Multimodal Semi-Automated Affect Detection from Conversational Cues, Gross Body Language, and Facial Features

Sidney K. D'Mello and Arthur Graesser

Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152, USA

[sdmello|a-graesser]@memphis.edu

Abstract
We developed and evaluated a multimodal affect detector that combines conversational cues, gross body language, and facial features. The multimodal affect detector uses feature-level fusion to combine the sensory channels and linear discriminant analyses to discriminate between naturally occurring experiences of boredom, engagement/flow, confusion, frustration, delight, and neutral. Training and validation data for the affect detector were collected in a study where 28 learners completed a 32-minute tutorial session with AutoTutor, an intelligent tutoring system with conversational dialogue. Classification results supported a channel × judgment type interaction, where the face was the most diagnostic channel for spontaneous affect judgments (i.e., at any time in the tutorial session), while conversational cues were superior for fixed judgments (i.e., every 20 seconds in the session). The analyses also indicated that the accuracy of the multichannel model (face, dialogue, and posture) was statistically higher than the best single-channel model for the fixed but not spontaneous affect expressions. However, multichannel models reduced the discrepancy (i.e., variance in the precision of the different emotions) of the discriminant models for both judgment types. The results also indicated that the combination of channels yielded superadditive effects for some affective states, but additive, redundant, and inhibitory effects for others. We explore the structure of the multimodal linear discriminant models and discuss the implications of some of our major findings.

Key words: multimodal affect detection, conversational cues, gross body language, facial features, superadditivity, AutoTutor, affective computing, human-computer interaction.

Declaration: This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism

1. Introduction

The field of affective computing aspires to narrow the communicative gap between the highly emotional human and the emotionally challenged computer by developing computational systems that recognize and respond to the affective states (i.e., emotions) of the user (Picard, 1997). Affective computing is motivated by the realization that humans are more than mere cognitive processors; instead, emotions are inextricably bound to cognitive processes such as memory, causal reasoning, deliberation, goal appraisal, and planning (Bower, 1981; de Rosis, Castelfranchi, Goldie, & Carofiglio, in press; Mandler, 1976, 1984; Russell, 2003). This complex interplay between cognition and affect suggests that a computer interface that is sensitive to users' affective and cognitive states is expected to be more usable, useful, naturalistic, social, and enjoyable. Consequently, the last decade has witnessed an eruption of research activities aimed at making computer interfaces sensitive to users' affective and cognitive states in an attempt to develop more effective, user-friendly, and naturalistic applications (Conati & Maclaren, 2009; Forbes-Riley, Rotaru, & Litman, 2008; Hudlicka & McNeese, 2002; Madsen, el Kaliouby, Goodwin, & Picard, 2008; McQuiggan, Mott, & Lester, 2008; Paiva, Prada, & Picard, 2007; Porayska-Pomsta, Mavrikis, & Pain, 2008; Prendinger & Ishizuka, 2005).

Achieving affect sensitivity is, however, a very complex endeavor and a number of obstacles need to be overcome before affect-sensitive systems can be successfully implemented. On the forefront of these challenges lies the ability for computers to reliably detect users' affective states. The affect detection problem is very challenging because emotions are psychological constructs with notoriously noisy, murky, and fuzzy boundaries that are compounded with individual differences and contextual influences in experience and expression. Despite these

difficulties, reliable affect detection is critical because the success of any affect-sensitive interface will ultimately depend upon how accurately users' affective states can be recognized. Expectations are raised when humans recognize that a computer system is attempting to communicate at their level (i.e., with enhanced cognitive and emotional intelligence), far beyond traditional interaction paradigms (i.e., WIMP—window, icon, menu, pointing device). When these expectations are not met, users often get discouraged, disappointed, or even frustrated (Norman, 1994; Shneiderman & Plaisant, 2005).

In recent years, researchers in the field of affective computing have risen to the affect detection challenge by developing a variety of novel sensors, algorithms, tools, and applications. Recent review papers have surveyed the state of the art affect detection research (Jaimes & Sebe, 2007; Pantic & Rothkrantz, 2003; Zeng, Pantic, Roisman, & Huang, 2009), so we will not provide an extensive literature review here.

It is important to point out, however, that one commonality that emerges from the literature reviews is that most systems primarily focus on a single modality[1] such as facial expressions or acoustic-prosodic features of speech. Although relying on a single modality for affect detection is an obvious and useful first start, there are three significant problems with this approach. First, it is unclear whether all emotions are expressed via facial expressions and paralinguistic features of speech. For example, there is some evidence that naturalistic episodes of boredom and engagement, two affective states that are ubiquitous in almost any task, cannot be reliably detected from the face (Craig, D'Mello, Witherspoon, & Graesser, 2008; Craig, D'Mello, Witherspoon, Sullins, & Graesser, 2004; McDaniel et al., 2007). Hence, alternate channels might be needed to detect the subtle expressions associated with these emotions.

The second problem is that facial expressions and speech patterns can be controlled by a deceptive user. Emotional expressions are a highly socially reactive phenomenon, so it is conceivable that users may attempt to disguise certain negative emotions. For example, frustration is a state that is typically associated with significant physiological arousal, yet facial features are not very diagnostic of this emotion (McDaniel et al., 2007). It might be the case that people do not readily display frustration, perhaps due to the negative connotations associated with this emotion. This finding is consistent with Ekman's theory of social display rules (Ekman, 2002; Ekman & Friesen, 1969), in which social pressures may result in the disguising of negative emotions such as frustration. Other channels such as body language, lexical features, and contextual cues are more resilient to attempts to conceal these emotions (D'Mello, Picard, & Graesser, 2007; Ekman & Friesen, 1969).

Perhaps the most important limitation on relying on facial and acoustic-prosodic features is that barring a few exceptions that are discussed below, most systems consider these channels in isolation. But naturalistic emotional expressions are rarely unimodal; instead they involve a complex symphony of multimodal expression. For example, consider Damasio's eloquent description of a patient experiencing profound sadness: "The patient stopped her ongoing conversation quite abruptly, cast her eyes down and to her right side, then leaned slightly to the right and her emotional expression became one of sadness. After a few seconds she suddenly began to cry. Tears flowed and her entire demeanor was one of profound misery. Soon she was sobbing. As this display continued, she began talking about how deeply sad she felt, how she had no energies left to go on living in this manner, how hopeless and exhausted she was" (Damasio, 2003) (pp. 67-68). Though anecdotal, this example illustrates the richness of affect expressions and highlights the importance of considering a broad array of channels for affect detection.

1.1. *Survey of Multimodal Affect Detection Systems*

There have been a small number of systems that have explored multimodal affect detection. These primarily include amalgamations of physiological sensors and combinations of audio-visual features (Caridakis et al., 2006; Chen, Huang, Miyasato, & Nakatsu, 1998; Dasarathy, 1997; Picard, Vyzas, & Healey, 2001; Yoshitomi, Sung-Ill, Kawano, & Kilazoe, 2000; Zeng et al., 2006). Another approach involves a combination of acoustic-prosodic, lexical, and discourse features for affect detection (Ang, Dhillon, Krupski, Shriberg, & Stolcke, 2002; Lee & Narayanan, 2005; Liscombe, Riccardi, & Hakkani-Tür, 2005; Litman & Forbes-Riley, 2004).

Of greater interest are the handful of research efforts that have attempted to monitor three or more modalities. For example, Scherer and Ellgring (2007) considered the possibility of combining facial, vocal features, and body movements (posture and gesture) to discriminate 14 emotions (e.g., hot anger, shame). Since there was a uniform distribution of emotion instances in their data set, classification accuracy expected by chance is 7.14% (i.e., 1/14).

[1] The present paper uses the term "modality" and "channel" interchangeably. Both terms are used quite broadly. For example, facial features are one modality for affect detection. Acoustic-prosodic features from a spoken utterance can be considered to be a separate modality from lexical features derived by transcribing the speech.

Their results indicated that single-channel classification accuracies from 21 facial features and 16 acoustic-prosodic features were substantially greater than chance when evaluated on the full training set (52.2% and 52.5%, from face and speech, respectively). The paper did not report single-channel accuracy rates from body movements, although this channel was included in a three-channel multimodal model. This combined 37-feature model yielded a classification accuracy of 79% when evaluated on the full training set; the accuracy dropped to 49.1% when the model was cross-validated. It is difficult to assess whether the combined model led to enhanced or equivalent classification scores when compared to the single-channel models, because cross-validated accuracy scores were not provided for all of the single-channel models.

More recently, Castellano and colleagues considered the possibility of detecting eight emotions (some basic emotions plus irritation, despair, etc.) by monitoring facial features, speech contours, and gestures (Castellano, Mortillaro, Camurri, Volpe, & Scherer, 2008). Classification accuracy rates for the single channel classifiers were 48.3%, 67.1%, and 57.1%, for the face, gestures, and speech, respectively. The best multimodal classifier achieved an accuracy of 78.3%, representing a 17% improvement over the best single-channel system.

Although Scherer's and Castellano's systems demonstrate that there are some advantages to multichannel affect detection, the systems were trained and validated on context-free, acted, emotional expressions. However, practical applications require the detection of naturalistic emotional expressions that are grounded in the context of the interaction. There are also some important differences between real and posed affective expressions (Afzal & Robinson, 2009; Cohn & Schmidt, 2004; Ekman, Friesen, & Davidson, 1990; Pantic & Patras, 2006), hence, insights gleaned from studies on acted expressions might not generalize to real contexts.

On the more naturalistic front, Kapoor and colleagues developed a contextually grounded probabilistic system to infer a child's interest level on the basis of upper and lower facial feature tracking, posture patterns (current posture and level of activity), and some contextual information (difficulty level and state of the game) (Burleson & Picard, 2007; Kapoor & Picard, 2005). The combination of these modalities yielded a recognition accuracy of 86%, which was quantitatively greater than that achieved from the facial features (67% upper face, 53% lower face) and contextual information (57%). However, the posture features alone yielded an accuracy of 82%, which would indicate that the other channels are redundant with posture.

Kapoor and colleagues have extended their system to include a skin conductance sensor and a pressure-sensitive mouse in addition to the face, posture, and context (Kapoor, Burleson, & Picard, 2007). This system predicts self-reported frustration while children engaged in a Towers of Hanoi problem solving task. They system yielded an accuracy score of 79%, which is a substantial improvement over the 58.3% accuracy base line. An analysis of the discriminability of the 14 affect predictors indicated that mouth fidgets (i.e., general mouse movements), velocity of the head, and ratio of postures were the most diagnostic features. Unfortunately, the Kapoor et al. (2007) study did not report single channel classification accuracy, hence, it is difficult to assess the specific benefits of considering multiple channels.

In a recent study, Arroyo and colleagues considered a combination of context, facial features, seat pressure, galvanic skin conductance, and pressure exerted on a mouse to detect levels of confidence, frustration, excitement, and interest of students in naturalistic school settings (Arroyo et al., 2009). Their results indicated that two-parameter face + context models explained 52% and 29% of the variance for confidence and interest, respectively. A four parameter (face + context) accounted for 69% of the variance for excited, while a four-parameter seat pressure + context model was the best (46% of variance) model for predicting frustration. Their results support the conclusion that in most cases monitoring facial plus contextual features yielded the best fitting models, and the other channels did not provide any additional advantages.

One important issue with the aforementioned approaches to multichannel affect detection is that the number of features in the combined model increases as additional channels are considered. Therefore, the number of estimated parameters is very different in single versus multiple channel models. This makes it difficult to defend comparisons of the classification accuracy of the combined model with the single-channel models. Of course, more features do not necessarily indicate enhanced classification accuracy, and sometimes too many features can negatively affect performance. Nevertheless, when feature sets are unbalanced, any effect in favor of the multichannel models could be attributed to a mere increase in the size of the feature set, and not to any real additive effects of the different channels.

1.2. *Present Research*

Aside from the handful of systems that have investigated multimodal affect detection, multimodal systems for affect detection have been widely discussed but rarely implemented (Jaimes & Sebe, 2007; Pantic & Rothkrantz, 2003; Zeng et al., 2009). Our own research has also singularly focused on single channel affect detection. We have

developed systems that classified affective states by (a) examining contextual features from tutorial dialogues (D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008), (b) tracking gross body language (D'Mello & Graesser, 2009), (c) monitoring facial features (McDaniel et al., 2007), and (d) analyzing lexical, semantic, and cohesion relationships in tutorial dialogues (D'Mello, Dowell, & Graesser, 2009). This limitation is addressed in the current paper by considering the possibility of classifying emotions on the basis of a combination of conversational cues, gross body language, and facial features. We did not include the text-based features (i.e. lexical, semantic, and cohesion features) because research on these channels is still preliminary.

One important question that arises during the design of a multisensory emotion classification system involves determining the appropriate level of abstraction at which to fuse the output from the sensors. In general, feature-level fusion and decision-level fusion are the most commonly used methods (Pantic & Rothkrantz, 2003). Fusion at the feature level involves grouping features from the various sensors before attempting to classify emotions. Alternatively, in decision-level fusion, the affective states would first be classified from each sensor and would then be integrated to obtain a global view across the various sensors. Although decision-level fusion is more common in HCI applications (Marsic, Medl, & Flanagan, 2000; Pentland, 2000), several have questioned the validity of using decision-level fusion in the affective domain because audio, visual, and tactile signals of a user are typically displayed in conjunction and with some redundancy (Jaimes & Sebe, 2007; Pantic & Rothkrantz, 2003). Therefore, important details regarding the coordination of features from different modalities will be overlooked if each channel is analyzed separately. Consequently, this paper exclusively focuses on feature-level fusion in developing a composite (or multichannel) affect detector[2].

The fact that more than one affect sensor will be monitored raises the issue of what the results will exhibit. One intriguing hypothesis is that classification performance from multiple channels will exhibit *super-additivity*, that is, classification performance from multiple channels will be superior to an additive combination of individual channels. Simply put, the whole will be greater than the sum of the parts. Another possibility is that the combined model will result in *additive* effects, where the performance of multiple channels is equivalent to an additive combination of individual channels. It is also possible that there will be *redundancy* between the channels. In this situation, the addition of one channel to another channel yields negligible or zero incremental gains; the features of the two channels are manifestations of very similar mechanisms. Yet another possibility is that a combination of channels will result in *inhibitory* effects, where the composite models result in substantially lower classification rates than the individual channels.

These possibilities were explored in the current paper by developing a multimodal system to detect the affective states that learners experience during tutoring sessions with intelligent tutoring systems (ITSs). ITSs have emerged as valuable systems to promote active learning by capitalizing on the benefits of one-on-one tutoring in an automated fashion (Dodds & Fletcher, 2004; Koedinger & Corbett, 2006; VanLehn et al., 2007). ITSs have for many years tailored their learning support to students' needs in a variety of ways, including identifying the reasons behind student errors, and mastery learning through assessments of the probability that the student knows each skill relevant to the system (Anderson, Corbett, Koedinger, & Pelletier, 1995; Anderson, Douglass, & Qin, 2005; Gertner & VanLehn, 2000; Graesser, McNamara, & VanLehn, 2005; Graesser, VanLehn, Rose, Jordan, & Harter, 2001; Koedinger, Anderson, Hadley, & Mark, 1997; VanLehn, 1990; VanLehn et al., 2005). Although ITSs have typically focused on the learner's cognitive states, they can be far more than mere cognitive machines. ITSs can be endowed with the ability to recognize, assess, and react to learners' affective states. This has been a long desired goal for ITS developers. (Conati, 2002; Conati & Maclaren, 2009; D'Mello et al., 2005; D'Mello et al., 2007; De Vicente & Pain, 2002; Forbes-Riley et al., 2008; Kort, Reilly, & Picard, 2001; McQuiggan et al., 2008; Woolf, Burleson, & Arroyo, 2007). The current research contributes to the broader goal of developing affect-sensitive ITSs by developing multimodal detectors of boredom, engagement/flow, confusion, frustration, and delight; the affective states that accompany learning at deeper levels of comprehension (Baker, D'Mello, Rodrigo, & Graesser, 2010; D'Mello, Craig, Sullins, & Graesser, 2006; Lehman, D'Mello, & Person, 2008; Lehman, Matthews, D'Mello, & Person, 2008).

There are a number of differences between the research reported in this paper and the existing multimodal affect detection systems described above. First, while the earlier approaches primarily relied on the face and included a small number of contextual and posture features, we consider a larger array of novel contextual and posture features, thereby putting these alternate channels on equal footing with more traditional facial feature tracking. Second, we attempt to classify a different (i.e., learning centered) and larger (N = 6) set of affective states. Third, the number of features in the single and multichannel models were equivalent, thereby affording meaningful evaluations of superadditive, additive, redundant, and inhibitory effects. Fourth, instead of simply declaring that a combination of

[2] We also investigated decision-level fusion. Classification accuracy rates were similar to feature-level fusion.

channels yielded enhanced effects, we perform a deeper analysis to determine under what conditions and for what emotions was the combination superadditive versus redundant. Finally, we explore the structure of the multimodal affect detection models to understand how the different channels combine to discriminate the emotions.

We begin with a description of the multimodal corpus used to train and validate the classifiers (Section 2). Next we discuss our feature selection method and proposed criterion to evaluate multimodal classification effects (Section 3). Section 4 discusses the accuracy by which the single and multichannel models could discriminate between boredom, confusion, engagement/flow, frustration, delight, and neutral. Section 5 examines the structure of the multimodal linear discriminant models used for affect detection. We conclude with a discussion of our major findings, some limitations, and directions for future work.

2. Training and Validation Data

The data used to train the affect detectors were obtained from an existing study (Graesser et al., 2006) involving 28 learners being tutored by AutoTutor, an intelligent tutoring system with conversational dialogue, as will be described below. The subsequent discussion briefly describes AutoTutor, the data collection procedure, and the features extracted for affect detection.

2.1. *AutoTutor*

AutoTutor is a fully automated tutor for Newtonian physics, computer literacy, and critical thinking (Graesser, Chipman, Haynes, & Olney, 2005; Graesser et al., 2004). AutoTutor's dialogues are organized around difficult questions and problems (called main questions) that require reasoning and explanations in the answers. For example, the following is an example of a challenging question in computer literacy: "How can John's computer have a virus but still boot to the point where the operating system starts?" In order to correctly answer a main question, students need to articulate 3-7 sentential expressions, which normally involves between 25 and 100 turns in a conversation. When presented with these questions, students typically respond with answers that are only one word to two sentences in length. In order to guide students in their construction of an improved answer, AutoTutor actively monitors learners' knowledge states and engages them in a turn-based dialogue. AutoTutor adaptively manages the tutorial dialogue by providing feedback on the learner's answers (e.g. "good job", "not quite"), pumping the learner for more information (e.g. "What else"), giving hints (e.g. "What about X"), prompting (e.g. "X is a type of what "), correcting misconceptions, answering questions, and summarizing answers. Using these strategies, AutoTutor adheres to constructivist theories of pedagogy (Biggs, 1995; Chi, Roy, & Hausmann, 2008; Jonassen, Peck, & Wilson, 1999; Moshman, 1982) by allowing students to chart their own course through the tutorial dialogue and to build their own answers to difficult questions. The AutoTutor interface along with an excerpt of an actual dialogue between a college student and AutoTutor is presented in Figure 1.

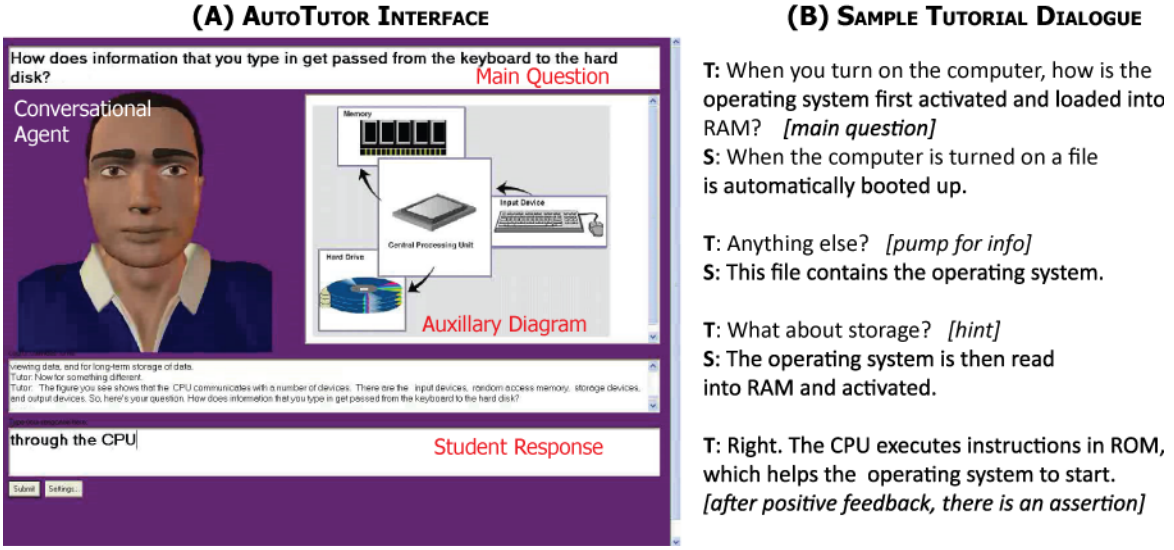

Figure 1. (A) AutoTutor interface and (B) Sample dialogue from a tutorial session

AutoTutor can keep the dialogue on track because it is always comparing what the student says to the anticipated input (i.e., the expectations and misconceptions in the curriculum script). Pattern matching operations and pattern completion mechanisms drive the comparison. These matching and completion operations are based on symbolic interpretation algorithms (Rus & Graesser, 2007) and statistical semantic matching algorithms (Landauer, McNamara, Dennis, & Kintsch, 2007). These include an Inverse Weighted Word Frequency Overlap algorithm, Latent Semantic Analysis (Landauer et al., 2007), and a Speech Act Classifier (Olney et al., 2003). Details on the actual mechanisms that AutoTutor uses to interpret the learner's contributions are presented in previous publications (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007).

The impact of AutoTutor in facilitating the learning of deep conceptual knowledge has been validated in more than a dozen experiments on college students (Graesser et al., 2004; Storey, Kopp, Wiemer, Chipman, & Graesser, in press; VanLehn et al., 2007). Tests of AutoTutor have produced learning gains of .4 to 1.5 sigma (a mean of .8 which is approximately equal to one letter grade) depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor.

2.2. *Data Collection Procedure*

2.2.1. *Interaction with AutoTutor*. The data for the present study came from 28 participants that interacted with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, internet, or operating systems. Three streams of information were recorded during the participant's interaction with AutoTutor (see Figure 2). A video of the participant's face was captured using the IBM® blue-eyes camera (Morimoto, Koons, Amir, & Flickner, 1998). Posture patterns were captured by the Tekscan® Body Pressure Measurement System (Tekscan, 1997). A screen-capturing software program called Camtasia Studio (developed by TechSmith) was used to capture the audio and video of the participant's entire tutoring session with AutoTutor. The captured audio included the speech generated by the AutoTutor agent. It is important to note that the left and right monitors (see Figure 2) were turned off during the AutoTutor interaction in order to prevent the participant from being distracted by the displays of their face and posture patterns.

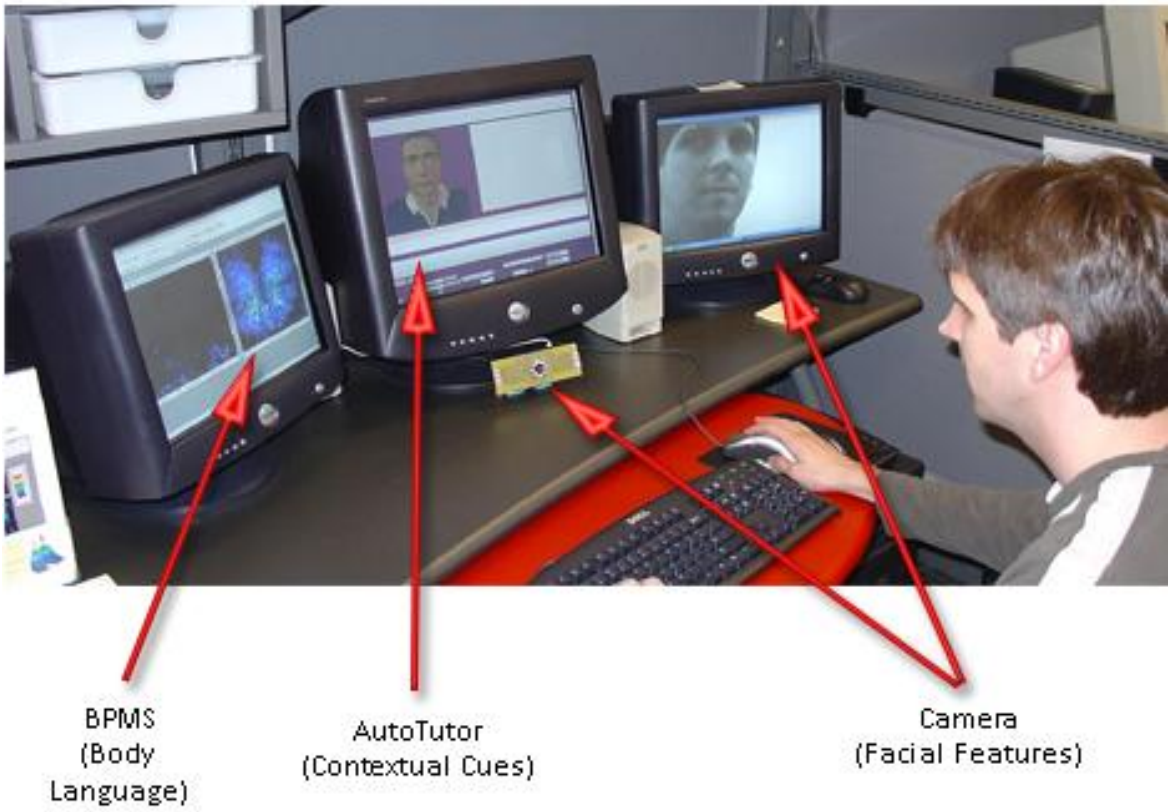

Figure 2. Learner interacting with AutoTutor

2.2.2. *Judging Affective States*. Two trained judges independently coded the learners' affective states (Graesser et al., 2006). The judges were trained on how to detect facial features with the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). The trained judges also had considerable experience interacting with AutoTutor. Hence, their emotion judgments were based on contextual dialogue information as well as the FACS system.

Judgments were based on the video streams from the participant's face and the AutoTutor screen. The screen capture included the tutor's animated speech, printed text, students' responses, dialogue history, and images, thereby providing the context of the participant's tutorial interaction. The alternatives for affective states were boredom, confusion, engagement/flow, frustration, delight, surprise, and neutral.

There were two different judgment points for each AutoTutor session. *Fixed* judgments were made at regular 20-second intervals, where the two video streams were automatically paused (freeze framed). *Spontaneous* judgments were obtained at any time between the 20-second points by manually pausing the videos. Judges were instructed to mark the state that was most pronounced at each judgment point.

2.3. *Distribution of Emotions and Interrater Reliability*

It is important to emphasize two important differences between these two judgment types[3]. First a different distribution of emotions was elicited for each judgment type. For the fixed judgments, the most common affective state was neutral (.40), followed by confusion (.27), engagement/flow (.15), and boredom (.16); the remaining states of delight, frustration, and surprise totaled .02 of the observations. The spontaneous points had a rather different distribution. The most prominent affect state was confusion (.42), followed by frustration (.18), delight (.28), and boredom (.04), whereas the remaining affective states comprised .08 of the observations; this includes a proportion score of .01 for neutral. So the subtle states of engagement/flow and boredom are more frequently observed at the fixed points when compared to the more obvious states of frustration and delight, which are routinely observed at the spontaneous points. Confusion is prominent at both judgment points.

The second difference between the two judgment types is that the emotion labeling task is more difficult if judges are asked to make emotion judgments at regularly polled timestamps (fixed points), rather than being able to stop a video display to make spontaneous judgments. Interrater reliability (kappa) (Cohen, 1960) scores between the two judges were .31 for the fixed points and .71 for the spontaneous points[4]. Kappa scores for individual emotions for the fixed judgments were .25, .36, .30, and .27, for boredom, confusion, flow, and frustration; respectively; there was an insufficient number of observations to compute kappas for delight, surprise, and frustration. Kappa scores for the spontaneous emotions that occurred with adequate frequency were .33, .76, .79, and .52, for boredom, confusion, delight, and frustration, respectively.

These kappas indicate that the states at fixed judgment points are much less salient so there is minimal information to base the judgments, compared with those points when affective states were voluntarily detected by the judge (spontaneous points). Training on facial expressions makes judges more mindful of relevant facial features and transient facial movements, but judges can do this only if the expressions have enough information to fortify these judgments.

Although kappas for the spontaneous points were higher than kappas for the fixed points, there is a concern that the spontaneous kappas might have been artificially inflated. In particular, our analysis only included spontaneous instances where both judges voluntarily made a judgment at approximately the same time (i.e., within one-second). The problem arises when one judge spontaneously identified an affective state that was overlooked by the other judge. Including these judgment points in the analysis would presumably lower the kappa.

We addressed this concern by performing a follow-up analysis where the two trained judges re-coded spontaneous observations that were identified by one or both of the judges. That is, the observation was included if any judge believed that an emotion occurred. The pattern of means, in descending order, for these observations was confusion (.36), boredom (.17), neutral (.15), frustration (.15), delight (.12), flow (.05), and surprise (.00). The kappa for this second round of spontaneous annotations was .49. Kappa scores for the emotions that occurred with sufficient frequency were .44, .59, .58, .37, and .31 for boredom, confusion, delight, frustration, and neutral, respectively. This second round of spontaneous judgments yielded a pattern of proportion scores and kappas that appear to be an amalgamation of the patterns for the fixed and first round of spontaneous observations. The subsequent analyses only consider spontaneous judgments from this second round of affect coding.

[3] The distribution of emotions and kappas have been previously reported in Graesser et al., (2006).

[4] A square agreement matrix is required to compute a kappa. The agreement matrix is rectangular in the rare cases when one of the judges does not rate all of the emotion categories. This problem was addressed by inserting zeros in the cells that corresponded to the missing category prior to computing the kappa.

2.4. *Creating Labeled Data Sets*

The development of a system that automatically detects the action units (AUs) is a challenging task because the coding system was originally created for static pictures rather than changing expressions over time. Although there has been remarkable progress in this area, the reliability of current automatic AU detection systems do not match humans (Asthana, Saragih, Wagner, & Goecke, 2009; Brick, Hunter, & Cohn, 2009; Hoque, el Kaliouby, & Picard, 2009). As an initial step, two trained raters (different than affect judges) coded a sample of the observations of emotions on facial action units (described in more detail below). Since manual annotation of facial features is a time-intensive endeavor, a random sample of data points was used for the current analyses (described below). Furthermore, we focused on points where the two trained judges agreed on the learners' emotions, so we have some confidence in the fidelity of the judgments.

There is a reason to suspect that the fidelity of the different data channels varies as a function of the judgment type. Since the spontaneous judgment points consist of the more obvious cases of emotion expressions, it is plausible that the face plays a more salient role for these points. On the other hand, it is possible that the context (i.e., dialogue features) is the most reliable channel for the fixed judgment points. Separate data sets for the fixed and spontaneous judgments were constructed in order to assess the hypothesis that the fidelity of each individual channel varies as a function of the judgment type. There was also a difference in the emotions in each data set. The fixed judgment data set consisted of boredom, confusion, engagement/flow, and neutral, while boredom, confusion, delight, frustration, and neutral were included in the spontaneous data set.

2.4.1. *Fixed Data Set.* Each trained judge provided 96 fixed judgments for each of the 28 learners, yielding 2668 judgments in all. There were 1350 data points in which both trained judges agreed on the learners' emotions (approximately half the time). A subset of 317 instances was randomly sampled from these 1350 points (about 25%). These points were sampled to approximate a uniform distribution of the different emotions across participants. Specifically, an approximately equal number of observations were obtained from each participant and for each of the affective states of boredom, confusion, engagement/flow, and neutral. There were 85 instances of boredom, 80 of confusion, 74 of engagement/flow, and 78 of neutral.

2.4.2. *Spontaneous Data Set.* There were only 1133 points of spontaneous observations. A subset of 407 emotion episodes were randomly sampled from the data points where the two trained judges agreed on learners' emotions. An approximately uniform distribution of the different emotions was obtained from each participant. There were 76 instances of boredom, 100 of confusion, 78 of delight, 84 of frustration, and 69 of neutral.

2.5. *Extracting Features from Sensors*

In order to prevent overfitting of the models due to the relatively small number of samples in each data set, a subset of features from each channel was considered for affect detection. These included features that were the most diagnostic of the learners' emotions and were derived from our previous research that considered each channel independently (D'Mello, Craig et al., 2008; D'Mello & Graesser, 2009; McDaniel et al., 2007). The large number of features considered in this paper renders a detailed description of each feature beyond the scope of the paper. Hence, the subsequent discussion provides an overview of the 29 features; detailed descriptions appear in D'Mello, Craig et al., 2008; D'Mello & Graesser, 2009, and McDaniel et al., 2007.

2.5.1. Conversational Cues (Dialogue Features). At the end of each student turn, AutoTutor maintains a log file that captures the student's response, a variety of assessments of the response, the feedback provided, and the tutor's next move. We extracted nine features from each student-tutor turn by parsing AutoTutor's log files. These included temporal measures, measures of response verbosity, measures of the conceptual quality of learners' responses, and measures of the tutor's feedback and level of directness.

The temporal features included the *subtopic number, turn number, and response time.* The subtopic number was the number of main questions answered. It provides a global measure of sequential position within the entire tutorial session. The turn number, on the other hand, provides a local temporal measure; it is the n^{th} turn of the student in the current question (subtopic). Response time was the elapsed time between the presentation of the question by AutoTutor and the student submitting an answer.

Assessments of response verbosity included the *number of characters* (letters, numbers) in the student's response and the *speech act.* The speech act specifies whether the student's response was a contribution towards an answer (coded as a 1) versus a nonsubstantive frozen expression, e.g., "I don't know", "Uh huh" (coded as –1).

The conceptual quality of the student's response was evaluated by Latent Semantic Analysis (LSA) (Graesser et al., 2007; Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2008). LSA-based measures included a *local good score* (the conceptual similarity between the student's current response and the expected answers) and a *global good score* (the similarity between all the student responses towards the current problem and the expected answers). Hence, the local score provides a measure of immediate progress (or understanding), while the global score is indicative of general performance towards the current problem (or subtopic).

AutoTutor's major dialogue moves were ordered onto a scale of conversational *directness,* ranging from –1 to 1, in terms of the amount of information the tutor explicitly provides the student: summary > assertion > prompt > hint > pump. AutoTutor's short *feedback* (negative, neutral negative, neutral, neutral positive, positive) was aligned on a scale ranging from –1 (negative feedback) to 1 (positive feedback).

Dialogue features were extracted for each turn and temporally aligned to the sample of fixed and spontaneous emotion judgments. More specifically, the emotion judgment that immediately followed a dialogue move (within a 15-second interval) was bound to that dialogue move.

2.5.2. *Gross Body Language (Posture Features).* Learners' posture patterns were automatically tracked with the Tekscan's Body Pressure Measurement System (BPMS). The BPMS consists of a thin-film pressure pad with a rectangular grid of sensing elements that is enclosed in a protective pouch. Pressure matrices (38×41) of participants' back and seat pressure while seated in a chair were recorded at 4Hz (see leftmost monitor in Figure 2).

There were four pressure-related features and one feature related to the pressure coverage for the back and the seat, yielding ten features in all. Each feature was computed by examining the pressure map during the emotional episodes for the fixed and spontaneous data sets (called the *current frame*). The pressure-related features included the *average pressure*, which measured the mean pressure exerted on the pad. The *prior change* and *post change* scores measured the difference between the average pressure in the current frame and the frame three seconds earlier and later, respectively. Finally, the *average pressure change* measured the mean change in pressure across a predefined window, typically four seconds, which spanned two seconds before and two seconds after an emotion judgment. The coverage feature was the proportion of non-negative sensing units (*average coverage*) on the pad.

2.5.3. *Facial Action Units (Facial Features).* Previous studies reduced the set of 58 action units to a subset of 10 AUs that were most diagnostic of the learning-centered emotions (Craig et al., 2008; Craig et al., 2004; McDaniel et al., 2007). These AUs occurred approximately 85% of the time with the learning-centered emotions, whereas the remaining 48 action units were much rarer. Therefore, instead of coding for all 58 AUs, the raters focused on the 10 AUs listed in Table 1[5]. The table also lists the proportion of each of the AUs aggregated across the two human coders. Kappa scores between the two coders for each AU are also presented. It should be noted that the majority of the activity of the facial features during emotional experiences occurred on the upper face. The kappa scores also indicate that the level of agreement achieved by the AU judges in coding the target action units ranged from fair to excellent (κ_{fixed} = .624, $\kappa_{spontaneous}$ = .733) (Robson, 1993).

[5] The affect judges and action unit coders were different researchers.

Table 1. Description of action units, proportional occurrence, and kappa scores.

Facial Action Unit			Fixed		Spontaneous	
			Proportion	*Kappa*	*Proportion*	*Kappa*
Upper Face	AU1	Inner Brow Raiser	.034	.632	.078	.642
	AU4	Brow Lowerer	.040	.800	.070	.779
	AU7	Lid Tightener	.040	.597	.113	.59
	AU43	Eye Closure	.003	-[a]	.071	.605
	AU45	Blink	.455	.745	.234	.681
Lower Face	AU12	Lip Corner Puller	.054	.520	.112	.707
	AU14	Dimpler	.064	.394	.041	-[a]
Lip Parting/ Jaw Opening	AU25	Lips Part	.114	.742	.134	.912
	AU26	Jaw Drop	.030	.452	.098	.851
Eye Position	AU64	Eyes Down	.165	.736	.049	.833

Note. [a] Kappa could not be computed as only one judge observed this action unit.

3. Multimodal Affect Detection with Feature-Level Fusion Models

Feature-level fusion created a multichannel feature vector by appending features from each individual channel. There are seven models that can be constructed from the three sensory channels. The first three models, referred to as single-channel models or individual models, consider each feature set individually. There was an *F* model for facial features, a *D* model for dialogue features, and a *P* model for posture features. The next three models were constructed by combining two feature sets. This yields three two-channel models: *FD*, *FP*, and *DP*. Finally, there was one model that was constructed via an additive combination of all three sensory channels, the *FDP* model.

3.1. *Feature Selection*

When each channel was considered independently, there were 10 predictors for the *F* and *P* models, and nine for the *D* model (see above). If the complete feature set of each one-channel model were used in the construction of the two-channel models, then the two-channel models would have 19 or 20 features. The three-channel model would have 29 features. As articulated in the Introduction, we consider the imbalance in the number of features to be somewhat problematic for evaluating superadditive versus additive effects.

One strategy to alleviate this problem is to construct the combined feature vector with a subset of features from the individual channels. It is obviously desirable that each channel contributes an equivalent number of features to the combined model so that each channel is equally represented. If m sensors are being considered, and each sensor has a varying number of features f_i, then the number of channels in the composite model would be: $f_c = \min(f_1, f_2, \ldots, f_m)$. In a uniform distribution, the number of features that each channel contributes is: $f_{equal} = f_c/m$.

As an example, consider a situation where $f_{face} = 10$, $f_{dialogue} = 9$, and $f_{posture} = 10$. For the face + dialogue+ posture model, $m = 3$, $f_c = \min(10,9,10) = 9$, and $f_{equal} = 9/3 = 3$. Hence, each channel would contribute three features and the full model would have nine features.

This procedure raises the question of how to select the f_{equal} features from each channel. One would obviously like to select the most diagnostic features. This can be achieved by using any feature selection algorithm such as stepwise selection, entropy reduction, and information gain (Hocking, 1976; Mitchell, 1997). The current analysis used the *F-ratio* from a univariate ANOVA (analysis of variance) that tested each feature to determine if its mean significantly differed across the different emotions and thereby was capable of discriminating emotions. Features with a higher F-ratio suggest that they have a greater potential in discriminating among the different emotions than features with lower F-ratio.

The results of the feature selection procedure are presented in Table 2. Each one-channel model used its entire feature set of nine or ten features, but the two-channel models were constructed by considering five of the ten

features from each individual channel. The three-channel model was constructed with three features from each channel[6]. As an example consider the process of constructing the *FD* model for fixed judgments. The feature selection procedure can be divided into two steps. First, each variable from the face was considered independently and ten F-ratios were computed. This process was repeated for the dialogue features. Next the five features with the highest F-ratio from the face were retained for the *FD* model (AU12, AU26, AU64, AU43, AU25 in descending order of F-ratio). Similarly, the five features with the highest F-ratio from the dialogue were included in the *FD* model (No. Characters, Subtopic No. Directness, Turn No., Reaction Time in descending order of F-ratio). The same idea can be extended to construct the *FDP* model. Only the three features with the highest F-ratios were included in the model.

Table 2. Features included in the various classification models.

	Fixed								**Spontaneous**							
Feature	F_{ratio}	*F*	*D*	*P*	*FD*	*FP*	*DP*	*FDP*	F_{ratio}	*F*	*D*	*P*	*FD*	*FP*	*DP*	*FDP*
Face																
AU1	1.68	×							4.58	×						
AU4	1.42	×							15.1	×			×	×		
AU7	1.42	×							21.9	×			×	×		×
AU43	2.17	×			×	×			2.37	×						
AU45	0.48	×							4.60	×						
AU12	5.83	×			×	×		×	84.5	×			×	×		×
AU14	0.57	×							1.19	×						
AU25	2.00	×			×	×			18.7	×			×	×		×
AU26	3.75	×			×	×		×	14.2	×			×	×		
AU64	3.12	×			×	×		×	0.98	×						
Dialogue																
Subtopic	10.1		×		×		×	×	18.1		×		×		×	×
Turn	8.19		×		×		×		6.07		×		×		×	
Reaction Time	6.00		×		×		×		1.10		×					
Characters	19.9		×		×		×	×	1.30		×					
Speech Act	2.64		×						0.57		×					
Local Good	2.63		×						2.61		×					
Global Good	4.97		×						4.49		×		×		×	
Directness	8.31		×		×		×	×	6.74		×		×		×	×
Feedback	2.96		×						17.7		×		×		×	×
Posture																
B Pressure	1.02			×					1.05			×				
B Prior Change	2.04			×		×	×		1.00			×				
B Post Change	0.55			×					0.38			×				
B Avg. Change	0.30			×					7.20			×		×	×	×
B Coverage	0.23			×					1.44			×		×	×	
S Pressure	6.40			×		×	×	×	0.79			×				
S Prior Change	4.36			×		×	×	×	1.71			×		×	×	
S Post Change	3.91			×		×	×		3.11			×		×	×	×
S Avg. Change	0.87			×					5.90			×		×	×	×
S Coverage	5.67			×		×	×	×	1.08			×				

Notes. × indicates that feature was included as a predictor in the model. B = Back, S = Seat, Avg. = Average. F = Face, D = Dialogue, P = Posture, FD = Face + Dialogue, FP = Face + Posture, DP = Dialogue + Posture, FDP = Face + Dialogue + Posture.

[6] It should be noted that comparisons of multichannel models with restricted feature sets to models with complete feature sets yielded similar results. Hence, the process of limiting each composite model to 9 or 10 features, instead of the full set of 20 or 30 features, does not cripple the models. Please see Appendix A for comparisons between partial and complete feature sets.

There is the important concern that our feature set might suffer from multicollinearity problems. This would cause some instability in the discriminant used to classify the emotions (described in the next section). We addressed this concern by performing a tolerance analysis. In particular, we computed the residual variance when each feature was regressed on the remaining features. Tolerance is a measure of unexplained variance, hence, a feature with high tolerance cannot be predicted from the remaining features. It has been suggested that a tolerance value lower than 0.4 is diagnostic of multicollinearity problems (Allison, 1999). Fortunately, only four out of the 29 features in our predictor set had a tolerance value less than 0.4 (but greater than 0.2), which would suggest that multicollinearity is not a major concern in our feature set.

3.2. *Evaluating Superadditive, Additive, Redundant, or Inhibitory Effects*

There is the important question of how to assess whether the combination of multiple channels has resulted in substantial rather than incremental effects over the individual channels. The requirement that the accuracy score of the combined model be statistically greater than the individual models is one initial evaluation criterion. It is easy to test whether two classification models are statistically different if the kappa statistic is used as the performance metric (Cohen, 1960). An important property of the kappa statistic is that a *z* score can be computed according to Eq. 1, which asymptotically approximates a normal distribution (Fleiss, 1981).

$$z = \frac{k}{se(k)} \qquad \textbf{Eq. 1}$$

This property of the kappa statistic allows us to derive a *z* score for the difference between two kappas, k_1 and k_2 (see Eq. 2). The cumulative density function of the normal distribution can then be consulted to obtain a significance value for the *z* score. Hence, if k_{1+2} is the kappa score for the combined model, and k_1 and k_2 are kappa scores for the individual models, then k_{1+2} should be significantly greater than both k_1 and k_2.

$$z = \frac{k_1 - k_2}{\sqrt{\frac{se(k_1)^2 + se(k_2)^2}{2}}} \qquad \textbf{Eq. 2}$$

After statistical significance has been established, the next step is to assess the size of the combined effect. One could consider incremental gains obtained above and beyond an additive combination of individual sensors to assess *superadditivity*. The Kappa score expected from an additive combination of two sensors is $k_1 + k_2 - k_1 \times k_2$ (the probability of the union between two independent events is $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A) \times \Pr(B)$). In line with this notion, Eq. 3 presents a threshold for superadditivity (s_{1+2}) that is sensitive to an additive combination of effects.

$$s_{1+2} = k_1 + k_2 - k_1 \times k_2 \qquad \textbf{Eq. 3}$$

In summary, there are two conditions under which superadditivity can be declared. Condition 1 requires that k_{1+2} is significantly greater than k_1 and k_2. Also, k_{1+2} must be significantly greater than s_{1+2} (Condition 2). If the first condition is satisfied, but $k_{1+2} \leq s_{1+2}$, then the feature fusion has resulted in *additivity*. *Redundancy* occurs to the extent that k_{1+2} is statistically equivalent to k_1 and k_2. There is one more possibility, namely an *inhibitory* effect. This occurs when a combination of multiple sensors results in accuracy scores that are significantly lower than the individual sensors, that is, if k_{1+2} is statistically lower than k_1 and k_2.

The superadditivity threshold specified in Eq. 3 can be extended to a three-channel classification problem. If k_3 is the accuracy of the third channel, then the threshold for superadditivity (s_{1+2+3}) is:

$$s_{1+2+3} = k_1 + k_2 + k_3 - k_1k_2 - k_1k_3 - k_2k_3 + k_1k_2k_3 \qquad \textbf{Eq. 4}$$

4. Classification Accuracy of Feature-Level Fusion Models

Linear discriminant analyses (LDA)[7] were conducted on the fixed and spontaneous data sets (Klecka, 1980). Linear discriminant analyses are a widely used classification procedure that consists of finding a linear combination of variables that best discriminates between the emotions. A leave-one-out cross validation method was used to gauge the accuracy by which the various models could discriminate between the emotions. According to this validation method, a single instance is removed from the training set and used to validate a model constructed on the remaining instances. The process is repeated so all training instances are individually used for validation.

The LDA analyses for the fixed data set performed a four-way discrimination between boredom, confusion, engagement/flow, and neutral, while a five-way discrimination (boredom, confusion, delight, frustration, and neutral) was considered for the spontaneous data set. There were seven analyses for each dataset, as specified above. Classification accuracy scores are listed in Table 3 and Figure 3.

Table 3. Overall classification results

	Fixed		**Spontaneous**	
Channel	*Cohen's Kappa*	*Proportion Correct*	*Cohen's Kappa*	*Proportion Correct*
Baserate (chance)	0	.250	0	.200
Single-Channel				
F	-.058	.205	.374	.499
D	.220	.416	.171	.346
P	.107	.331	.110	.300
Two-Channels				
FD	.271	.454	.391	.514
FP	.205	.401	.361	.489
DP	.242	.432	.198	.367
Three-Channels				
FDP	.288	.467	.382	.506

Note. Correct = Proportion correct

4.1. *Single-Channel Models*

Let us begin by considering the single-channel models. The accuracy of these models can be used as a lower bound on the performance of the multichannel models. Kappa scores for the fixed judgments were –0.058, 0.220, and 0.107 for the face, dialogue, and posture, respectively. The kappa scores for the dialogue and posture were significantly greater than zero ($p < .001$ unless specified otherwise), but the kappa score for the face was statistically indistinguishable from zero ($p > .05$), which is consistent with random guessing. So clearly, the face does not provide sufficient cues to discriminate the subtle emotional expressions during these fixed judgment points. The kappa scores for the dialogue model were significantly greater than the kappa scores for the posture model ($p = .002$). So it is the discourse context that plays a major role in discriminating among the emotions for the fixed data set. The posture features were about half as reliable as the dialogue features.

A rather different pattern of results was found for the spontaneous judgments. Although all three channels could significantly discriminate between the five emotions, the face was the most diagnostic with a kappa score of .374. This kappa score was significantly greater than kappas for the dialogue (.171) and posture (.110). Kappa scores for dialogue were significantly greater than kappas for posture ($p = .038$). Although performance of the face was poor

[7] A quadratic discriminant analysis yielded accuracies similar to linear discriminant functions. Hence, only results from the linear discriminant analysis are reported here.

for the fixed judgments, it is quite accurate for the spontaneous judgments, at least when compared to the other channels.

These results are important because they question the importance of the face as the primary communicative channel for emotional expressions. The face can be highly diagnostic of emotions when the expressions are accompanied by heightened activity in the face that can be visibly detected by the judges (i.e., spontaneous points). But the face is less useful when the emotion judgments are obtained at regularly polled timestamps. Although these events are not accompanied by vigorous facial activity, there is some confidence in the fidelity of these judgments because the fixed data set used for the discriminant analysis only consisted of judgments in which both trained judges agreed on the learners' emotions. These results highlight the widely acknowledged, but rarely realized, importance of using multiple modalities for affect detection. It appears that there is an interaction between the channel (face, dialogue, posture) and judgment type (fixed, spontaneous). The next step is to determine whether fusing data from multiple channels yields classification accuracies that are superior to the best individual channel.

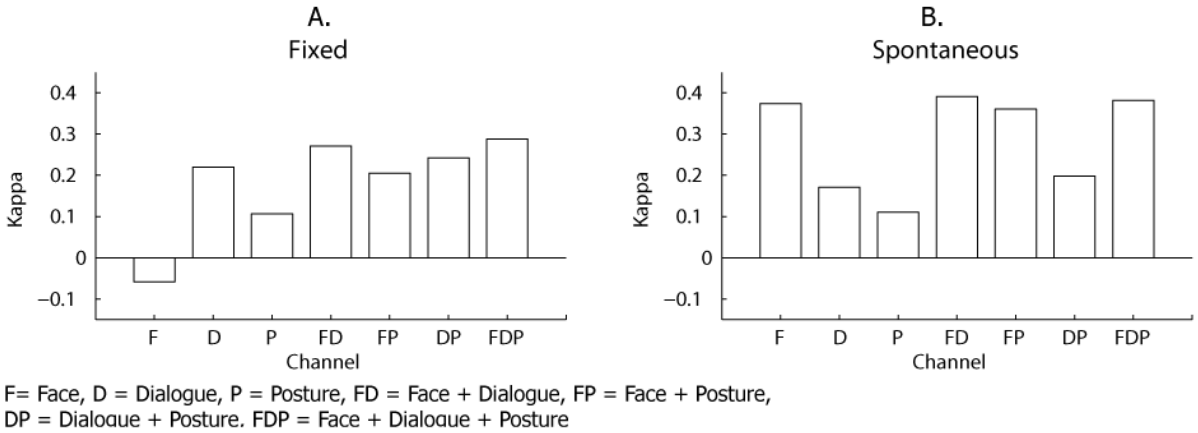

Figure 3. Overall classification results for feature-level fusion

4.2. *Multichannel Models*

Results of the fixed models are presented in Figure 3A. Consider first the two-channel models. The kappa scores for these models were statistically indistinguishable (κ_{FD} = .271, κ_{FP} = .205, and κ_{DP} = .242). The FD model was superior to the F model but equivalent to the D model. The DP model was superior to the P model but equivalent to the D model ([FD = D] > F[8] and [DP = D] > P). So adding the face or posture to dialogue clearly does not result in superadditive effects. Instead the results supported a combined model that is superior to the face and posture, but equivalent to the dialogue ([FD = DP = D] > P > F).

In contrast, a different effect was discovered for an additive combination of the face and posture. The FP model was statistically greater than the individual F and P models. It appears that adding facial features to posture features doubles the diagnosticity of the posture model (κ_F = –.058, κ_P = .107, κ_{FP} = .205; $p < .01$). Furthermore, the kappa for the FP model clearly exceeds the superadditivity threshold for two-channel models (see Eq. 3), which is consistent with a superadditive effect.

Let us now consider the three-channel FDP model that yielded a kappa score of .288. The accuracy of this model was significantly higher than the F and P individual models ($p < .01$), and significantly greater than the D model (p = .068, which is significant on a one-tailed test). But does this model resonate with superadditivity? According to Eq. 4, the superadditivity threshold for this three-channel model is .263. So it appears that combining features from the face, dialogue, and posture yields effects that surpass an additive combination of the individual channels.

Results of spontaneous models are presented in Figure 3B. Recall that the face was the dominant channel for the spontaneous data set, and it appears to have preserved its dominance for the two-channel models. The kappa scores for the two-channel models that include the face were on par with each other (κ_{FD} = .391, κ_{FP} = .361) and significantly higher than the model without the face (κ_{DP} = .198). The following pattern of results is obtained when

[8] [*A* = *B*] > *C* implies that *A* and *B* are statistically equivalent and are significantly greater than *C*.

the two-channel models that include the face were compared to the single channel models: [FD = FP = F] > D > P. The pattern of results for the models that exclude the face was: [DP = D] > P.

The three-channel spontaneous FDP model yielded a kappa of .388. This kappa was significantly greater than the individual dialogue and posture models, but not the face. So unlike the fixed models, superadditivity effects were not discovered for the spontaneous models.

In summary, four major patterns emerge when the fixed and spontaneous multichannel models were compared to their single channel counterparts. First, for the fixed judgments, two-channel models encompassing dialogue outperformed single channel posture and face models, but not the combined face and posture model. Second, the fixed three-channel FDP model outperformed the single channel models and yielded a superadditive effect. Third, for the spontaneous judgments, models that included the face outperformed models that excluded the face. Finally, the spontaneous FDP model was redundant with the single channel face model, so there are negligible improvements when other channels were added to the dominant face model.

4.3. *Discrepancy Reduction*

The results suggest that the composite models have no merits for the spontaneous judgments. In essence, the face prevails. Before this conclusion is accepted too cavalierly, it is important to consider alternative performance metrics. For example, one alternative is to compare the classification patterns of the best single-channel models with the composite models. These classification patterns can be expressed as the precision scores of each emotion. While the kappa score provides a measure of the overall level of agreement when considering all emotions, the precision scores for the individual emotions provide an indication of how well each emotion was classified.

Figure 4 offers a comparative view of the precision scores for the best single model (dialogue for fixed and face for spontaneous) along with the associated three-channel models. The precision scores have been corrected for base rate biases, as there were minor differences in the distributions of the various emotions[9].

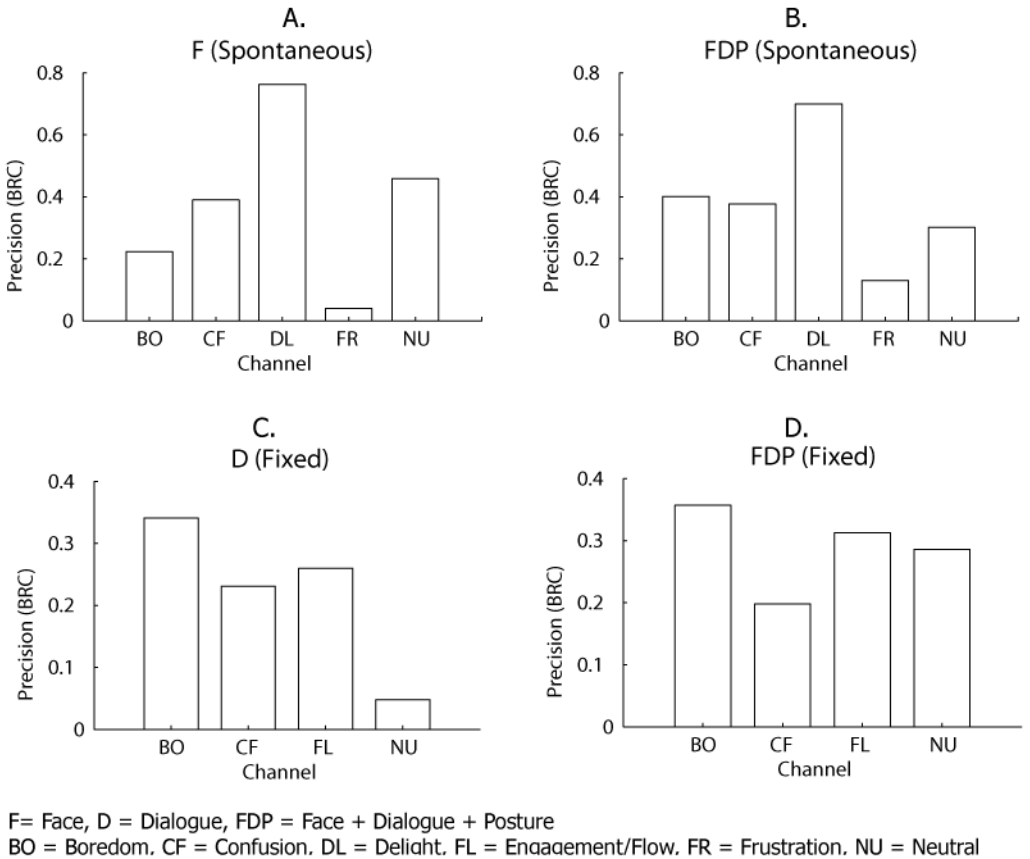

Figure 4. Single versus multi channel precision scores for individual emotions.

[9] BRC = Base rate corrected.

According to Figure 4A, the face was very successful in detecting delight for the spontaneous judgments, but was poor for frustration. In fact the precision for frustration was approximately zero. An interesting phenomenon occurs when the composite model is considered (Figure 4B). Although the overall classification rates between these two models were equivalent, there were differences in the precision patterns. Delight preserves its dominance, but frustration is now being detected with a modicum of accuracy. The accuracy of boredom also improved, whereas the precision of neutral was somewhat lower.

A more striking effect was observed for the fixed judgments. In this case, the precision of boredom, confusion, and engagement/flow were approximately equivalent for the single channel dialogue model, but the precision of neutral was much lower (Figure 4C). The precision of neutral substantially improved for the composite model (Figure 4D), at the expense of a small negative shift in the precision of confusion.

Although the question of how the patterns of precision shift when additional channels are concerned is indeed relevant, the most important finding is that the addition of multiple channels reduced the *discrepancy* of the model. Discrepancy, in this context, refers to the degree to which the precision scores for the various emotions fluctuate (i.e., more fluctuation = more discrepancy). Ideally a model would have zero or low discrepancy; each channel would be classified with approximately equivalent precision. A highly discrepant model might be excellent at classifying certain emotions but poor for others. Although the overall classification rates for models with high versus low discrepancy might be equivalent, the latter is definitely more desirable in the current context.

An obvious metric to quantify the degree of discrepancy of a model is the population variance of the precision scores for the individual emotions (see Table 4). The variance for the F and FDP spontaneous models were .058 and .034, respectively. Therefore, adding the additional channels reduced the discrepancy of the single-channel F model by a factor of 1.7. A much larger effect was observed for the fixed judgments. In this situation, the composite model reduced the discrepancy of the best single channel model by a factor of 4 (discrepancy for dialogue = .012; discrepancy for FDP model = .003). Hence, discrepancy reduction appears to be the real advantage of the composite models.

Table 4. Discrepancy scores for fixed and spontaneous models

Discrepancy	**F**	**D**	**P**	**FD**	**FP**	**DP**	**FDP**
Fixed	.015	**.012**	.005	.004	.003	.012	**.003**
Spontaneous	**.058**	.062	.036	.037	.050	.027	**.034**

4.4. *Effects for Individual Emotions*

The fact that the composite models reduced the discrepancy of the single channel models without necessarily increasing the overall classification scores (particularly for spontaneous judgments) suggests that the composite models have different effects on different emotions. Fusing features from multiple sensors increases the precision of some emotions but reduces the precision for others (see Table 5). The influence of the composite models on the precision of each emotion was assessed by comparing precision scores from the best one, two, versus three channel models. The D and F models were the best single-channel models for the fixed and spontaneous judgments, respectively. The FD model was the best two-channel model for both judgments types. So the analyses focused on the D, FD, and FDP models for the fixed judgments and the F, FD, and FDP models for the spontaneous judgments.

The analyses investigated whether the precision scores for each emotion substantially increased (superadditive effects), decreased (inhibitory effects), or were unchanged (redundant effects) when features from multiple channels were included in the discriminant analysis. In order to avoid conducting a large number of significance tests and risk committing Type I errors, a less stringent metric for superadditivity is adopted for the analyses on individual emotions. Because we have narrowed the focus to three models for each judgment type, we mainly focus on the gain in classification accuracy as one moves from a single-channel to a two-channel model and from a two to a three-channel model. In particular, the metric depicted in Eq. 5 was used to evaluate the degree of superadditivity afforded by fusing channels. Large values are consistent with superadditive effects, values near zero indicate redundant effects, while negative values suggest inhibitory effects.

$$\%\ superadditivity = \frac{k_{combined} - k_{single}}{k_{single}} * 100 \qquad \textbf{Eq. 5}$$

4.4.1. *Boredom*. It appears that the two-channel FD model had no enhanced effect for fixed boredom judgments (see Figure 5A). However, the FD model did yield superadditive effects for the spontaneous data set. The FD model yielded an impressive 79.8% improvement over the F model. So the face does not provide sufficient cues to detect learners' boredom for the spontaneous judgments, compared to the situation when the discourse context is considered (i.e. FD model).

Whereas the D fixed model was better than the F spontaneous model, the FD and FDP models were approximately equivalent. So the addition of D features to the F spontaneous model allows precision scores for spontaneous boredom to match the precision of fixed boredom. It should also be noted that the addition of posture to the FD model did not result in enhanced boredom detection for either data set, indicating that a two-channel FD model is sufficient to detect this emotion.

Table 5. Base rate corrected precision scores for emotions

	Fixed						**Spontaneous**					
Channel	*Bor*	*Con*	*Del*	*Flo*	*Fru*	*Neu*	*Bor*	*Con*	*Del*	*Flo*	*Fru*	*Neu*
F	–.205	.064	-	.066	-	–.156	**.223**	**.390**	**.762**	-	**.040**	**.459**
D	**.341**	**.231**	-	**.260**	-	**.048**	.304	.111	.048	-	.565	–.170
P	.132	.064	-	.207	-	.031	–.036	.443	.127	-	–.110	.127
FD	**.325**	**.215**	-	**.348**	-	**.201**	**.401**	**.403**	**.730**	-	**.175**	**.232**
FP	.180	.181	-	.295	-	.167	.272	.390	.746	-	.055	.336
DP	.357	.248	-	.295	-	.065	.353	.257	.064	-	.370	–.048
FDP	**.357**	**.198**	-	**.313**	-	**.286**	**.401**	**.377**	**.699**	-	**.130**	**.302**

Note. The critical models are in bold typeface.

4.4.2. *Confusion*. The face was a good indicator of confusion for the spontaneous points (.390) but not the fixed points (.064). A reverse effect was observed for contextual discourse cues. Here, the dialogue was diagnostic of confusion for the fixed points (.231) but not the spontaneous points (.111). Figure 5B reveals that the FD models resulted in redundant effects for both judgment types. Furthermore, in contrast to boredom where the composite models resulted in approximately equivalent precision scores for both judgment types, confusion was always detected more accurately at the spontaneous points.

These results suggest that judges rely on two different criteria in detecting confusion. Although confusion has well defined facial and discourse correlates (Craig et al., 2008; D'Mello, Craig et al., 2008; McDaniel et al., 2007), judges appear to be consulting each channel independently in judging confusion. They pay attention to the face for the spontaneous judgments and concentrate on the dialogue context for the fixed judgments.

4.4.3. *Delight*. The face was very accurate in detecting delight (.762), so one could expect inhibitory effects when additional channels are recruited to classify this emotion. Fortunately, the results do not support this conclusion. Instead redundant effects are observed when the FD and FDP models attempted to diagnose delight (see Figure 5C).

4.4.4. *Engagement/Flow*. Engagement/flow was only considered in the fixed data set, so the composite model was compared to the D model. It appears that the addition of facial features to the single channel D model resulted in 33.9% increase in accuracy. Adding posture to this FD model did not result in an additional improvement (Figure 5D). So it is the face and the dialogue that collectively signaled heightened engagement akin to an engagement/flow experience.

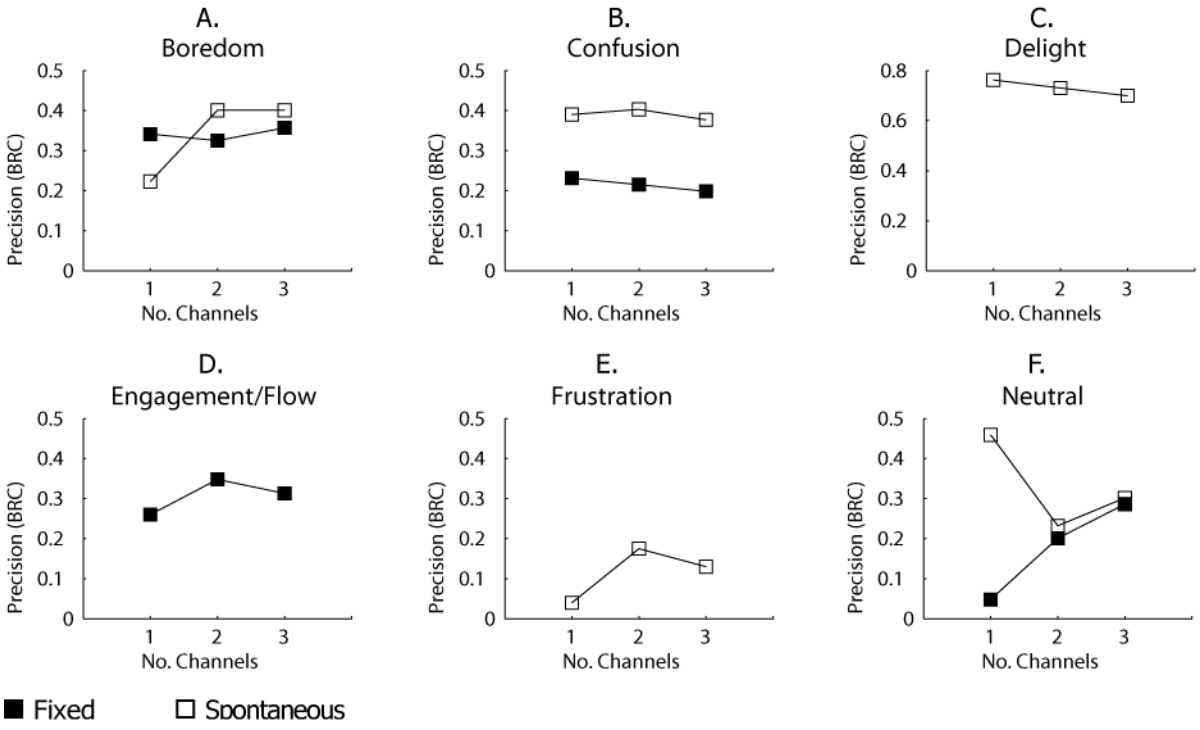

Figure 5. Precision scores for each emotion for the best one, two, and three channel models.

4.4.5. *Frustration.* A more striking pattern was observed for frustration (Figure 5E). Frustration is a state that is typically associated with significant physiological arousal, yet the facial features that were tracked were not very good at detecting this emotion (precision = .04). This finding is consistent with Ekman's theory of display rules (Ekman & Friesen, 1975), in which social pressures may result in the disguising of negative emotions such as frustration. However, an additive combination of discourse and facial features resulted in an impressive 333.9% superadditive effect. So although learners might attempt to disguise their frustration on the face, an examination of the discourse history betrays their frustration.

4.4.6. *Neutral.* The composite models had contradictory effects for the fixed versus spontaneous judgments. The precision of neutral and the number of channels were linearly related for the fixed judgments. The precision score of the FD model was consistent with a 320.1% superadditive effect over the D model. Furthermore, adding posture to the FD model led to a 42.3% increase in the precision over the FD model. So although the F, D, and P models individually provided poor precision scores (–.156, .048, and .031), the combined FDP model had a comparatively impressive precision of .286.

The results for the spontaneous models were less impressive. Here, the inclusion of dialogue features to the face model resulted in a 49.5% reduction in the precision, which is consistent with an inhibitory effect. So the face by itself is sufficiently reliable at classifying neutral, presumably with a marked decrease in facial activity. But the inclusion of dialogue complicates the situation and the precision of neutral suffers. Fortunately, the addition of posture yielded a small effect of 30.1% over the FD model so that the precision of the FDP spontaneous model matched the FDP fixed model.

4.5. *Practical Significance of Results*

According to Table 3, the cross-validated kappa scores for the fixed and spontaneous FDP models were .288 and .382, respectively. Classification accuracy (% correct) scores were 46.7% and 50.6% for fixed and spontaneous judgments, respectively; approximately half of the instances were correctly classified. The accuracy scores of these FDP models can be considered to represent "moderate" accuracy, since a reliable upper bound on classification accuracy is undetermined, and it is unlikely that perfect accuracy will ever be achieved. These results are significant

because affect detection is a difficult problem that is compounded with noise data, fuzzy categories, and individual differences in the experience and expression of emotion.

It is important to note, however, that our results might be undesirably influenced by methodological artifacts introduced by the leave-one-out cross validation procedure. The problem arises because cross-validation procedures do not provide a clear person-level separation between the training and test sets. In particular, different instances from the same participant can be present in both the training and the test set, thereby potentially inflating the accuracy scores and providing no guarantees for generalizability across individuals.

Since individual differences play an important role in affect expression, what is needed is an evaluation method that transcends individual differences. We addressed this concern by assessing the classification accuracy of the fixed and spontaneous FDP models with a split-half evaluation method. The analyses proceeded as follows. 14 of the 28 participants were randomly selected and their instances were assigned to the training set. Instances from the remaining 14 participants were assigned to the test set. Discriminant models were constructed from the training instances and evaluated on the testing instances. For the fixed FDP model, kappa scores for the training and test sets were .355 and .323, respectively. Kappa scores for the spontaneous model were .469 and .353 for the training and test sets, respectively. The results are positive because they indicate that there is a negligible loss in classification accuracy between the training and test set for the fixed FDP model. Although, there was a reduction in test-set accuracy for the spontaneous FDP model, the split-half kappa for this model was similar to the cross-validated kappa (.382). This suggests that individual differences did not negatively impact the performance of the cross-validated models described in sections 4.1-4.4. The discriminant models correctly classified approximately half of the instances irrespective of whether a leave-one-out method or a more conservative split-half method was used to evaluate the models. Hence, these accuracy scores can be expected in real-world situations where the affect detector has to classify the emotions of unknown learners.

5. Structure of Multimodal Discriminant Models

Taking a step back from the classification accuracy of the discriminant models, it is important to investigate *how* the predictors in the FDP model discriminant between the emotions. This can be achieved by analyzing the structure of the discriminant functions generated for the FDP model, as is described below.

5.1. *Fixed FDP Model*

Three discriminant functions were generated for the fixed FDP model that attempted to classify four emotions[10]. These functions were all statistically significant, $\chi^2(27) = 174.6$, $p < .001$ for Function 1; $\chi^2(16) = 73.4$, $p < .001$ for Function 2; and $\chi^2(7) = 17.8$, $p = .013$ for Function 3. The first two functions accounted for 90.8% of the variance, with 60.2% of the variance explained by the first function and the remaining 30.6% of the variance attributed to Function 2. Function 3 contributed a mere 9.2% of the variance, so subsequent discussion exclusively focuses on Function 1 and Function 2.

The centroids for the four emotions projected on the first two discriminant functions are depicted in Figure 6. The plot is consistent with the basic valence-arousal model (Barrett, 2006; Russell, 2003) where Function 1 represents the valence dimension and Function 2 is the arousal dimension. Valence increases from left to right and arousal increases from top to bottom. As could be expected, the neutral state is located near the origin. Boredom is a state with negative valence and low arousal and is located on the top left quadrant. The centroids of boredom and neutral are close to each other, suggesting that the FDP model has some difficulty in discriminating between these emotions. Similar to boredom, engagement/flow also has low arousal but positive valence, and is well segregated from the other emotions. Finally, confusion is a state with heightened arousal and a modicum of positive valence. The somewhat positive valence associated with confusion is puzzling because confusion is typically considered to be a negative emotion, although it does correlate with learning gains.

[10] The number of functions required to classify g groups is $g - 1$ (Klecka, 1980).

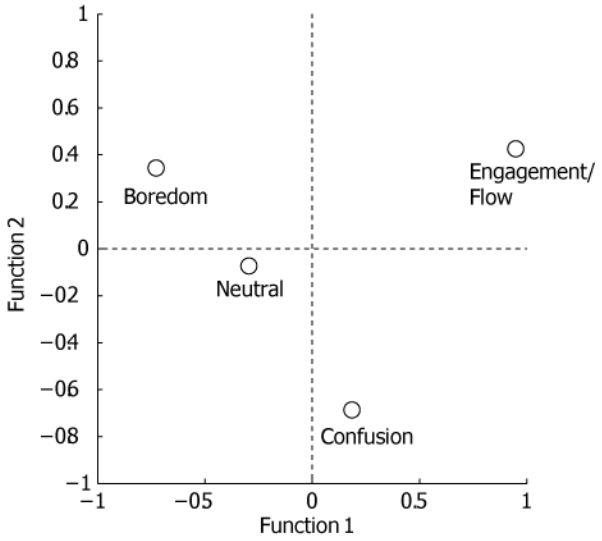

Figure 6. Group centroids for fixed FDP model.

According to Figure 6, Function 1 distinguishes boredom and neutral from engagement/flow and confusion. Function 2 segregates emotions with lower arousal such as boredom, neutral, and engagement/flow from the highly aroused state of confusion. An examination of the relationship between the predictors and the functions provides some insights into how the FDP model discriminates between the affective states. The predictor-function relationship is characterized by the *Structure Matrix* (see Table 6). Each cell in the matrix represents the pooled within-groups correlations between predictors and standardized canonical discriminant functions (Klecka, 1980).

It should be noted that discriminant functions do not represent individual affective states. Hence, a correlation between a feature and a discriminant function should not be interpreted as a correlation between a feature and an affective state. The discriminant functions carve up the feature space, hence, correlations between features and discriminant functions provide an indication how the predictors discriminate between the affective states. For example, Function 2 discriminates emotions with low arousal (such as boredom and confusion) from the highly aroused state of confusion. The fact that AU12 correlates with this function (see Table 6), suggests that this feature plays an important role in performing an arousal based discrimination.

The structure matrix indicated that the number of characters in the learner's responses and the coverage on the seat of the chair were positively correlated with Function 1 (Table 6). Tutor directness negatively correlated with Function 1. So Function 1 is indicative of the student leaning forward (i.e., high coverage on seat) and providing verbose responses (i.e., large number of characters) to the tutor's pumps, hints, and prompts (i.e., low directness). On the basis of this description, one might characterize this function as the *active-student* function, because students that take initiative lean forward and do most of the talking (student is highly verbose while tutor is less direct). Furthermore, this function discriminates confusion and engagement/flow from boredom and neutral.

Function 2 is characterized by an increase in the subtopic number, a lack of a lip corner puller, and a negative prior change in seat pressure (see Table 6). So Function 2 is indicative of lack of arousal in both the face and the body as the tutorial session drags on (high subtopic number). Function 2 might be characterized as the *passive-student* function, where the student leans back (negative seat coverage) and acts as a passive receptacle of information while the tutor does most of the talking (i.e., high tutor directness). Note that the student does contribute some information as evident by the .350 correlation between the number of characters and discriminant Function 2. However, this correlation is much lower than the .654 correlation between the number of characters and Function 1, indicating that Function 1 is consistent with more student verbosity than Function 2.

Table 6. Structure matrix for fixed FDP model.

	Discriminant Function		
Feature	*Function 1*	*Function 2*	*Function 3*
Number of Characters	.654*	.350	.147
Tutor Directness	–.398*	.283	–.207
Seat Coverage	.344*	–.117	–.312
Subtopic Number	–.177	.596*	.504
Lip Corner Puller (AU12)	–.015	–.490*	.382
Seat Prior Change	–.185	–.356*	–.245
Seat Average Pressure	.315	.049	–.618*
Jaw Drop (AU26)	.125	–.259	.532*
Eyes Down (AU64)	.256	–.002	.278*

Note. * Largest absolute correlation between each variable and any discriminant function (e.g., number of characters has the largest correlation with Function 1, compared to the other functions). The features in this table have been sorted with respect to how they correlate with the discriminant functions. Since Function 1 explains more variance than the other functions, the features that strongly correlate with Function 1 are listed first (number of characters, tutor directness, seat coverage). Function 2 explains the second largest amount of variance, hence, the features that load (i.e., are correlated with) onto this function are listed next.

5.2. *Spontaneous FDP Model*

Four discriminant functions were generated for the spontaneous FDP model that attempted to classify five emotions. These functions were all statistically significant at $p < .001$, $\chi^2(36) = 498.4$ for Function 1; $\chi^2(24) = 205.6$ for Function 2; $\chi^2(14) = 105.0$ for Function 3, and $\chi^2(6) = 28.1$ for Function 4. The first three functions were able to account for 95.6% of the variance, with 65.4% of the variance explained by Function 1, 17.3% by Function 2, and 12.8% by Function 3. Since Function 4 only explained 4.4% of the variance, the subsequent discussion focuses on the first three functions.

Figure 7 provides three different views of the distribution of the emotion centroids over the discriminant space. As evident in Figure 7A and Figure 7B, Function 1 segregates frustration and delight from boredom, neutral and confusion. The Structure Matrix listed in Table 7 indicates that the activation of the lip corner puller correlates strongly with Function 1. This implies that Function 1 utilizes information from the *lower face* (i.e., the mouth) to segregate frustration and delight from the other states. The apparent importance of mouth movements to identify frustration is consistent with Kapoor's earlier finding that mouth fidgets were the best predictor of frustration, although there was considerable variability across subjects (Kapoor et al., 2007).

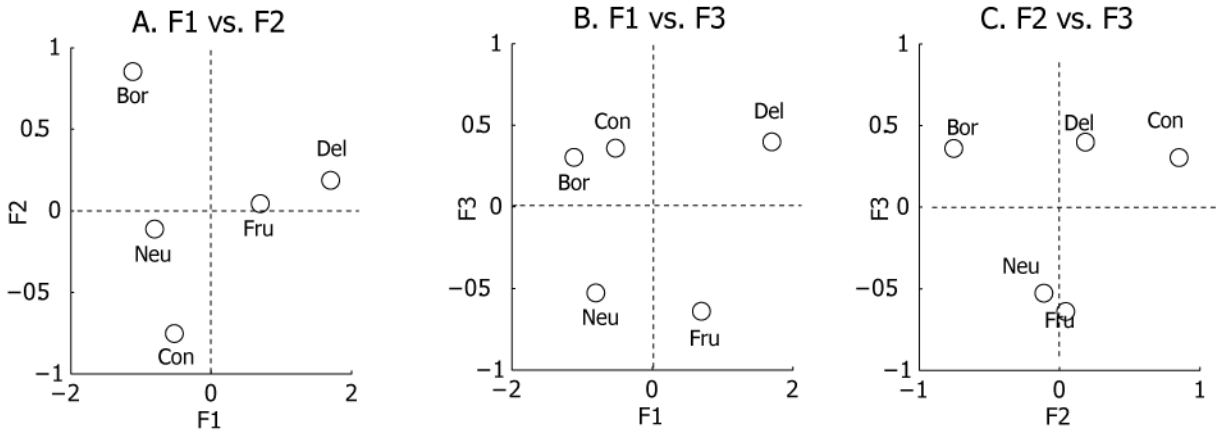

Figure 7. Group centroids for spontaneous FDP model.

Function 2 appears to segregate boredom from the other emotions. The Structure matrix indicates that high subtopic numbers (i.e., time on task) and arousal on the back and the seat of the chair correlate with this function (Table 7). Perhaps these patterns are indicative of *fidgeting* as the session progresses.

Although Function 3 explains less variance than the other two functions, it appears to separate neutral and frustration from boredom, delight, and confusion (Figure 7B and Figure 7C). The Structure matrix suggests that Function 3 primarily represents an amalgamation of activity in the lower face (lips part), upper face (lid tightener), and seat (post change) (see Table 7).

Table 7. Structure matrix for spontaneous FDP model.

	Discriminant Function			
Feature	*Function 1*	*Function 2*	*Function 3*	*Function 4*
Lip Corner Puller	.871*	.252	–.037	–.154
Subtopic Number	–.253	.597*	.092	.308
Back Avg Change	.127	.407*	.165	–.108
Seat Avg Change	.119	.328*	.240	–.041
Lid Tightener	.302	–.397	.549*	.370
Lips Part	.331	.211	.500*	.151
Seat Post Change	–.141	.067	.180*	.122
Feedback	–.252	.171	.526	–.747*
Directness	–.163	.272	–.117	.442*

Note. * Largest absolute correlation between each variable and any discriminant function

6. General Discussion

The results support a number of conclusions about how the affective states are manifested in the face, dialogue, and posture, and how these channels combine during emotional episodes. The subsequent discussion lists some of the most significant findings, followed by an analysis of some of the limitations, possible avenues for future work, and applications of the multimodal affect detector.

6.1. *Summary of Major Findings*

Our major findings can be aligned with respect to the following four goals: (1) To compare the accuracy by which the individual channels classify the emotions, (2) To compare single-channel affect detection with multichannel affect detection, (3) To identify conditions in which a combination of channels yields superadditivity versus situations that indicate redundancy, and (4) To analyze the structure of multimodal discriminant models. Item 4 has already been discussed in substantial detail, hence, this discussion exclusively focuses on the first three goals.

6.1.1. *Comparing Face, Dialogue, and Posture.* Psychological investigations of the saliency of the face versus contextual information for emotion communication have resulted in conflicting conclusions. On one hand, proponents of basic emotions and face dominance have consistently maintained that facial signatures of the basic emotions are innate, universal, and cross cultural (Ekman, 1984, 1992; Ekman & Friesen, 1975; Keltner & Ekman, 2000; Panksepp, 2000; Turner & Ortony, 1992). On the other hand, opponents suggest that emotional expressions are always modulated by context and might be best understood via a two-dimensional valence-arousal model, instead of the basic emotion categories (Barrett, 2006; Ortony & Turner, 1990; Russell, 1994, 2003; Turner & Ortony, 1992). The results partially confirm both positions and support a channel × judgment type interaction, where

the face was the most diagnostic channel for the spontaneous judgments, while the dialogue was superior for the fixed judgments.

The critical insight that the results support is that an emotion can be expressed with or without significant facial cues. The face clearly dominates when the degree of facial arousal is so pronounced that judges can spontaneously identify an emotion. But of equal importance are situations where the face is unexpressive, even though the learner is in the midst of an affective experience. It is in these situations, where performance of facial feature tracking is subpar, that the contextual information illuminates the correct emotion.

There is yet another situation in which the surrounding context rescues the face. Frustration is a state that is typically associated with significant physiological arousal, yet the facial features that were tracked were not very good at detecting this emotion. A closer look at how the face discriminates between the emotions can explain this result. Consider delight, an emotion with vigorous facial activity with excellent precision scores. Previous studies (on spontaneous judgments) have revealed that a number of action units that span the entire face are diagnostic of delight (McDaniel et al., 2007). In particular, the presence of AU 7 (lid tightener), AU 12 (lip corner puller), AU 25 (lips part), and AU 26 (jaw drop) coupled with an absence of AU 45 (blink) are diagnostic of this emotion (these patterns are generally consistent with a smile). In contrast to delight, the only significant correlation with frustration was obtained for AU 12 (lip corner puller)—perhaps indicative of a half smile that is similar to delight. This may be an attempt by the learner to disguise an emotion associated with negative connotations in society (Ekman & Friesen, 1975) and it clearly poses problems for facial feature based emotion detection. But precision scores for frustration significantly increase when contextual information is included, further supporting the importance of context in detecting naturalistic emotion expressions.

6.1.2. *Single-Channel versus Multiple-Channel Affect Detection.* The results revealed that the accuracy of multichannel FDP model was statistically higher (albeit marginally) than the best single-channel dialogue model for the fixed judgments. This FDP model also reduced the discrepancy (i.e., variance in the precision of the different emotions) of the D model. So a multichannel model for tracking emotional states at the fixed polled timestamps would be recommended.

Having decided to adopt a multichannel model, there is the issue of deciding whether all three channels are essential or if a two-channel model would suffice. For the fixed judgments, the FP model had lower accuracy scores than the FDP model, and hence cannot replace the FDP model. Performances of the FD and DP models were on par with each other and statistically equivalent to the FDP model. Hence, either two-channel model can replace the FDP model.

The next decision is to select either the FD or DP model. Although the overall classification accuracy of both models was statistically equivalent, there are two reasons to select the FD model over the DP model. First, although the precision scores for both models were approximately equivalent for boredom, confusion, and engagement/flow, there was a difference in the precision scores for neutral in favor of the FD model (.201 and .065 for FD and DP, respectively; see Table 5). The second reason to select the FD model over the DP model is that its discrepancy score was about three times lower (see Table 4).

What about the more salient emotional states, namely the ones that occur at spontaneous timestamps? Here, the accuracy of the FDP model was statistically equal to the accuracy of the best single-channel F model. However, the FDP model did reduce the discrepancy of the F model, and for this reason, can be considered to be superior to the F model. Another reason to consider the FDP model is that its precision for frustration (.130) is much higher than the near zero (.040) precision obtained by the F model (see Table 5). Since frustration is a state that has the potential to substantially disrupt the learning process, it is advisable to select the model that can detect this state with a modicum of precision.

Unlike the fixed judgments where models involving the dialogue were in competition, it is the models that involve the face that compete for the spontaneous judgments. Here, the kappa scores for the DP model were significantly lower than all other competing models (FD, FP, and FDP), essentially rendering DP noncompetitive. Kappa scores for the FD and FP models were on par with each other as well as with the more complex FDP model. Hence, the two competing models are the FD and FP models. The FD model enjoyed an inherent advantage over the FP model since it was the winner for the fixed judgments. It also had slightly lower discrepancy scores than the FP model (see Table 4) and nonzero precision scores for frustration.

In summary, it appears that tracking facial features with contextual cues is the best emotion detection strategy. Hence, it would appear that posture is redundant with these two channels in both the fixed and the spontaneous contexts. This finding is similar to a previous study reported by Arroyo and colleagues. Their results also indicated that a combination of facial features and contextual cues yielded the best models to predict confidence, excited, and

interested (Arroyo et al., 2009). Posture, skin conductance, and pressure exerted on the mouse were largely redundant with the face and context.

6.1.3. *Superadditive Effects*. The results indicated that superadditive effects were discovered for the fixed judgments but not the spontaneous judgments. There are two possible interpretations of this finding. The first position states that when a single channel is very efficient at classification, then adding additional channels results in redundancy instead of superadditivity. The face was the most superior channel for the spontaneous judgments, so adding dialogue and posture had no effect. But this position would not be able to explain why superadditivity effects were discovered for the fixed judgments, where the dialogue was the most reliable channel. In fact, the degree of face and dialogue dominance for the spontaneous and fixed data sets is equivalent. For the spontaneous data set, the percent improvement of the face over the next best channel (i.e., dialogue) is 119%[11]. For the fixed data set, the percent improvement of dialogue over posture (i.e., the next best channel) is 106%. Therefore, this position cannot explain why superadditivity effects were discovered for the fixed but not for the spontaneous judgments, since the degree of single channel dominance for both judgments types is approximately equivalent.

The alternate position states that the lack of superadditivity for the spontaneous judgments can be attributed to a simple difference between the fixed and spontaneous judgments. It is not a stretch to assume that judges primarily relied on the face when they provided spontaneous judgments. From an information-theoretic position, the rate of change of information on the face is far greater than the dialogue, because facial expressions change spontaneously while contextual changes are turn based. So judges will obviously focus more on the face. The fact that facial information is readily available at the spontaneous points eliminates the need for considering the additional channels, which result in redundant effects. On the other hand, the face is not a very reliable source of information for the fixed points. Here, the judges need to carefully monitor the additional channels in making their judgments, thereby yielding superadditive effects.

One remarkable finding about the fixed points was the additive combination of face and posture resulted in superadditivity. Some important insights can be gleaned by examining how the addition of the face to posture drastically improves its classification accuracy. Posture, by itself, was able to detect boredom (base rate corrected precision = .132) and engagement/flow (.207), but not confusion (.064) and neutral (.031) (see Table 5). The addition of facial features resulted in a small improvement in the precision of boredom (.180) and engagement/flow (.295), but a drastic improvement in detecting confusion (.181) and neutral (.167) (see Table 5). Although neither channel is accurate at detecting any emotion, the combined model is quite accurate, at least when compared to the individual channels. This effect is akin to a form of *emergence*, where the face and posture unite to create a model, whose effects cannot be explained by the sum of the individual channels alone. This is a bona-fide example of the whole being greater than the sum of the parts.

6.2. *Limitations*

There are five primary limitations with this study. The first two limitations are associated with the inclusion of the facial features with the other channels. Although the facial features are good predictors of affect in certain contexts, the classification results of models that include the face should be interpreted with some caution. This is because trained human judges annotated the facial action units (AUs) of the learners, so one might expect some reduction in accuracy when the AUs are automatically coded by a computer. This is a rather important limitation because automatic AU detection is obviously required for real time affect detection.

The second downside of expanding the scope of the current analyses to include facial features is that this reduced the sizes of the data sets. Since manual annotation of facial features is a tedious and laborious process, it was necessary to proceed with a random sample of observations, instead of the complete data sets. Although the ability to generalize is somewhat reduced by the smaller data set, this is not considered to be a major problem because: (a) the sampling procedure ensured that an approximately random distribution of emotions was selected from each participant, and (b) only 9-10 parameters were included in each model, which eliminated any serious overfitting concerns, and (c) there was a sufficient number of observations to perform the statistical analyses with adequate power.

[11] For spontaneous: $k_F = .374$; $k_D = .171$; $\%Improvement = \frac{\kappa_F - \kappa_D}{\kappa_D} * 100$

For fixed: $k_D = .220$; $k_P = .107$; $\%Improvement = \frac{\kappa_D - \kappa_P}{\kappa_P} * 100$

The third limitation with this study is primarily methodological. Although the results indicated that posture was redundant with the other channels, a critic could attribute this effect to the emotion judging methodology. Since the affect judges retrospectively provided ratings of the learners' emotions from videos of the learners' face and computer screen (but not posture via the BPMS), it is reasonable to expect that features from these channels correlate with their judgments at a higher rate than the posture features. Simply put, the judges were more mindful of the face and dialogue than posture.

It is possible to argue that although the output of the BPMS was not explicitly provided to the judges, learners' body language could be inferred from the videos of the face. If one assumes that the primary indices of body movement are attentiveness (i.e., forward lean versus backward lean) and arousal (magnitude of gross movement) (Bull, 1987), then it is reasonable to expect these indices to be derived at a crude grain size from the videos of the learners' face and upper torso. Of course, there is no evidence to suggest that the judges relied on this attentiveness-arousal framework in their interpretation of body language derived from video. Hence, as it stands, the fact that posture replays were not included in the retrospective affect judgment protocol is a limitation in this study.

The fourth limitation pertains to the finding that the face was the most effective channel for the spontaneous judgments. Although information from both the context and the face was available to the judges, it is plausible that they focused more on the face when providing spontaneous judgments. A critic could rightly claim that it is no surprise that classifiers that monitored facial movements were predictive of learner affect; judges were simply more sensitive to facial movements when making their judgments. An important question to ask is whether the face will retain its status as the most effective predictor for spontaneous judgments when these points are identified without any facial input. For example, one can envision a paradigm where judgment points are first identified by monitoring learner physiology (e.g., galvanic skin response, electromyography) or by some other method (e.g., emote-aloud). Judges are subsequently asked to provide affect annotations at these points by monitoring the face, the context, or both face and context. It remains to be seen whether facial features still outperform contextual cues in this affect annotation paradigm.

The fifth limitation pertains to the fact that multiple perspectives were not incorporated in the affect judgment procedure. The present study relied on affect annotations by two trained judges, and it is likely that a different pattern of results might have emerged if additional perspectives (e.g., self reports, peer judgments) were considered. We did collect self-reported affect data in the present study, however, this data was not used to develop the affect detectors because of sparseness issues and other related problems. On the positive front, we do have some evidence that some of our diagnostic features generalize across individual differences in affect judges (D'Mello, Craig, & Graesser, 2009). Whether this is the case for the broad set of predictors considered in this paper is a question for future research.

6.3. *Future Work*

There are two primary avenues of future research. The first consists of using automated facial feature coding systems to annotate the videos of learners' faces to see if the observed patterns replicate with automated facial coding. Another important line of research to pursue involves integrating acoustic-prosodic features with the face, posture, and context. Paralinguistic features of speech have been shown to be a viable channel for affect detection (Banse & Scherer, 1996; Fernandez & Picard, 2005; Johnstone & Scherer, 2000; Lee & Narayanan, 2005; Scherer, 2003; Scherer & Ellgring, 2007; Scherer, Johnstone, & Klasmeyer, 2003), so there is the question of whether their inclusion will yield superadditive effects.

We have already begun to make some progress along this front. In a replication of the current study participants verbally expressed their contributions to AutoTutor rather than typing them in (D'Mello, King, Entezari, Chipman, & Graesser, 2008). Future analyses will assess the reliability in detecting the affective states from acoustic-prosodic features as well as composite models that combine vocal information with conversational cues, gross body language, and facial features. Whether a combination of these four channels yields superadditive effects above and beyond those obtained in the current analysis awaits future technological development and empirical evaluation.

6.4. *Applications*

We have recently developed two new versions of AutoTutor that detect and respond to learners' boredom, confusion, and frustration (D'Mello, Jackson et al., 2008; D'Mello, Craig, Fike, & Graesser, 2009). Appropriate responses to these states could potentially have a positive impact on engagement and learning outcomes. These affect-sensitive versions of AutoTutor have a set of production rules that were designed to map dynamic assessments of the student's cognitive and affective states with tutor actions to address the presence of the negative emotions. Hence, the learner and the tutor are embedded into an affective loop that involves *detecting* the learner's

affective states, *responding* to the detected states, and *synthesizing* emotional expressions via animated pedagogical agents. The multimodal affect detection system described in this paper plays a major role in the affect-sensitive versions of AutoTutor.

In addition to its application with AutoTutor, the multimodal affect detector can be integrated into any application that requires the detection of boredom, flow/engagement, confusion, frustration, and neutral. The use of contextual cues raises a challenge for widespread integration with different systems, because, to some extent, these features are system specific. Although we expect that several of the contextual features will generalize to other domains (e.g., verbosity, feedback, response time), it is likely that a new set of contextual cues will be required for every new domain. This is expected because affect can never be divorced from context (Barrett, Mesquita, Ochsner, & Gross, 2007; Russell, 2003). The question of how much reengineering will be required when moving to new domains and how effective the affect detector will be in these domains awaits further research and development.

Acknowledgments

We thank our research colleagues in the Emotive Computing Group and the Tutoring Research Group (TRG) at the University of Memphis (http://emotion.autotutor.org; http:// www.iismemphis.org), and at the Affective Computing Group at MIT. We would also like to thank the three anonymous reviewers for their valuable suggestions that significantly improved the paper.

This research was supported by the National Science Foundation (ITR 0325428, and HCC 0834847). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

References

Afzal, S., and P. Robinson: 2009, 'Natural Affect Data - Collection and Annotation in a Learning Context'. *International Conference on Affective Computing and Intelligent Interaction.* Amsterdam, The Netherlands

Allison, P. D.: 1999, 'Multiple Regression'. Thousand Oaks, CA: Pine Forge Press.

Anderson, J., Corbett, A., Koedinger, K., and Pelletier, R.: 1995, 'Cognitive tutors: Lessons learned'. *Journal of the Learning Sciences*, **4**, 167-207.

Anderson, J., S. Douglass, and Y. Qin: 2005, 'How should a theory of learning and cognition inform instruction'? In A. Healy (ed.): *Experimental cognitive psychology and it's applications*, Washington, DC.: American Psychological Association. pp. 47-58.

Ang, J., R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke: 2002, 'Prosody-based automatic detection of annoyance and frustration in human-computer dialog'. *International Conference on Spoken Language Processing*, Denver, CO.

Arroyo, I., B. Woolf, D. Cooper, W. Burleson, K. Muldner, and R. Christopherson: 2009, 'Emotion Sensors Go To School'. In V. Dimitrova, R. Mizoguchi, B. Du Boulay and A. Graesser (eds.): *14th International Conference on Artificial Intelligence In Education*. Amsterdam: IOS Press.

Asthana, A., J. Saragih, M. Wagner, and R. Goecke: 2009, 'Evaluating AAM Fitting Methods for Facial Expression Recognition'. *International Conference on Affective Computing and Intelligent Interaction*. Amsterdam, The Netherlands

Baker, R., S. D'Mello, M. Rodrigo, and A. Graesser:2010, 'Better to be frustrated than bored: The incidence and persistence of affect during interactions with three different computer-based learning environments'. *International Journal of Human-Computer Studies*, **68(4)**, 223-241.

Banse, R., and K. Scherer:1996, 'Acoustic profiles in vocal emotion expression'. *Journal of Personality and Social Psychology*, **70**, 614-636.

Barrett, L.: 2006, 'Are Emotions Natural Kinds'? *Perspectives on Psychological Science* **1**, 28-58.

Barrett, L., B. Mesquita, K. Ochsner, and J. Gross: 2007, 'The experience of emotion'. *Annual Review of Psychology*, **58**, 373-403.

Biggs, J.: 1995, 'Enhancing teaching through constructive alignment'. 20th *International Conference on Improving University Teaching*, Hong Kong, Hong Kong.

Bower, G.: 1981, 'Mood and memory'. *American Psychologist*, **36**, 129-148.

Brick, T., M. Hunter, and J. Cohn.: 2009, ‚Get The FACS Fast: Automated FACS face analysis benefits from the addition of velocity'. *International Conference on Affective Computing and Intelligent Interaction*. Amsterdam, The Netherlands

Bull, P.: 1987, 'Posture and Gesture'. Oxford Pergamon Press.

Burleson, W., and R. Picard: 2007, 'Evidence for Gender Specific Approaches to the Development of Emotionally Intelligent Learning Companions'. *IEEE Intelligent Systems*, **22**, 62-69.

Caridakis, G., L. Malatesta, L. Kessous, N. Amir, A. Paouzaiou, and K. Karpouzis: 2006, 'Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition'. *International Conference on Multimodal Interfaces*. Cambridge, Massachusetts

Castellano, G., M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer: 2008, 'Automated Analysis of Body Movement in Emotionally Expressive Piano Performances'. *Music Perception*, **26**, 103-119.

Chen, L., T. Huang, T. Miyasato, and R. Nakatsu: 1998, 'Multimodal human emotion/expression recognition.' *Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 366-371.

Chi, M., M. Roy, and R. Hausmann: 2008, 'Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning'. *Cognitive Science*, **32**, 301-341.

Cohen, J.: 1960, 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement*, **20**, 37-46.

Cohn, J., and K. Schmidt: 2004, 'The timing of facial motion in posed and spontaneous smiles'. *International Journal of Wavelets, Multiresolution and Information Processing*, **2**, 1-12.

Conati, C.: 2002, 'Probabilistic assessment of user's emotions in educational games'. *Applied Artificial Intelligence*, **16**, 555-575.

Conati, C., and H. Maclaren: 2009, 'Empirically building and evaluating a probabilistic model of user affect'. *User Modeling and User-Adapted Interaction*, **19**, 267-303.

Craig, S., S. D'Mello, A. Witherspoon, and A. Graesser: 2008, 'Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive-affective states during learning'. *Cognition and Emotion*, **22**, 777-788.

Craig, S., S. D'Mello, A. Witherspoon, J. Sullins, and A. Graesser: 2004, 'Emotions during learning: the first step toward an affect sensitive intelligent tutoring system'. *International Conference on eLearning*, Orlando, Florida, pp. 284-288.

D'Mello, S., S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser: 2005, 'Integrating affect sensors in an intelligent tutoring system'. *The Computer In The Affective Loop Workshop At 2005 International Conference On Intelligent User Interfaces*, New York: AMC Press**,** pp. 7-13.

D'Mello, S., S. Craig, and A. Graesser: 2009, 'Multi-method assessment of affective experience and expression during deep learning'. *International Journal of Learning Technology*, **4**, 165-187.

D'Mello, S., S. Craig, J. Sullins, and A. Graesser: 2006, 'Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue'. *International Journal of Artificial Intelligence in Education*, **16**, 3-28.

D'Mello, S., S. Craig, A. Witherspoon, B. McDaniel, and A. Graesser: 2008, 'Automatic detection of learner's affect from conversational cues'. *User Modeling and User-Adapted Interaction*, **18**, 45-80.

D'Mello, S., and A. Graesser: 2009, 'Automatic detection of learners' affect from gross body language'. *Applied Artificial Intelligence*, ***23***, pp.123-150.

D'Mello, S., G. Jackson, S. Craig, B. Morgan, P. Chipman, H. White, et al.: 2008, 'AutoTutor detects and responds to learners affective and cognitive states'. *Workshop on Emotional and Cognitive issues in ITS held in conjunction with the Ninth International Conference on Intelligent Tutoring Systems*, Montreal, Canada.

D'Mello, S., B. King, O. Entezari, P. Chipman, and A. Graesser: 2008, 'The impact of automatic speech recognition errors on learning gains with AutoTutor'. *Annual meeting of the American Educational Research Association*, New York, New York.

D'Mello, S., R. Picard, and A. Graesser: 2007, 'Towards an affect-sensitive AutoTutor'. *Intelligent Systems, IEEE*, **22**, 53-61.

D'Mello, S., S. Craig, K. Fike, and A. Graesser: 2009, 'Responding to learners' cognitive-affective states with supportive and shakeup dialogues'. In J. Jacko (ed.): *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction* Berlin/Heidelberg: Springer, pp. 59-604.

D'Mello, S., N. Dowell, and A. Graesser: 2009, 'Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States'. In V. Dimitrova, R. Mizoguchi, B. du Boulay and A. Graesser (eds.): *14th International Conference on Artificial Intelligence In Education.* Amsterdam: IOS Press, pp. 9-16

Damasio, A.: 2003, 'Looking for Spinoza: Joy, sorrow, and the feeling brain'. Harcourt Inc.

Dasarathy, B.: 1997, 'Sensor fusion potential exploitation: Innovative architectures and illustrative approaches'. *IEEE*, **85**, pp.24-38.

de Rosis, F., C. Castelfranchi, P. Goldie, and V. Carofiglio: (in press), 'Cognitive Evaluations And Intuitive Appraisals: Can Emotion Models Handle Them Both'? *HUMAINE Handbook*. Berlin: Springer.

De Vicente, A., and H. Pain: 2002, 'Informing the detection of the students' motivational state: An empirical study. In S. A. Cerri, G. Gouarderes and F. Paraguacu: (eds.): *6th International Conference on Intelligent Tutoring Systems*, San Sebastian, Spain, pp. 933-943.

Dodds, P., and J. Fletcher: 2004, 'Opportunities for new "smart" learning environments enabled by next-generation web capabilities'. *Journal of Educational Multimedia and Hypermedia*, **13**, 391-404.

Ekman, P.: 1984, 'Expression and the nature of emotion'. In K. Scherer and P. Ekman (eds.), *Approaches to emotion*, Hillsdale, NJ: Erlbaum, pp. 319-344.

Ekman, P.: 1992, 'An Argument for Basic Emotions'. *Cognition and Emotion*, **6**, 169-200.

Ekman, P.: 2002,. 'Darwin, deception, and facial expression'. *Conference on Emotions Inside Out, 130 Years after Darwin's the Expression of the Emotions in Man and Animals*, New York, New York.

Ekman, P., and W. Friesen: 1969, 'Nonverbal Leakage and Clues to Deception'. *Psychiatry*, **32**, 88-105.

Ekman, P., and W. Friesen.: 1975, 'Unmasking the face: A guide to recognizing emotions from facial expressions'. Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P., and W. Friesen: 1978, 'The Facial Action Coding System: A technique for the measurement of facial movement'. Palo Alto: Consulting Psychologists Press.

Ekman, P., W. Friesen, and R. Davidson: 1990, 'The Duchenne Smile - Emotional Expression and Brain Physiology .2'. *Journal of Personality and Social Psychology*, **58**, 342-353.

Fernandez, R., and R. Picard: 2005, 'Classical and Novel Discriminant Features for Affect Recognition from Speech'. 9th European Conference on Speech Communication and Technology.

Fleiss, J.: 1981, 'Statistical Methods for Rates and Proportions' (2nd ed.). New York: John Wiley and Son.

Forbes-Riley, K., M. Rotaru, and D. Litman: 2008, 'The relative impact of student affect on performance models in a spoken dialogue tutoring system'. *User Modeling and User-Adapted Interaction*, **18**, 11-43.

Gertner, A., and K. VanLehn: 2000, 'Andes: A coached problem solving environment for physics'. In G. Gauthier, C. Frasson and K. VanLehn (eds.): *International Conference on Intelligent Tutoring Systems*, Berlin / Heidelberg: Springer, pp. 133-142.

Graesser, A., P. Chipman, B. Haynes, and A. Olney: 2005, 'AutoTutor: An intelligent tutoring system with mixed-initiative dialogue'. *IEEE Transactions on Education*, **48**, 612-618.

Graesser, A., S. L. Lu, G. Jackson, H. Mitchell, M. Ventura, A. Olney, et al.: 2004, 'AutoTutor: A tutor with dialogue in natural language'. *Behavioral Research Methods, Instruments, and Computers*, **36**, 180-193.

Graesser, A., B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and B. Gholson: 2006, 'Detection of emotions during learning with AutoTutor'. *28th Annual Conference of the Cognitive Science Society*, Vancouver, Canada.

Graesser, A., D. McNamara, and K. VanLehn: 2005, 'Scaffolding deep comprehension strategies through PointandQuery, AutoTutor, and iSTART,. *Educational Psychologist*, **40**, 225-234.

Graesser, A., P. Penumatsa, M. Ventura, Z. Cai, and X. Hu: 2007, 'Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language'. In T. Landauer, D. McNamara, S. Dennis and W. Kintsch (eds.): *Handbook of Latent Semantic Analysis* Mahwah, NJ: Erlbaum, pp. 243-262.

Graesser, A., K. VanLehn, C. P. Rose, P. W. Jordan, and D. Harter: 2001, 'Intelligent tutoring systems with conversational dialogue'. *AI Magazine*, **22**, 39-51.

Hocking, R.: 1976, 'Analysis and selection of variables in linear regression'. *Biometrics*, **32**, 1-49.

Hoque, M. E., R. el Kaliouby, and R. W. Picard: 2009, 'When Human Coders (and Machines) Disagree on the Meaning of Facial Affect in Spontaneous Videos'. *9th International Conference on Intelligent Virtual Agents*, Amsterdam.

Hudlicka, E., and M. McNeese: 2002, 'Assessment of user affective and belief states for interface adaptation: Application to an Air Force pilot task'. *User Modeling and User-Adapted Interaction*, **12**, 1-47.

Jaimes, A., and N. Sebe: 2007, 'Multimodal human-computer interaction: A survey'. *Computer Vision and Image Understanding*, **108**, 116-134.

Johnstone, T., and K. Scherer: 2000, Vocal communication of emotion. In M. Lewis and J. Haviland-Jones (eds.): *Handbook of Emotions* (2nd ed.), New York: Guilford Press, pp. 220-235.

Jonassen, D., K. Peck, and B. Wilson: 1999, 'Learning with technology: A constructivist perspective'. Upper Saddle River, NJ: Prentice Hall.

Kapoor, A., B. Burleson, and R. Picard: 2007, 'Automatic prediction of frustration'. *International Journal of Human-Computer Studies*, **65**, 724-736.

Kapoor, A., and R. Picard: 2005, 'Multimodal affect recognition in learning environments'. *13th annual ACM international conference on Multimedia*, Hilton, Singapore.

Keltner, D., and P. Ekman: 2000, 'Facial expression of emotion'. In R. Lewis and J. M. Haviland-Jones (eds.): *Handbook of emotions* (Vol. 2nd ed.), New York: Guilford, pp. 236–264.

Klecka, W.: 1980, 'Discriminant Analysis'. Beverly Hills, CA: Sage.

Koedinger, K., J. Anderson, W. Hadley, and M. Mark: 1997, 'Intelligent tutoring goes to school in the big city'. *International Journal of Artificial Intelligence in Education,* **8**, 30-43.

Koedinger, K., and A. Corbett: 2006, 'Cognitive tutors: Technology bringing learning sciences to the classroom'. In R. K. Sawyer (ed.): *The Cambridge handbook of the learning sciences*. New York, NY: Cambridge University Press, pp. 61-78.

Kort, B., R. Reilly, and R. Picard: 2001, 'An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion' *IEEE International Conference on Advanced Learning Technologies*, Madison, Wisconsin

Landauer, T., and S. Dumais: 1997, A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge'. *Psychological Review,* **104**, 211-240.

Landauer, T., D. McNamara, S. Dennis, and, W. Kintsch (eds.): 2007, 'Handbook of Latent Semantic Analysis'. Mahwah, NJ: Erlbaum.

Landauer, T., D. McNamara, S. Dennis, and, W. Kintsch (eds.): 2008, 'Handbook of Latent Semantic Analysis'. Mahwah, NJ: Erlbaum.

Lee, C., and S. Narayanan: 2005, 'Toward detecting emotions in spoken dialogs'. *IEEE Transactions on Speech and Audio Processing, 13*, pp. 293-303.

Lehman, B., S. D'Mello, and N. Person: 2008, 'All Alone with your Emotions: An Analysis of Student Emotions during Effortful Problem Solving Activities'. *Workshop on Emotional and Cognitive issues in ITS at the Ninth International Conference on Intelligent Tutoring Systems*. Montreal, Canada

Lehman, B., M. Matthews, S. D'Mello, and N. Person: 2008, 'What are you feeling? Investigating student affective states during expert human tutoring sessions'. In B. Woolf, A. E., N. R. and L. S. (eds.):, *9th International Conference on Intelligent Tutoring Systems*, Montreal, Canada, pp. 50-59.

Liscombe, J., G. Riccardi, and D. Hakkani-Tür: 2005, 'Using Context to Improve Emotion Detection in Spoken Dialog Systems'. *9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.

Litman, D., and K. Forbes-Riley: 2004, 'Predicting student emotions in computer-human tutoring dialogues'. *42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Madsen, M., R. el Kaliouby, M. Goodwin, and R. Picard: 2008, 'Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder'. *10th ACM Conference on Computers and Accessibility*. Halifax, Canada.

Mandler, G.: 1976, 'Mind and emotion'. New York: Wiley.

Mandler, G.: 1984, 'Mind and Body: Psychology of Emotion and Stress'. New York: W.W. Norton and Company.

Marsic, I., A. Medl, and J. Flanagan: 2000, 'Natural communication with information systems'. *IEEE,* **88**, 1354-1366.

McDaniel, B., S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser: 2007, 'Facial Features for Affective State Detection in Learning Environments'. In D. McNamara and G. Trafton (eds.): *29th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society, pp. 467-472.

McQuiggan, S., B. Mott, and J. Lester: 2008, 'Modeling self-efficacy in intelligent tutoring systems: An inductive approach'. *User Modeling and User-Adapted Interaction,* **18**, 81-123.

Mitchell, T.: 1997, 'Machine Learning'. Mc-Graw-Hill.

Moshman, D.: 1982. 'Exogenous, endogenous, and dialectical constructivism'. *Developmental Review, **2***, 371-384.

Norman, D.: 1994, 'How might people interact with agents'. *Communications of the ACM, **37***, 68-71.

Olney, A., M. Louwerse, E. Mathews, J. Marineau, H. Hite-Mitchell, and A. Graesser: 2003, 'Utterance classification in AutoTutor.' *Human Language Technology - North American Chapter of the Association for Computational Linguistics Conference*, Edmonton, Canada.

Ortony, A., and T. Turner: 1990, 'What's Basic About Basic Emotions'. *Psychological Review,* **97**, 315-331.

Paiva, A., R. Prada, and R. Picard,: (eds.): 2007, 'Affective Computing and Intelligent Interaction'. Heidelberg: Springer.

Panksepp, J.: 2000, '*Emotions as natural kinds within the mammalian brain*'. In M. Lewis and J. M. Haviland-Jones (eds.): *Handbook of emotions* (2nd ed.). New York: Guilford, pp. 137-156.

Pantic, M., and I. Patras: 2006, 'Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences'. *IEEE Transactions on Systems, Man, and Cybernetics, Part B.*, **36**, 433-449.

Pantic, M., and L. Rothkrantz: 2003, 'Toward an affect-sensitive multimodal human-computer interaction'. *Proceedings of the IEEE*, **91**,1370-1390.

Pentland, A.: 2000, 'Looking at people: Sensing for ubiquitous and wearable computing'. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, pp.107-119.

Picard, R.: 1997, 'Affective Computing. Cambridge', Mass: MIT Press.

Picard, R., E. Vyzas, and J. Healey: 2001, 'Toward machine emotional intelligence: Analysis of affective physiological state'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**, 1175-1191.

Porayska-Pomsta, K., M. Mavrikis, and H. Pain: 2008, 'Diagnosing and acting on student affect: the tutor's perspective'. *User Modeling and User-Adapted Interaction*, **18**, 125-173.

Prendinger, H., and M. Ishizuka: 2005, 'The empathic companion: A character-based interface that addresses users' affective states'. *Applied Artificial Intelligence*, **19**, 267-285.

Robson, C.: 1993, 'Real world research: A resource for social scientist and practitioner researchers'. Oxford: Blackwell.

Rus, V., and A. Graesser: 2007, 'Lexico-syntactic subsumption for textual entailment'. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.): *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 200*. Amsterdam: John Benjamins Publishing Company, pp. 187-196.

Russell, J.: 1994, 'Is There Universal Recognition of Emotion from Facial Expression - a Review of the Cross-Cultural Studies'. *Psychological Bulletin*, **115**, 102-141.

Russell, J.: 2003, 'Core affect and the psychological construction of emotion'. *Psychological Review*, **110**, 145-172.

Scherer, K.: 2003, 'Vocal communication of emotion: A review of research paradigms'. *Speech Communication*, **40**, 227-256.

Scherer, K., and H. Ellgring: 2007, 'Multimodal expression of emotion: Affect programs or componential appraisal patterns'? *Emotion*, **7**, 158-171.

Scherer, K., T. Johnstone, and G. Klasmeyer: 2003, 'Vocal expression of emotion'. In R. J. Davidson, K. R. Scherer and H. Goldsmith (eds.): *Handbook of the Affective Sciences*. New York and Oxford: Oxford University Press, pp. 433–456.

Shneiderman, B., and C. Plaisant: 2005, 'Designing the user interface: Strategies for effective human-computer interaction'. Reading, MA: Addison-Wesley.

Storey, J., K. Kopp, K. Wiemer, P. Chipman, and A. Graesser (in press), 'Critical thinking tutor: Using AutoTutor to teach scientific critical thinking skills'. *Behavioral Research Methods*.

Turner, T., and A. Ortony: 1992, 'Basic Emotions - Can Conflicting Criteria Converge'. *Psychological Review*, **99**, 566-571.

VanLehn, K.: 1990, Mind bugs: The origins of procedural misconceptions'. Cambridge, MA: MIT Press.

VanLehn, K., A. Graesser, G. Jackson, P. Jordan, A. Olney, and C. Rose: 2007. 'When are tutorial dialogues more effective than reading'? *Cognitive Science*, **31**, 3-62.

VanLehn, K., C. Lynch, K. Schulze, J. Shapiro, R. Shelby, L. Taylor et al.: 2005, 'The Andes physics tutoring system: five years of evaluations'. *International Journal of Artificial Intelligence in Education* **15**, 147-204.

Woolf, B., W. Burleson, and I. Arroyo: 2007, *Emotional intelligence for computer tutors. Workshop on Modeling and Scaffolding Affective Experiences to Impact Learning at 13th International Conference on Artificial Intelligence in Education*, Los Angeles, California.

Yoshitomi, Y., K. Sung-Ill, T. Kawano, and T. Kilazoe: 2000, 'Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face'. IEEE International Workshop on *Robots* and Human Interactive Communications, Osaka, Japan.

Zeng, Z., Y. Hu, G. Roisman, Z. Wen, Y. Fu, and T. Huang: 2006, 'Audio-visual emotion recognition in adult attachment interview'. *International Conference on Multimodal Interfaces*, Alberta, Canada.

Zeng, Z., M. Pantic, G. Roisman, and T. Huang: 2009, 'A survey of affect recognition methods: Audio, visual, and spontaneous expressions'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 39-58.

Appendix A. Comparison between partial and full feature sets for composite models

Table 8. Kappa scores for partial and full feature sets

			Kappa			
	No. Features		**Fixed**		**Spontaneous**	
Channel	*Partial*	*Full*	*Partial*	*Full*	*Partial*	*Full*
Two-Channels						
FD	10	19	.271	.247	.391	.384
FP	10	20	.205	.146	.361	.351
DP	10	19	.242	.234	.198	.223
Three-Channels						
FDP	9	29	.288	.326	.382	.390
Mean	**-**	**-**	**.252**	**.236**	**.333**	**.337**

Authors' Vitae

Sidney D'Mello
University of Memphis, Institute for Intelligent Systems,
202 Psychology Building, Memphis, TN, 38152, USA

Sidney D'Mello is Postdoctoral Researcher in the Institute for Intelligent Systems at the University of Memphis. D'Mello received his B.S. in Electrical Engineering from Christian Brothers University, his M.S. in Mathematical Science, and his Ph.D. in Computer Science from the University of Memphis. D'Mello has worked in several areas of artificial intelligence and cognitive science, including affective computing, intelligent tutoring systems, speech and language processing, human like learning in machines, and computational models of human cognition. He has authored over 70 papers in these areas.

Arthur Graesser
University of Memphis, Institute for Intelligent Systems,
202 Psychology Building, Memphis, TN, 38152, USA

Art Graesser is a professor in the Department of Psychology and co-director of the Institute for Intelligent Systems at the University of Memphis. Graesser received his B.A. in Psychology at Florida State University and his Ph.D. in psychology from the University of California at San Diego. Graesser has worked in several areas of cognitive science, artificial intelligence, and discourse processing, including knowledge representation, question asking and answering, tutoring, text comprehension, inference generation, conversation, reading, memory, and human-computer interaction. He has authored over 400 technical articles, two books, and edited nine books.