posture. Multimodal sequences were found to be associated with tutorial outcomes of engagement, frustration, and learning gain. Additionally, incoming student characteristics of general and computer science self-efficacy were linked to other multimodal sequences. Among these results, one-hand-to-face gestures were found to occur more frequently with positive phenomena of engagement, learning, and general self-efficacy, while two-hands-to-face gestures occurred more frequently with frustration.

This line of research has improved automated understanding of learning-centered affect, with particular insights into how affect unfolds from moment to moment during tutoring. The next generation of affect-responsive intelligent tutoring systems will leverage such information to provide affective interventions at beneficial moments. This may result in systems that treat student affect not as transient states, but instead as interconnected links in a student's path toward learning.

References

Afzal, S., & Robinson, P. (2009). Natural Affect Data - Collection & Annotation in a Learning Context. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction* (pp. 1–7).

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. M. (2009). Emotion Sensors Go To School. In *14th International Conference on Artificial Intelligence in Education* (pp. 17–24).

Bakeman, R., & Gottman, J. (1997). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.

Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241.

Baker, R. S. J. d., Rodrigo, M. M. T., & Xolocotzin, U. E. (2007). The Dynamics of Affective Transitions in Simulation Problem-Solving Environments. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction* (pp. 666–677).

Bosch, N., D'Mello, S. K., & Mills, C. (2013). What Emotions Do Novices Experience during Their First Computer Programming Learning Session? In *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (pp. 11–20).

Boyer, K. E., Ha, E. Y., Wallis, M., Phillips, R., Vouk, M. A., & Lester, J. C. (2009). Discovering tutorial dialogue strategies with hidden Markov models. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 141–148).

Boyer, K. E., Phillips, R., Ingram, A., Ha, E. Y., Wallis, M., Vouk, M. A., & Lester, J. C. (2010). Characterizing the Effectiveness of Tutorial Dialogue with Hidden Markov Models. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 55–64). Springer.

Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a New General Self-Efficacy Scale. *Organizational Research Methods*, *4*(1), 62–83.

Cooper, D. G., Muldner, K., Arroyo, I., Woolf, B. P., & Burleson, W. (2010). Ranking Feature Sets for Emotion Models used in Classroom Based Intelligent Tutoring Systems. In *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization* (pp. 135–146).

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. NY: Harper-Row.

D'Mello, S. K., & Calvo, R. A. (2011). Significant Accomplishments, New Challenges, and New Perspectives. In R. A. Calvo & S. K. D'Mello (Eds.), *New Perspectives on Affect and Learning Technologies* (pp. 255–271). New York, NY: Springer.

D'Mello, S. K., Craig, S. D., & Graesser, A. C. (2009). Multimethod Assessment of Affective Experience and Expression during Deep Learning. *International Journal of Learning Technology*, *4*(3/4), 165–187.

D'Mello, S. K., Dale, R., & Graesser, A. C. (2012). Disequilibrium in the Mind, Disharmony in the Body. *Cognition & Emotion*, *26*(2), 362–374.

D'Mello, S. K., & Graesser, A. C. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157.

D'Mello, S. K., & Kory, J. (2012). Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 31–38).

D'Mello, S. K., Lehman, B., Pekrun, R., & Graesser, A. C. (2014). Confusion Can Be Beneficial for Learning. *Learning & Instruction*, *29*, 153–170.

D'Mello, S. K., Lehman, B., & Person, N. (2010). Monitoring Affect States During Effortful Problem Solving Activities. *International Journal of Artificial Intelligence in Education*, *20*(4).

D'Mello, S. K., Taylor, R. S., & Graesser, A. C. (2007). Monitoring Affective Trajectories during Complex Learning. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 203–208).

Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The Role of Emotion in Computer-Mediated Communication: A Review. *Computers in Human Behavior*, *24*(3), 766–785.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002a). *Facial Action Coding System*. Salt Lake City, USA: A Human Face.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002b). *Facial Action Coding System: Investigator's Guide*. Salt Lake City, USA: A Human Face.

El Kaliouby, R., & Robinson, P. (2005). The Emotional Hearing Aid: an Assistive Tool for Children with Asperger Syndrome. *Universal Access in the Information Society*, *4*(2), 121–134.

Fournier-Viger, P., Gomariz, A., Soltani, A., Lam, H., & Gueniche, T. (2014). SPMF: Open-Source Data Mining Platform. Retrieved from http://www.philippe-fournier-viger.com/spmf/

Fournier-Viger, P., Nkambou, R., & Nguifo, E. M. (2008). A Knowledge Discovery Framework for Learning Task Models from User Interactions in Intelligent Tutoring Systems. In *7th Mexican International Conference on Artificial Intelligence* (pp. 765–778).

Gee, J. P. (2004). *Situated Language and Learning: A Critique of Traditional Schooling*. Psychology Press.

Gottman, J., Markman, H., & Notarius, C. (1977). The Topography of Marital Conflict: A Sequential Analysis of Verbal and Nonverbal Behavior. *Journal of Marriage and Family*, *39*(3), 461–477.

Graesser, A. C., & Olde, B. A. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, *95*(3), 524–536.

Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2011). Predicting Facial Indicators of Confusion with Hidden Markov Models. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (pp. 97–106).

Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2012). Toward a Machine Learning Framework for Understanding Affective Tutorial Interaction. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 52–58).

Grafsgaard, J. F., Fulton, R. M., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2012). Multimodal Analysis of the Implicit Affective Channel in Computer-Mediated Textual Communication. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 145–152).

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013a). Automatically Recognizing Facial Expression: Predicting Engagement and Frustration. In *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 43–50). Memphis, Tennessee.

Grafsgaard, J. F., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2013b). Automatically Recognizing Facial Indicators of Frustration: A Learning-Centric Analysis. In *Proceedings of the 5th International Conference on Affective Computing and Intelligent Interaction* (pp. 159–165).

Ha, E. Y., Grafsgaard, J. F., Mitchell, C. M., Boyer, K. E., & Lester, J. C. (2012). Combining Verbal and Nonverbal Features to Overcome the "Information Gap" in Task-Oriented Dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (pp. 247–256). Seoul, South Korea.

Hardy, M., Wiebe, E. N., Grafsgaard, J. F., Boyer, K. E., & Lester, J. C. (2013). Physiological Responses to Events During Training: Use of Skin Conductance to Inform Future Adaptive Learning Systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 2101–2105).

Harrigan, J. A., & O'Connell, D. M. (1996). How Do You Look When Feeling Anxious? Facial Displays of Anxiety. *Personality and Individual Differences*, *21*(2), 205–212.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 904–908).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). Amsterdam: Elsevier Science.

Ho, J., Lukov, L., & Chawla, S. (2005). Sequential Pattern Mining with Constraints on Large Protein Databases. In *Proceedings of the 12th International Conference on Management of Data* (pp. 89–100).

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic Prediction of Frustration. *International Journal of Human-Computer Studies*, *65*(8), 724–736.

Kapoor, A., & Picard, R. W. (2005). Multimodal Affect Recognition in Learning Environments. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* (pp. 677–682).

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis*. Guilford Press.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, *39*(10), 1123–1134.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2013). A Contextualized, Differential Sequence Mining Method to Derive Students' Learning Behavior Patterns. *Journal of Educational Data Mining*, *5*(1), 190–219.

Kleinsmith, A., & Bianchi-Berthouze, N. (2012). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing*.

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies* (pp. 43–46).

Lee, D. M., Rodrigo, M. M. T., Baker, R. S. J. d., Sugay, J., & Coronel, A. (2011). Exploring the Relationship Between Novice Programmer Confusion and Achievement. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (pp. 175–184).

Lepper, M. R., & Woolverton, M. (2002). The Wisdom of Practice: Lessons Learned from the Study of Highly Effective Tutors. In J. Aronson (Ed.), *Improving Academic Achievement* (pp. 135–158). Elsevier.

Littlewort, G., Bartlett, M. S., Salamanca, L. P., & Reilly, J. (2011). Automated Measurement of Children's Facial Expressions during Problem Solving Tasks. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 30–35).

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J. R., & Bartlett, M. S. (2011). The Computer Expression Recognition Toolbox (CERT). In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 298–305).

Mahmoud, M., & Robinson, P. (2011). Interpreting Hand-Over-Face Gestures. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (pp. 248–255).

Marcoccia, M., & Atifi, H. (2008). Text-centered versus multimodal analysis of instant messaging conversation. *Language@Internet*, *5*.

Marx, J. D., & Cummings, K. (2007). Normalized Change. *American Journal of Physics*, *75*(1), 87–91.

Matsumoto, D., & Ekman, P. (2004). The Relationship Among Expressions, Labels, and Descriptions of Contempt. *Journal of Personality and Social Psychology*, *87*(4), 529–540.

McNeill, D. (2005). *Gesture & Thought*. Chicago: The University of Chicago Press.

McQuiggan, S. W., Lee, S., & Lester, J. C. (2007). Early Prediction of Student Frustration. In *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction* (pp. 698–709).

McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective Transitions in Narrative-Centered Learning Environments. *Educational Technology & Society*, *13*(1), 40–53.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007). Analysis of affect expressed through the evolving language of online communication. *Proceedings of the 12th International Conference on Intelligent User Interfaces - IUI '07*, 278.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). Affect Analysis Model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, *17*(01), 95–135.

O'Brien, H. L., & Toms, E. G. (2010). The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology*, *61*(1), 50–69.

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. (2006). Human Computing and Machine Understanding of Human Behavior: A Survey. In *Proceedings of the 8th International Conference on Multimodal Interaction* (pp. 239–248).

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review*, *18*(4), 315–341.

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic Emotions and Student Engagement. In *Handbook of Research on Student Engagement* (pp. 259–282). Springer US.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rodrigo, M. M. T., & Baker, R. S. J. d. (2011). Comparing Learners' Affect while using an Intelligent Tutor and an Educational Game. *Research and Practice in Technology Enhanced Learning*, *6*(1), 43–66.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 305–311).

Shanabrook, D. H., Cooper, D. G., Woolf, B. P., & Arroyo, I. (2010). Identifying High-Level Student Behavior Using Sequence-based Motif Discovery. In *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 191–200).

Symonds, M. R. E., & Moussalli, A. (2010). A Brief Guide to Model Selection, Multimodel Inference and Model Averaging in Behavioural Ecology using Akaike's Information Criterion. *Behavioral Ecology and Sociobiology*, *65*(1), 13–21.

Vail, A. K., & Boyer, K. E. (2014a). Adapting to Personality Over Time: Examining the Effectiveness of Dialogue Policy Progressions in Task-Oriented Interaction. In *Proceedings of the 15th Annual SIGDIAL Meeting on Discourse and Dialogue* (pp. 41–50).

Vail, A. K., & Boyer, K. E. (2014b). Identifying Effective Moves in Tutoring: On the Refinement of Dialogue Act Annotation Schemes. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems* (pp. 199–209).

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When Are Tutorial Dialogues More Effective Than Reading? *Cognitive Science*, *31*(1), 3–62.

Vygotsky, L. S. (1978). Interaction between Learning and Development. In *Mind in Society: The Development of Higher Psychological Processes* (pp. 79–91). Harvard University Press.

Walther, J. B. (1992). Interpersonal Effects in Computer-Mediated Interaction: A Relational Perspective. *Communication Research*, *19*(1), 52–90.

Whitehill, J., Serpell, Z., Foster, A., Lin, Y.-C., Pearson, B., Bartlett, M. S., & Movellan, J. R. (2011). Towards an Optimal Affect-Sensitive Instructional System of Cognitive Skills. In *Proceedings of the Computer Vision and Pattern Recognition Workshop on Human Communicative Behavior* (pp. 20–25).

Wiebe, E. N., Williams, L., Yang, K., & Miller, C. (2003). Computer Science Attitude Survey. *North Carolina State University Technical Report TR-2003-1*.

Woolf, B. P., Burleson, W., Arroyo, I., Dragon, T., Cooper, D. G., & Picard, R. W. (2009). Affect-Aware Tutors: Recognising and Responding to Student Affect. *International Journal of Learning Technology*, *4*(3-4), 129–164.

Wu, T., Butko, N. J., Ruvolo, P., Whitehill, J., Bartlett, M. S., & Movellan, J. R. (2012). Multi-Layer Architectures for Facial Action Unit Recognition. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, *42*(4), 1027–1038.

Appendices

Appendix A: Tutor Post-Session Survey

Created for use in JavaTutor Study II.

Tutor Post-Session Survey (reported using a 5-point scale)

Rate your agreement with the following statements:

1: Overall, the session was successful.
2: I felt like I provided cognitive support this session.
3: I helped the student finish the programming exercises more quickly than they would have on their own.
4: I helped the student master the most important concepts better than they would have on their own.
5: I was able to help the student finish the session with less effort than they would have on their own.
6: I felt like I provided emotional support this session.
7: I felt like my student was in the flow of the task.
8: I felt like my student thought the task was worthwhile.
9: I felt like my student found the task fun.
10: The student understood the computational thinking concepts.
11: The student understood the written task instructions.
12: The student understood my directions.

The student experienced the following during the lesson:

13: Anxiety (worried or uneasy about the lesson)
14: Boredom (not interested in the lesson or learning programming concepts)
15: Confusion (uncertain about some aspect of the lesson)
16: Contempt (scornful of the tutor, the lesson, or him/herself)
17: Excitement (enthusiastic or eager about the lesson)
18: Frustration (annoyed at difficulties with the tutor, the lesson, or him/herself)
19: Joy (happy about the tutor, the lesson, or him/herself)

I experienced the following during the lesson:

20-26: [tutor experience of emotion terms from items 13-19]
27: Open-ended response.

Appendix B: User Engagement Survey

Adapted for use in JavaTutor Study II, from the original (O'Brien & Toms, 2010).

User Engagement Survey (Likert items)

Focused Attention

1: I lost myself in this learning experience.
2: I was so involved in my learning task that I lost track of time.
3: I blocked out things around me when I was working.
4: When I was working, I lost track of the world around me.
5: The time I spent working just slipped away.
6: I was absorbed in my learning task.
7: During this learning experience I let myself go.

Endurability

8: Working on this learning task was worthwhile.
9: I consider my learning experience a success.
10: My learning experience was rewarding.
11: I would recommend using this tutoring system to my friends and family.

Felt Involvement

12: I was really drawn into my learning task.
13: I felt involved in this learning task.
14: This learning experience was fun.

Appendix C: NASA-TLX

Adapted for use in JavaTutor Study II from the original (Hart & Staveland, 1988).

NASA-TLX (Response scale from 0-100)

Mental Demand
1: How mentally demanding was the task?

Physical Demand
2: How physically demanding was the task?

Temporal Demand
3: How hurried or rushed was the pace of the task?

Performance
4: How successful were you in accomplishing what you were asked to do?

Effort
5: How hard did you have to work to accomplish your level of performance?

Frustration Level
6: How insecure, discouraged, irritated, stressed, and annoyed were you?

Appendix D: General Self-Efficacy

This is the New General Self-Efficacy instrument (Chen, Gully, & Eden, 2001).

General Self-Efficacy (Likert items)

1: I will be able to achieve most of the goals that I have set for myself.
2: When facing difficult tasks, I am certain that I will accomplish them.
3: In general, I think that I can obtain outcomes that are important to me.
4: I believe that I can succeed at most any endeavor to which I set my mind.
5: I will be able to successfully overcome many challenges.
6: I am confident that I can perform effectively on many different tasks.
7: Compared to other people, I can do most tasks very well.
8: Even when things are tough, I can perform quite well.

Appendix E: Computer Science Self-Efficacy

This is the Confidence sub-scale of the Computer Science Attitude survey (Wiebe, Williams, Yang, & Miller, 2003).

Computer Science Self-Efficacy (Likert items)
1: Generally I have felt secure about attempting computer programming problems.
2: I am sure I could do advanced work in computer science.
3: I am sure that I can learn programming.
4: I think I could handle more difficult programming problems.
5: I can get good grades in computer science.
6: I have a lot of self-confidence when it comes to programming.

Appendix F: Fine-Grained Dialogue Acts

These dialogue acts were created from the JavaTutor Study II corpus (Vail & Boyer, 2014b). These are the dialogue acts that appear in the multimodal data streams (Chapter 11).

Student	**Tutor**
Acknowledge	Acknowledge
Confirmation Question	Correction
Correction	Directive
Direction Question	Evaluative Question
Explanation	Explanation
Extra Domain Answer	Extra Domain Answer
Extra Domain Question	Extra Domain Question
Extra Domain Statement	Extra Domain Statement
Feedback Not Understanding	Factual Question
Feedback Understanding	Greeting
Greeting	Information Question
Information Question	Information Statement
Observation	Negative Elaborated Feedback
Ready Answer	Negative Feedback
Ready Question	Observation
WH Answer	Open Question
Yes/No Answer	Other Elaborated Feedback
	Other Feedback
	Positive Elaborated Feedback
	Positive Feedback
	Probing Question
	Query for Questions
	Ready Answer
	Ready Question
	Reassurance
	WH Answer
	Yes/No Answer

Appendix G: Multimodal Sequence Alphabet

These sequence event symbols were used in the itemset representation for multimodal differential sequence mining on the JavaTutor Study II corpus (Chapter 11). There are a total of 63 alphabet symbols.

COMPILE ERROR
COMPILE SUCCESS
FACE: AU1
FACE: AU14
FACE: AU2
FACE: AU4
FACE: AU7
FAR POSTURE
MID POSTURE
NEAR POSTURE
ONEHANDTOFACE
POSTURE SHIFT
RUN PROGRAM
START CODING
STOP CODING
STOP TYPING MESSAGE
STUDENT ACKNOWLEDGE
STUDENT CONFIRMATION QUESTION
STUDENT CORRECTION
STUDENT DIRECTION QUESTION
STUDENT EXPLAIN
STUDENT EXTRA DOMAIN ANSWER
STUDENT EXTRA DOMAIN QUESTION
STUDENT EXTRA DOMAIN STATEMENT
STUDENT FEEDBACK NOT UNDERSTANDING
STUDENT FEEDBACK UNDERSTANDING
STUDENT GREETING
STUDENT INFORMATION QUESTION
STUDENT NOT TYPING
STUDENT OBSERVATION
STUDENT READY ANSWER
STUDENT READY QUESTION
STUDENT WH ANSWER
STUDENT YES/NO ANSWER

TUTOR ACKNOWLEDGE
TUTOR CORRECTION
TUTOR DIRECTIVE
TUTOR EVALUATIVE QUESTION
TUTOR EXPLAIN
TUTOR EXTRA DOMAIN ANSWER
TUTOR EXTRA DOMAIN QUESTION
TUTOR EXTRA DOMAIN STATEMENT
TUTOR FACTUAL QUESTION
TUTOR GREETING
TUTOR INFORMATION QUESTION
TUTOR INFORMATION STATEMENT
TUTOR NEGATIVE ELABORATED FEEDBACK
TUTOR NEGATIVE FEEDBACK
TUTOR OBSERVATION
TUTOR OPEN QUESTION
TUTOR OTHER ELABORATED FEEDBACK
TUTOR OTHER FEEDBACK
TUTOR POSITIVE ELABORATED FEEDBACK
TUTOR POSITIVE FEEDBACK
TUTOR PROBING QUESTION
TUTOR QUERY FOR QUESTIONS
TUTOR READY ANSWER
TUTOR READY QUESTION
TUTOR REASSURANCE
TUTOR WH ANSWER
TUTOR YES/NO ANSWER
TWOHANDSTOFACE
TYPING MESSAGE