

## **NOTE TO USERS**

**This reproduction is the best copy available.**

**UMI<sup>®</sup>**



AFFECT MEASUREMENT BY MAN AND MACHINE

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Memphis

Sidney K. D'Mello

August, 2009

UMI Number: 3400141

All rights reserved

**INFORMATION TO ALL USERS**

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3400141

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.

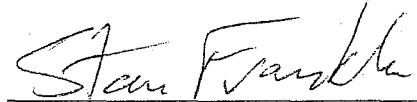


ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

Copyright © 2009 Sidney K. D'Mello  
All rights reserved

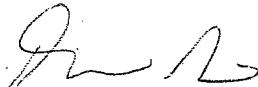
To the Graduate Council:

I am submitting herewith a dissertation written by Sidney Keith D'Mello entitled "Affect Measurement by Man and Machine." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Computer Science.

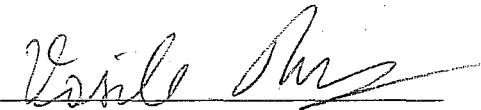


Stan Franklin, Ph.D.  
Major Professor

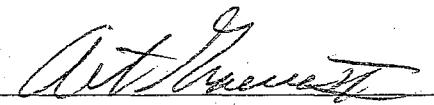
We have read this dissertation and  
recommend its acceptance:



King-Ip Lin, Ph.D.

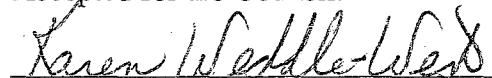


Vasile Rus, Ph.D.



Arthur Graesser, Ph.D.

Accepted for the Council:



Karen D. Weddle-West, Ph.D.  
Vice Provost for Graduate Programs

## Acknowledgements

This dissertation would not have been possible without my colleagues in the Emotive Computing Group at the University of Memphis and the Affective Computing Group at MIT. I gratefully acknowledge Scotty Craig, Amy Witherspoon, Jeremiah Sullins, Bethany McDaniel, Kristy Tapp, and Patrick Chipman for their invaluable assistance in data collection, emotion coding, and facial expression coding. Thanks to Barb Hertel for proofreading this dissertation.

I would like to thank Steelcase Inc. for providing the Tekscan Body Pressure Measurement System and the National Science Foundation for funding this work.

I am deeply indebted to my advisor Stan Franklin for his support, patience, guidance, encouragement, and overall grace and brilliance over the past seven years. This dissertation would not be possible without him.

I am also deeply indebted to my co-advisor and collaborator Art Graesser for introducing me to the wonderful world of emotion, cognition, and learning, and then setting me free to explore my own ideas. This dissertation would not have been possible without his intellectual and emotional support over the last five years.

Finally, I would like to thank my parents and my family for their enduring love and support.

# **Abstract**

D'Mello, Sidney Keith. PhD. The University of Memphis. August/2009 Computer Science. Affect Measurement by Man and Machine. Stan Franklin

The field of affective computing aspires to narrow the communicative gap between the highly expressive human and the socially challenged computer by developing computational systems that recognize and respond to the affective states of the user. This dissertation addresses one of the fundamental goals of affective computing by developing computational mechanisms that automatically detect users' affective states (or emotions) during naturalistic interactions with computer interfaces.

This dissertation describes systems that discriminate between the affective states (e.g., boredom, flow/engagement, confusion, frustration, delight, surprise, and neutral) by monitoring conversational cues, gross body language, and facial features. Training and validation data for the affect detectors were obtained from a study in which 28 learners completed a tutorial session with an intelligent tutoring system after which their affective states were judged by the learners themselves, untrained peers, and two trained judges.

Affect detectors that automatically classified learners' affective states in real time by monitoring lexical and semantic features from the tutorial dialogue were developed. A feature selection algorithm for situations in which there is ambiguity in the affect

categories was developed to select features that were most diagnostic of the affective states and that were expected to generalize above and beyond individual differences.

Two algorithms to detect affect from gross body language were developed. The first algorithm tracked the average pressure exerted, along with the magnitude and direction of changes in pressure during emotional experiences. The second algorithm monitored the spatial and temporal properties of naturally occurring pockets of pressure. A hierarchical classification algorithm motivated by the pandemonium model was also developed and evaluated.

A feature-level and a decision-level multi-modal affect detector that combined conversational cues, gross body language, and facial features was developed and evaluated. A naïve additive algorithm was considered for feature-level fusion, while a spreading activation network architecture with differential weighting was used to model decision-level fusion.

The results on human judgments of emotions indicated that (a) boredom, flow/engagement, confusion, and frustration were the major affective states that learners experienced, (b) affect detection accuracies of human judges was low, and (c) trained judges provided more accurate judgments than untrained peers.

Machine learning analyses that independently evaluated each sensory channel revealed that (a) monitoring gross body language was most effective for detecting boredom (74%) and flow/engagement (83%), (b) confusion (76%) and delight (90%) were best detected by monitoring facial features, (c) frustration was best detected by

examining the dialogue features in the tutoring context (78%), and (d) detection accuracies were 80% when particular emotions were aligned with the optimal sensors.

Classification results on combinations of sensory channels indicated that (a) the face was the most diagnostic channel for voluntary affect judgments, while conversational cues were superior for mandatory judgments, (b) combination of channels yielded superadditive effects for some states, but additive, redundant, and inhibitory effects for others, (c) multi-channel models reduced the discrepancy of single-channel models, and (d) decision-level fusion yielded accuracy scores that were equivalent with feature-level fusion.

This dissertation concludes that conversational cues and gross body language are serious contenders to existing affect detection methods that focus on facial features and acoustic-prosodic cues. Limitations, solutions, improvements, and extensions of this research are discussed. Finally, a case study that describes an application of the affect detectors developed in this research is presented.

# Table of Contents

Acknowledgements.....	iii
Abstract.....	iv
Table of Contents.....	vii
List of Tables .....	xiii
List of Figures.....	xvi
Chapter 1: Introduction to Affect Measurement .....	1
The Affective Age of Human–Computer Interaction .....	1
Affect-Sensitive Intelligent Tutoring Systems .....	3
Defining the Affect Measurement Problem.....	7
Challenges Facing Affect Measurement.....	10
Individual Differences in Affect Experience and Expression .....	10
Validity Concerns in Affect Measurement.....	12
Limitations of Current Affect Detection Systems .....	16
Problems with Sensors .....	17
Methodological Concerns.....	19
Concerns Pertaining to Learning Environments.....	21
An Interdisciplinary Framework for Automated Affect Measurement .....	22
Solutions to Sensor Problems: Tracking Conversational Cues and Gross Body Language	23
Solutions to Methodological Problems: Multiple Measures of Ground Truth.....	25

Solutions to Problems Concerning Affect Measurement in Learning Environments .....	26
Research Overview .....	27
Research Goals .....	27
Accomplishments and Pending Work .....	28
Research Novelty .....	32
Research Contributions .....	33
Dissertation Organization .....	34
Chapter 2: Affect Measurement by Humans.....	37
Introduction .....	37
The AutoTutor Learning Environment.....	39
The Structure of AutoTutor Dialogue .....	42
Interpreting Learners' Contributions.....	46
Methods .....	47
Participants .....	47
Content Covered in the Tutorial Session with AutoTutor .....	47
Materials and Equipment.....	48
Protocols.....	51
Procedure.....	53
Results and Discussion .....	55
What Emotions Occur During Complex Learning? .....	56
Inter-judge Agreement on JOEs .....	60
General Discussion .....	65
Chapter 3: Automatic Affect Detection from Conversational Cues .....	68
Introduction .....	68

Computational Challenges and Solutions .....	72
Features of AutoTutor's Mixed-Initiative Dialogue .....	77
Temporal Information .....	79
Response Information.....	79
Answer Quality Assessments .....	80
Tutor Directness .....	80
Tutor Feedback.....	81
Results and Discussion .....	82
Data Visualization and Preliminary Predictions.....	84
Illustration of Feature Selection Algorithm.....	87
Dimensionality Reduction.....	94
Classifying Affective States from Conversation Features.....	96
General Discussion .....	108
Research Overview.....	110
Limitations.....	113
Chapter 4: Automatic Affect Detection from Gross Body Language .....	116
Introduction .....	116
Architecture of the Posture Based Affect Detector .....	121
The Body Pressure Measurement System (BPMS) .....	121
High Level Pressure Features .....	122
Spatial-Temporal Features.....	126
Hierarchical Classification via an Associated Pandemonium .....	128
Measuring the Accuracy of the Posture Based Affect Detector .....	132
Experimental Setup .....	132

Trends in Classification Accuracy.....	135
Maximum Classification Accuracy .....	140
General Discussion .....	143
Chapter 5: Automated Affect Detection by Combining Sensors.....	145
Introduction .....	145
Mechanisms for Feature and Decision-Level Fusion .....	149
Feature-Level Fusion.....	149
Decision-Level Fusion.....	150
Superadditive, Additive, Redundant, or Inhibitory Effects .....	155
Data Sets .....	158
Data Sampling .....	160
Extracting Features from Sensors.....	161
Feature-Level Fusion Models .....	165
Single-Channel Models .....	169
Multi-Channel Models.....	171
Discrepancy Reduction.....	173
Effects for Individual Emotions .....	176
Structure of Composite Discriminate Models .....	181
Decision-Level Fusion.....	189
Weighting Schemes .....	190
Classification Accuracy.....	194
General Discussion .....	197
Comparing Face, Dialogue, and Posture .....	197
Single-Channel versus Multiple-Channel Affect Detection .....	199

Superadditive Effects.....	201
Limitations.....	203
Future Work .....	205
Chapter 6: Concluding Discussion on Affect Detectors .....	207
Introduction .....	207
Accuracy of Automated Affect Detectors .....	208
Incorporating Temporal Context in Affect Sensing Methods.....	211
Graded Difference in the Relative Durations of the Affective States.....	211
Transitions between the Affective States .....	214
Scalability of Affect Detectors .....	218
Text-Based (Language-Based) Affect Detectors.....	218
Camera-Based Body Position and Motion Tracking.....	221
Individual Differences in Affective Experience and Expression.....	224
Applications of Affect Detectors .....	228
Strategies to Respond to Learners' Affective States .....	229
Evaluating the Affect-Sensitive AutoTutor.....	235
References.....	238
Appendix A. Affect Measurement by Humans .....	271
Appendix B. Affect Classification from Conversational Cues.....	274
Detailed classification results for 4-way emotion discriminations.....	274
Detailed classification results for affect-neutral discriminations .....	275
Appendix C. Affect Classification from Gross Body Language .....	279
Detailed classification results for affect-neutral discriminations .....	279
Detailed classification results for 2-way classifications.....	284

Detailed classification results for 3-way classifications.....	289
Detailed classification results for 4-way classifications.....	294
Detailed classification results for 5-way classifications.....	299

## List of Tables

Table 1. Sample conversation between AutoTutor and a student.....	42
Table 2. Descriptive statistics for proportions of emotions observed.....	59
Table 3. Descriptive statistics for kappas between raters for judging affect.....	62
Table 4. Simulated output of feature selection algorithm.....	76
Table 5. Description of the information mined from AutoTutor's log files. ....	78
Table 6. Frequency of affective states in each data set. ....	83
Table 7. Summaries of the multiple regression models for emotions in each data set. ....	89
Table 8. Significant predictors for the multiple regression models. ....	91
Table 9. Comparison of various classification techniques to detect learner's affect. ....	103
Table 10. F-Measure for emotions.....	104
Table 11. Comparisons of computer and human classifications.....	107
Table 12. Frequency of affective states in each data set. ....	133
Table 13. Maximum classification accuracy in detecting affect.....	141
Table 14. Description of action units, incidence of occurrence, and kappa scores.....	165
Table 15. Features included in the various classification models. ....	168
Table 16. Base rate corrected precision scores for emotions.....	178
Table 17. Structure matrix for mandatory FDP model. ....	186
Table 18. Structure matrix for voluntary FDP model. ....	188
Table 19. Modulation used in decision-level fusion. ....	193

Table 20. Mean difference in kappa scores for judge $\times$ judgment type interaction.....	271
Table 21. Mean difference in kappa scores for mandatory judgments.....	272
Table 22. Mean difference in kappa scores for voluntary judgments.....	273
Table 23. Classification accuracies for 4-way discrimination.....	274
Table 24. Classification accuracies in discriminating between boredom and neutral.....	275
Table 25. Classification accuracies in discriminating between confusion and neutral.....	276
Table 26. Classification accuracies in discriminating between flow and neutral.....	277
Table 27. Classification accuracies in discriminating between frustration and neutral.....	278
Table 28. Affect neutral classification for self judgments.....	279
Table 29. Affect neutral classification for peer judgments.....	280
Table 30. Affect neutral classification for judgments by trained judge 1.....	281
Table 31. Affect neutral classification for judgments by trained judge 2.....	282
Table 32. Affect neutral classification for judgments where trained judges agree.....	283
Table 33. Two-way classification for self judgments.....	284
Table 34. Two-way classification for peer judgments.....	285
Table 35. Two-way classification for judgments by trained judge 1.....	286
Table 36. Two-way classification for judgments by trained judge 2.....	287
Table 37. Two-way classification for judgments where trained judges agree.....	288
Table 38. Three-way classification for self judgments.....	289
Table 39. Three-way classification for peer judgments.....	290
Table 40. Three-way classification for judgments by trained judge 1.....	291
Table 41. Three-way classification for judgments by trained judge 2.....	292
Table 42. Three-way classification for judgments where trained judges agree.....	293
Table 43. Four-way classification for judgments by self judgments.....	294

Table 44. Four-way classification for judgments by peer judges.....	295
Table 45. Four-way classification for judgments by trained judge 1.....	296
Table 46. Four-way classification for judgments by trained judge 2.....	297
Table 47. Four-way classification for judgments where trained judges agree.....	298
Table 48. Five-way classification accuracies.....	299

## List of Figures

Figure 1. Sample facial displays of learning-centered emotions. ....	38
Figure 2. Screenshot of the AutoTutor interface.....	41
Figure 3. Sensors used during data collection.....	49
Figure 4. Two-dimensional representations of each emotion versus neutral.....	85
Figure 5. Two-dimensional representations of various combinations of the affective states. ....	86
Figure 6. Mean kappa across: (a) Affect Judge; (b) Emotions Classified; (c) Classifier Type....	100
Figure 7. Body pressure measurement system.....	122
Figure 8. Clustering pressure maps on the back and the seat with the EM algorithm. ....	128
Figure 9. Hierarchical classification via an associated pandemonium.....	130
Figure 10. Mean kappa for affect detection. ....	136
Figure 11. Sample activation spreading network for decision-level fusion.....	152
Figure 12. Overall classification results for feature-level fusion. ....	172
Figure 13. Single versus multi channel precision scores for individual emotions. ....	175
Figure 14. Precision scores for each emotion for the best one, two, and three channel models..	180
Figure 15. Group centroids for mandatory FDP model. ....	184
Figure 16. Group centroids for voluntary FDP model. ....	187
Figure 17. Weighting functions used for decision-level fusion. ....	191
Figure 18. Kappa scores for decision-level FDP models. ....	195
Figure 19. Precision scores for emotions, corrected for base rate. ....	196

Figure 20. Exponential decay curves for the learning centered emotions .....	214
Figure 21. Predicted affective transitions .....	215
Figure 22. Sample frames from motion tracking algorithm.....	223
Figure 23. Affect synthesis by embodied pedagogical agents.....	232

# **Chapter 1: Introduction to Affect Measurement**

## **The Affective Age of Human–Computer Interaction**

The field of interface development was radically transformed when design decisions began to encompass constraints of the body and mind of users in addition to technical concerns that had previously dominated system development. Stemming from the human factors movement in the early 50s and transcending to the cognitive revolution of the sixties and seventies, the impetus of man-machine interaction began to gradually shift away from the machine and more towards the man.

However, throughout the HCI revolution, affective experiences (emotions, moods, feelings) of the user were excluded from the realm of human factors, AI, and cognitive science (McNeese, 2003). Instead, design decisions seemed to primarily focus on the cognitive constraints of the user. But humans are more than mere cognitive machines because emotions are an inextricable part of everyday experience. Physiological arousal and its cognitive interpretation imply that there is a complex interplay between cognition and emotion (Mandler, 1976, 1984b). Therefore, an interface that is sensitive to a learner's affective state is expected to be more usable, useful, naturalistic, social, and enjoyable - all factors that guarantee wide use and acceptance.

A historic walkthrough along the vested research foci of HCI leads to two plausible reasons for the exclusion of affect in HCI. First, emotion has been traditionally considered to be a dominant factor in clinical psychology and considered to have no impact on intelligent behavior (i.e., feelings get in the way of rational thought). Second, the adoption of the information processing metaphor as a vestibule to understanding human cognition had little room for emotion (Boehner, DePaula, Dourish, & Sengers, 2007).

The recent inclusion of a social interactional approach to cognition and thereby HCI has led to a departure from the traditional information-processing paradigm toward the emotional realm (Boehner et al., 2007). Stemming from Picard's influential 1997 book, the last decade has witnessed a surge of research activities that support the notion of affective computing. Generally speaking, affective computing focuses on creating technologies that can monitor and appropriately respond to the affective states of the user (Picard, 1997). Such systems attempt to bridge the communicative gap between the emotionally expressive human and the emotionally deficit computer.

Consequently, there has been an eruption of research activities that aspire to incorporate the affective states of a user into the decision cycle of the interface in an attempt to develop more effective, user-friendly, and naturalistic applications (Hudlicka & McNeese, 2002; Klein, Moon, & Picard, 2002; Mandryk & Atkins, 2007; Prendinger & Ishizuka, 2005; Whang, Lim, & Boucsein, 2003). One application genre that is hypothesized to be a particularly good candidate for improvement by addressing affective states is intelligent tutoring systems (ITSs) (Aist, Kort, Reilly, Mostow, & Picard, 2002;

Burleson & Picard, 2007; Conati, 2002; D'Mello, Craig, Gholson et al., 2005; D'Mello, Picard, & Graesser, 2007; De Vicente & Pain, 2002; McQuiggan, Mott, & Lester, 2008; Woolf, Burleson, & Arroyo, 2007), particularly due to the well-known connections between affect and learning (Arnold, 1999; Bower, 1992; Meyer & Turner, 2006; Stein & Levine, 1991; Sylwester, 1994).

## Affect-Sensitive Intelligent Tutoring Systems

Intelligent tutoring systems 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). ITSs have emerged as valuable systems to promote active learning with learning gains associated with sophisticated ITSs at around 1.0 sigma<sup>1</sup> improvement (Dodds & Fletcher, 2004; Koedinger & Corbett, 2006; VanLehn et al., 2007), higher than those of inexperienced human tutors (0.4 SD) but not quite as good as expert human tutors have achieved (2.0 sigma) (Bloom, 1984).

---

<sup>1</sup> Effect size in standard deviation units. One sigma is approximately equal to a letter grade.

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 a learner's affective state. This has been a long desired goal for ITS developers. For example, one of the first suggestions for endowing computer tutors with a degree of empathy or affect was made by Lepper and Chabay (1988). They claim that ITSs should include a mechanism for motivating the learner, detecting the learner's emotional/motivational state, and appropriately responding to that state (Issroff & del Soldato, 1996; Lepper & Chabay, 1988; Lepper & Woolverton, 2002). Progress in achieving the primary goal requires an interdisciplinary integration of computer science, psychology, artificial intelligence, and artifact design.

De Vicente and Pain (2002) have argued that motivation components are as important as cognitive components in tutoring strategies, and that important benefits would arise from considering techniques that track the learner's motivation and emotions. There is some evidence, for example, that tracking and responding to human emotions on a computer increases students' persistence (Aist et al., 2002). Kim conducted a study that demonstrated that the interest and self-efficacy of a learner significantly increased when the learner was accompanied by a pedagogical agent acting as a virtual learning companion that is sensitive to the learner's affect (Kim, 2005). It has also been reported that the posttest scores of physics understanding decreased as a function of negative affect during learning (Linnenbrink & Pintrich, 2002). Craig and colleagues reported that increased levels of boredom were negatively correlated with the learning of computer literacy, whereas increased levels of confusion and the state of flow [being absorbed in

the learning process (Csikszentmihalyi, 1990)] were positively correlated with learning in an AutoTutor learning environment (Craig, Graesser, Sullins, & Gholson, 2004). More recently, it has been documented that in constructivist learning contexts, learner affect (certainty and frustration) were stronger predictors of learning than answer correctness and other generic parameters (Forbes-Riley, Rotaru, & Litman, 2008).

Over the last few years there has been work toward incorporating assessments of the learner's affect into intelligent tutoring systems (Conati, 2002; Conati & Maclarens, in press; D'Mello, Craig, Gholson et al., 2005; D'Mello, Jackson et al., 2008; D'Mello, Picard et al., 2007; De Vicente & Pain, 2002; Forbes-Riley et al., 2008; Kort, Reilly, & Picard, 2001; McQuiggan et al., 2008; Woolf et al., 2007). For example Kort, Reilly, and Picard (2001) proposed a comprehensive four-quadrant model that explicitly links learning and affective states. This model was used in the MIT group's work on their *affective learning companion*, a fully automated computer program that recognizes a learner's affect by monitoring facial features, posture patterns, and onscreen keyboard/mouse behaviors (Burleson & Picard, 2007).

De Vicente and Pain (2002) developed a system that could track several motivational and emotional states during use of an intelligent tutoring system, building on judgments of emotion by expert coders. Conati (2002) developed a probabilistic system that can reliably track multiple learner affective states (including joy and distress) of the learner during interactions with an educational game, and used these assessments to drive the behavior of an intelligent pedagogical agent (Conati & Maclarens, in press; Conati & Zhou, 2004). Another example of research aimed at developing affect-sensitive

ITSs involves Litman and Forbes-Riley's work with the ITSPOKE (Litman & Silliman, 2004) conceptual physics ITS has used a combination of discourse markers and acoustic-prosodic cues to detect and respond to a learner's affective states (Forbes-Riley & Litman, 2007, 2009; Forbes-Riley et al., 2008; Litman & Forbes-Riley, 2004, 2006)

There are a number of ways in which artificial tutors (and other types of computerized learning environments) might adaptively respond to the learner's affective states in the course of enhancing learning (D'Mello, Craig, Sullins, & Graesser, 2006; D'Mello, Jackson et al., 2008; Dweck, 2002; Lepper & Woolverton, 2002). If the learner is frustrated, for example, the tutor can give hints to advance the learner in constructing knowledge or make supportive empathetic comments to enhance motivation. If the learner is bored, the tutor needs to present more engaging or challenging problems for the learner to work on. The tutor would probably want to lay low and stay out the learner's way when the learner is in a state of flow (Csikszentmihalyi, 1990), i.e., when the learner is so deeply engaged in learning the material that time and fatigue disappear. The flow experience is believed to occur when the learning rate is high and the learner has achieved a high level of mastery at the *region of proximal learning* (Metcalfe & Kornell, 2005).

In general, an affective interaction involves the immersion of a user into an affective loop or a *detect-select-synthesize cycle*. This involves the real-time *detection* of the user's affective states relevant to the domain, the *selection* of appropriate actions by the system to optimize task efficiency, and the *synthesis* of emotional expressions by the system so that the user remains engaged and the interaction is not compromised.

Although there are a number of obstacles that need to be overcome before functional affect-sensitive computer interfaces can be realized, the success of any affect-sensitive interface will ultimately depend upon the accuracy by which the user's affect can be detected. These interfaces are ultimately guided by the design goal of narrowing the communicative gap between the emotionally challenged computer and the emotionally rich human. 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). Therefore, robust recognition of the users' emotions is a crucial challenge that is hindering major progress toward the larger goal of developing affect-sensitive interfaces that work.

## **Defining the Affect Measurement Problem**

The problem of affect measurement (or detection or recognition or monitoring) is best framed in a machine learning paradigm. In its simplest form, the goal is to monitor a set of  $n$  emotions,  $E = [e_1, e_2, \dots, e_n]$ . The emotions will be monitored from data produced by sensor<sup>2</sup>  $S$ . Each instance of raw sensor data is denoted as  $[s_1, s_2, \dots]$ . An instance of this problem would involve the development of an algorithm  $A_k$  that selects an emotion

---

<sup>2</sup> This is the simplest case with a single sensor. A framework for the more complex case with multiple sensors is presented in Chapter 5.

$e_i$  when presented with sensor data  $s_j$ . An example would be a system that detects whether a user is happy or sad on the basis of facial expressions. In this example,  $E = [\text{happy}, \text{sad}]$ ,  $S$  is a camera, and  $s_j$  is a frame captured from the camera.

The process of developing algorithm  $A_k$  requires three fundamental steps. These include (a) feature identification, (b) feature selection, and (c) classifier selection.

*Feature identification* involves identifying and computing a set of  $m$  features  $F_j = [f_{j1}, f_{j2}, \dots, f_{jn}]$  from an instance of sensor data  $s_j$ . For example, if the sensor is a camera and  $s_j$  is a frame captured from the camera, then  $F_j$  could be a set of facial features such as the position of the eyebrows, eyelids, and mouth. Alternatively, if  $S$  is a microphone and  $s_j$  is a vocalization sample, then  $F_j$  could be a set of acoustic-prosodic features such as energy, pitch, loudness, etc. It is important to emphasize that there are no generalizable techniques for feature identification; that is, no domain independence can be established because the features identified are tightly coupled with the sensors in use.

Once a set of  $m$  features ( $F$ ) has been identified and computed from the sensory data, the next challenge involves selecting a subset  $G$  of these features ( $G \subseteq F$ ). The goal of this process, called *feature selection*, is to select the set of features that are most diagnostic of the emotions  $E$ . It is important to emphasize the importance of the feature selection phase because using the entire feature set ( $F$ ) is undesirable unless specified by prior theory. Generally speaking using the entire feature set ( $F$ ) can cause instability in the final system if potential problems such as overfitting (i.e., too many features in the model), multicollinearity (individual features that are correlated with each other), and

non-predictive features (i.e., a feature is not diagnostic of an emotion) are not adequately addressed.

The third phase, classifier selection, commences when feature identification and feature selection are complete. *Classifier selection* involves selecting an appropriate learning algorithm than can predict a particular emotion  $E_i$  on the basis of feature set  $G_j$  obtained from sensor data  $S_j$ . In a supervised learning environment a classifier is first trained on  $l$  samples of  $G_j$ . The emotion labels associated with each feature vector in  $G_j$  are also made available to the classifier. The classifier then attempts to build a predictive model based on statistical regularities between the feature vectors and emotion labels. The predictive model is then used to select the most likely emotion from  $E$  given an unseen feature vector. Different learning algorithms (i.e., different classifiers) yield different predictive models. Example learning mechanisms (i.e., classification schemes) include Bayesian classifiers, neural networks, lazy classifiers, decision tree classifiers, decision tables, and meta classification schemes (i.e., combining classifiers). Different classifiers yield different performance rates on different data sets and problem domains. Therefore, identifying the classifier that best discriminates among emotions  $E$  from feature vectors  $G$  provided by sensors  $S$  is an important question.

In summary, feature identification, feature selection, and classifier selection are the three critical phases required to develop automated affect detection systems. In addition to the common problems in supervised machine learning, classifier development in the affective domain has two additional complications as discussed below. This

dissertation will focus on developing and validating computational solutions to these challenges.

## **Challenges Facing Affect Measurement**

### *Individual Differences in Affect Experience and Expression*

Though early emotion research was rich with assumptions, expectations, and predictions of the universality of emotional experience, such claims have not withstood the careful eye of scientific scrutiny (Barrett, 2006). Although there is some evidence that people can recognize the emotions of others from different cultures on that basis of static facial displays (Elfenbein & Ambady, 2002a, 2002b) and vocal cues (Scherer, 2003), there is also evidence that people from the same culture share an in-group advantage (Russell, 1994). Furthermore, the studies demonstrating the universality of affective experience share several methodological issues related mostly to ecological validity concerns. Dominant among these problems are reliance on culled (carefully selected) stimuli, posed static displays of affect (by actors) rather than naturalistic dynamic experience from users, and the emotion recognition task being detached from the context of the experience (Barrett, 2006).

Differences in the age, gender, and personality of users are also a matter of some concern. For example, young children are generally more expressive than teens and adults. They are less likely to suppress and disguise their emotions due to societal pressures (Ekman, 2002; Ekman & Friesen, 1969). Therefore, the affect recognition

module, a critical component of a system that aspires to detect and respond to a user's affect, will have problems generalizing across different age groups. Furthermore, such a system may need to invoke different responses to appropriately manage the emotions of a user. For example, consider the situation of engagement, a crucial component of online learning environments, because a user is a mouse-click away from ending the interaction. Strategies to engage a bored male learner by launching educational but game-like simulation environments may not be as effective for a female learner.

Personality differences are also a critical factor in designing affect-sensitive interfaces. For example, within the context of online learning environments it is important to contrast adventuresome learners who want to be challenged with difficult tasks, take risks of failure, and manage negative emotions when they occur, with cautious learners who tackle easier tasks, take fewer risks, and minimize failure and its resulting negative emotions (Meyer & Turner, 2006).

In addition to the long list of individual differences surrounding emotional experiences, additional variations are observed across domains. For example, an immersive game-like learning environment will promote a different set of affective experiences than a classical computer assisted learning (CAL) system. It is reasonable to expect an enhanced state of engagement in a game environment than in a traditional CAL system. A sex education tutoring system is also likely to be more engaging for high school students than an algebra tutor. Differences can also be expected in situations when users interact with varying motivations. For example, a learning environment that prepares students for high stakes testing will invoke a different set of affective states than

a system deployed in classrooms for remedial students, even when the domain and underlying pedagogical framework of both systems are on par. In summary, the vast individual differences in affective expression and experience challenge the development of automated affect detection systems because they reduce the generalizability of such systems.

#### *Validity Concerns in Affect Measurement*

Akin to most psychological variables, affect is a construct (i.e., an inferred conceptual entity). Therefore, it cannot be directly measured and one can only approximate its true value. This approximation raises critical validity concerns in the measurement of human emotions. These include conclusion validity, internal validity, construct validity, and external (or ecological) validity (Rosenthal & Rosnow, 1984).

Conclusion validity pertains to the ability to infer a relationship (not necessarily causal) between any two variables of interest (e.g., are increased levels of happiness related to increased learning gains) (Shadish, Cook, & Campbell, 2002). *Internal validity* is concerned with establishing whether a relationship between two variables is causal (i.e., does happiness cause positive learning gains) (Shadish et al., 2002). *Construct validity* involves determining whether the operational definitions of a construct accurately reflect the construct (Campbell & Fiske, 1959; Fiske & Campbell, 1987). Simply put, are we measuring what we are claiming to be measuring? Finally, *external validity* (or ecological validity) is related to the extent that any observed relationship can generalize to other people, places, and times (Shadish et al., 2002). Although each of these validity

measures is pertinent to the scientific study of affect, this short discussion will focus on construct and external validity only. This is because the immediate concern is whether affect can be accurately measured (construct validity) and whether the measurements can be generalized (external validity). It would be impossible to develop a supervised affect classifier if affect labels for the training set could not be reliably obtained (construct validity). Similarly, an affect classifier would not be very functional if its predictions did not generalize above and beyond its training set (external validity).

Construct validity is an important concept in affect measurement because human emotions cannot be directly observed. So unlike developing a classifier in a domain where humans can reliably provide labels for the training set, the process is severely complicated in automated affect classification. Hence, one has to rely on operational measures such as self reports, observer judgments, or monitoring physiological and bodily correlates of affect expression. For example, consider the problem of measuring learners' happiness during a learning task. Since happiness cannot be directly measured (unlike physical measurements such as height, weight, etc.), it would have to be inferred from operational definitions such as self reports, observer judgments, or monitoring facial expressions. In this hypothetical situation, establishing construct validity pertains to the extent to which each of these operational definitions accurately reflects the desired construct (i.e., happiness). For example, simply monitoring the presence of smiles would be an unstable operational definition for happiness because people also smile when they are unhappy (e.g., a grimace), and there is a difference between true smiles (or Duchenne smiles) and forced smiles (Ekman, Friesen, & Davidson, 1990).

Although it is generally difficult to establish construct validity, (Campbell & Fiske, 1959) landmark paper provides useful guidelines. These guidelines include the demonstration of reliability, convergent validity, and discriminant validity.<sup>3</sup> Reliability implies that the same or similar measurement devices should produce measurements that are highly correlated. For example, affect judgments provided by two researchers observing a learner should be strongly correlated. Or an automated system that detects affect by monitoring facial features should demonstrate similar performance under varying conditions such as backgrounds, lighting, etc.

Convergent validity means that measurements produced by different measures that are theoretically related to a construct should be highly correlated. Therefore, to establish convergent validity in measuring affect, multiple measurement schemes should be employed and these should be strongly correlated. For example, subjective experiences of affect via self reports (Measure 1) should be correlated with facial expressions (Measure 2) (e.g., (Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005).

It is important to emphasize the difference between multiple measures of a construct (e.g., self reports + nonverbal behaviors) and multiple exemplars of the same measure (i.e., two researchers observing a student's emotions). For example, affect judgments made by two observers can highly correlate with each other (i.e., high reliability), yet this correlation is insufficient to establish any degree of convergent validity (i.e., multiple measures were not used). On the other hand, convergent validity

---

<sup>3</sup> Discriminant validity implies that measures that are not theoretically related should not be strongly correlated. Although not discussed here please see Campbell & Fiske (1959) for more details.

could be established if self reports of affect were correlated with judgments made by researchers.

Another important concern of research involving affect is the ecological validity of the studies. Ecological validity (or external validity) pertains to the ability to make inferences about behavior in the real world from controlled laboratory research. Barrett (2006) groups the threats to ecological validity into two main categories. The first issue is a sampling problem that plagues designs that adhere to *culled* sampling procedures. In these situations, participants judging affect are exposed to a small, carefully selected sample of available behavior. This selective sampling process results in overly optimistic recognition rates that do not reflect the realistic difficulty in decoding affect. For example, a meta-analysis by (Elfenbein & Ambady, 2002a, 2002b) reports that cross-cultural face perception accuracies were lower for studies that used the entire data set when compared to those using a culling sampling procedure.

According to Barrett (2006) the second important violation of ecological validity occurs when actors are used to portray affective expressions. Examples of these include caricatures of emotions as prominent in the facial expression literature, or actors posing various expressions (Banziger & Scherer, 2007; Scherer & Ellgring, 2007a). The use of actors is defended by virtue of the fact that they produce prototypical emotional expressions that resemble real emotional productions (Banse & Scherer, 1996). However, it is unclear whether actors portraying artificial affective expressions in controlled laboratory environments will ever yield affect classifiers that will generalize to naturalistic expressions in real-world environments [please see (Barrett, 2006) for an

extensive discussion on issues pertaining to threats to ecologically validity in affect measurement).

### **Limitations of Current Affect Detection Systems**

There appear to be three major limitations of contemporary affect detection systems. The major cause of these limitations appears to be a general reluctance to delve into the two challenges discussed above. The first two limitations refer to a set of concerns that apply to any affect detection system, independent of the domain in which it is deployed. These include problems with sensors and methodological concerns. The third set of problems applies to affect detection systems that aspire to be implemented in artificial learning environments.

It is important to highlight two salient aspects before delving into the discussion below. First, the list of problems addressed here is not comprehensive. Instead, it is restricted to issues that are directly addressed in this research. Second, and more importantly, it is important to emphasize that the goal of the discussion is to objectively list limitations of current affect-sensitive systems, knowing fully well that several of the limitations were only revealed once the systems were developed and tested.

### *Problems with Sensors*

*Intrusive Sensing Channels.* Many of the affect detection technologies analyze physiological signals for emotion detection (Picard, Vyzas, & Healey, 2001; Rani, Sarkar, & Smith, 2003; Whang et al., 2003). One potential pitfall to this approach is the reliance on obtrusive sensing technologies, such as skin conductance, heart rate monitoring, and measurement of brain activities. These obtrusive physiological sensors are acceptable in some applications, and it is true that users habituate to the presence of these sensors, but they are not satisfactory in environments where the sensors distract users and interfere with the primary tasks. This has motivated designers of affect-sensitive technologies to focus on facial feature tracking and acoustic-prosodic vocal features, two technologies that are unobtrusive (Paiva, Prada, & Picard, 2007; Pantic & Rothkrantz, 2003; Zeng, Pantic, Roisman, & Huang, 2009). However, as discussed below, it is unclear whether the face and voice capture the entire gamut of affective expression.

*Reliance on Facial and Acoustic-Prosodic Features.* Although the significance of non-verbal channels in affective expression is widely acknowledged (Andersen, 1999; Coulson, 2004; Darwin, 1872; Ekman, 1964, 1965a, 1965b; Ekman & Friesen, 1968, 1969; Mehrabian, 1968a, 1968b, 1971, 1972; Obudho, 1979), the impetus of non-verbal communication research has been almost entirely devoted to facial expressions and paralinguistic features of speech such as intonation and stress. A recent review by Pantic and Rothkrantz (2003) confirmed that the majority of affect detection systems rely on facial feature tracking or acoustic-prosodic feature monitoring. However, there are

several limitations to these channels. First, it is unclear whether all emotions are expressed through the face and speech. For example, there is some evidence that boredom and engagement, two affective states that are critical to affect-sensitive learning environments, cannot be reliably detected from the face (Craig, D'Mello, Witherspoon, & Graesser, 2008; Craig, D'Mello, Witherspoon, Sullins, & Graesser, 2004; McDaniel et al., 2007). Second, facial features and speech patterns can be controlled by a deceptive user. Emotional expressions are a highly socially reactive phenomenon, and it is quite conceivable that learners may attempt to disguise certain negative emotional expressions. 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.

*Reliance on Single Modality to Infer Affect.* Multimodal systems for affect detection and general user modeling have been widely discussed but rarely implemented (Jaimes & Sebe, 2007). This is mainly due to the inherent challenges with unisensory affect detection, which no doubt increase in multisensory environments. However, humans rarely express affect through a single modality (Feldman & Rime', 1991; Hinde, 1972; Knapp & Hall, 1997). Instead, the body demonstrates a remarkable degree of sophisticated coordination during emotional expressions (Scherer & Ellgring, 2007b).

Facial expressions are often accompanied by variations in prosody and inflections of posture. Therefore, systems that focus on a single modality are inherently handicapped in their ability to detect naturally occurring spontaneous affective displays. In contrast, multimodal affect detectors are expected to yield superior precision and recall.

### *Methodological Concerns*

*Threats to Construct Validity.* Although a handful of automated affect detection systems operate in an unsupervised fashion, supervised machine learning techniques are at the heart of most current affect detection systems. Providing accurate models of ground truth for a complex construct such as emotion is an important requirement for such supervised affect classifiers. However, as discussed above, establishing a sufficient degree of construct validity in affect measurement is quite challenging. Almost all research efforts toward affect detection have bypassed the problem entirely by relying on a single operational measure when inferring a learner's emotion. They have used self reports (De Vicente & Pain, 2002; Klein et al., 2002; Matsubara & Nagamachi, 1996) or ratings by independent judges (Liscombe, Riccardi, & Hakkani-Tür, 2005; Litman & Forbes-Riley, 2004; Mota & Picard, 2003). But a combination of measures to establish convergent validity for ground-truth affect categories has rarely been implemented. The significance of this problem cannot be overstated because any affect classifier trained on a data set with unreliable affect labels will produce spurious results.

*Threats to Generalizability and Ecological Validity.* It is unfortunate, but one does not have to do an extensive literature search to find examples of affect detection

systems that are rife with threats to ecological validity. In fact, ecological validity violations are the rule rather than the exception. Pantic and Rothkrantz (2003) review a list of 18 systems that attempt to recognize affective states from static facial images or sequences of facial images. Although affect recognition rates range from 40% to 97%, concerns pertaining to the size (#samples) and variation (#participants) of the data sets used to train and test the affect detectors raise some critical generalizability and ecological validity concerns. For example, the training sets of 61% of the studies on affect detection from static facial images included 10 or fewer participants. Similarly, an astonishing 78% of studies on affect detection from sequences of facial images had 10 or fewer participants. In another review of 14 studies on affect detection from acoustic-prosodic features, fewer than 30% of the studies utilized utterances from 10 or more participants. The lack of a sample size to ensure generalizability only scratches at the surface of ecological validity concerns. Other pressing problems include the use of actors for productions of affective expressions, context-free expressions, and invariance in sex, age, and racial profile of participants.

*Context-Free Affective Data Sets.* Another limitation of affect detection systems is that they are trained on data sets where the affective expressions are collected in a context-free environment. But the nature of affective expressions is not context free. Instead, it is highly situational dependent (Bachorowski & Owren, 1995; Barrett, 2006; Keltner & Haidt, 2001; Panksepp, 1998; Russell, 2003). Since affective expressions can convey different meanings in different contexts, training affect detection systems in context-free environments is unlikely to produce useful results.

### *Concerns Pertaining to Learning Environments*

*Focus on “Basic” Emotions.* An affect-sensitive learning environment needs to have a list of affective states to monitor and respond to. Identifying a relatively short but comprehensive set of affective states to monitor is quite challenging due to the lack of basic research on the affective activities during complex learning. There have been theories that link cognition and affect very generally, such as those of Mandler (1984), Bower (1981), Stein and Levine (1991), Ortony, Clore, and Collins (1988), Russell (2003), and several others (Barrett, 2006; Bower, 1981, 1992; Ekman, 1984; Lazarus, 1991; Mandler, 1976, 1984a, 1984b; Ortony, Clore, & Collins, 1988; Russell, 2003; Scherer, Schorr, & Johnstone, 2001; Smith & Ellsworth, 1985; Stein, Hernandez, & Trabasso, 2008; Stein & Levine, 1991). Although these theories convey general links between cognition and emotions, they do not directly explain or predict the sort of emotions that occur during complex learning, such as attempts to master physics, biology, or computer literacy. Some emotions presumably have a more salient role in learning than others (Linnenbrink & Pintrich, 2002). What are these emotions? How are they linked to cognition? These are the fundamental questions that surface in modeling affect in learning environments.

Ekman and Friesen (1992) have proposed six *basic* emotions that are ubiquitous in everyday experience. The six basic emotions include fear, anger, happiness, sadness, disgust, and surprise. However, many have called into question the relevance of these basic emotions to the learning process (Baker, D'Mello, Rodrigo, & Graesser, in review; Kort et al., 2001). But almost all contemporary affect detection systems focus on the

basic emotions (Paiva et al., 2007; Pantic & Rothkrantz, 2003; Zeng et al., 2009) and are not very relevant to learning environments (D'Mello & Graesser, 2006; D'Mello, Lehman, & Person, in review; Lehman, D'Mello, & Person, 2008; Lehman, Matthews, D'Mello, & Person, 2008).

*Focus on a Small Number of Affective States.* Some of the affect detection systems are limited in their applicability to learning environments because they focus on a small number of affective states such as neutral, negative, and positive (Litman & Forbes-Riley, 2004), negative versus positive/non-negative (Lee & Narayanan, 2005; Liscombe et al., 2005), or annoyance versus frustration (Ang, Dhillon, Krupski, Shriberg, & Stolcke, 2002). These contrasts may be suitable for some domains, but they are not sufficient to encompass the realistic gamut of learning (Conati, 2002). Additional complexities arise from the fact that a person's reaction to the presented material can change as a function of their goals, preferences, expectations, and knowledge state. Consequently, this dissertation advocates for a larger set of affective states within the arena of complex-learning. These include boredom, engagement/flow, confusion, frustration, delight, and surprise.

### **An Interdisciplinary Framework for Automated Affect Measurement**

Functional solutions to the problems raised above require an interdisciplinary framework that spans computer science, psychology, and education. This framework will be considered in this research. In general, three solutions to the problems raised above are

proposed. First, the problems documented with unimodal facial or vocal sensory channels are alleviated by a combination of two novel sensors that include conversational cues and gross body language. Methodological concerns are reduced by using multiple judges to establish a degree of construct validity in establishing ground truth for the affect categories needed for supervised machine learning. Finally, concerns pertaining to affect detection in learning environments are addressed by monitoring a larger number of learning-centered affective states (i.e., boredom, flow/engagement, confusion, frustration, delight, surprise, and neutral).

*Solutions to Sensor Problems: Tracking Conversational Cues and Gross Body Language*

As an alternative to intrusive sensors and the reliance on facial and vocal features this dissertation explores the possibility of affect detection from two relatively unexplored sensors: conversational cues (or dialogue features) and gross body language (or posture patterns). There are good reasons for expecting that dialogue features are diagnostic of affect detection in learning environments. The dialogue in one-on-one tutoring sessions yields a rich trace of contextual information, characteristics of the learner, episodes during the coverage of the topic, and social dynamics between the tutor and learner. Within the context of tutoring systems, a prediction is that dialogue features will provide a very robust diagnostic channel to infer a learner's affect because the conversational cues have a broad and deep feature set that covers deep meaning, world knowledge, and pragmatic aspects of communication.

The use of posture for affect detection is motivated by a number of embodied theories of cognition (Clark, 1997; de Vega, 2002; deVega, Glenberg, & Graesser, 2008; Glenberg, Havas, Becker, & Rinck, in press). Theories of embodied cognition postulate that cognitive processes are constrained substantially by the environment and by the coupling of perception and action. If embodied theories are correct, the cognitive and emotional states of a learner are implicitly or intentionally manifested through their gross body language. Therefore, the monitoring of posture patterns may lead to insights about the corresponding cognitive states and affective arousal. An added advantage of monitoring posture patterns is that they are ordinarily unconscious and thereby not susceptible to social editing, at least compared with speech acts in conversation.

The fact that two sensors of affect will be monitored raises the issue of how the features of the two channels will be combined. 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. An alternative hypothesis would be that there is redundancy between the channels. When there is redundancy, the addition of one channel to another channel yields negligible incremental gains; the features of the two channels are manifestations of very similar mechanisms.

### *Solutions to Methodological Problems: Multiple Measures of Ground Truth*

A number of researchers have relied on a single operational measure when inferring a learner's emotion, such as self reports (De Vicente & Pain, 2002; Klein et al., 2002; Matsubara & Nagamachi, 1996) or ratings by independent judges (Liscombe et al., 2005; Litman & Forbes-Riley, 2004; Mota & Picard, 2003). In contrast, the combination of several different measures of a learner's affect is proposed. The emotion measures included in this research incorporate judgments made by the learner, a peer, and two trained judges, as will be elaborated later. The use of multiple measures to infer a learner's affective state represents a significant difference from previous research and a positive step toward obtaining the requisite degree of construct validity in measuring affect.

In addition to using multiple measures for affect detection, an important goal of this dissertation is to ensure that training and testing data are collected in an ecologically valid setting. First, no actors were used to express emotions and no emotions were intentionally induced. Second, the problem with context-free affective expressions was eliminated by designing a data collection procedure where all affective expressions were recorded in context. Finally, the data set consisted of approximately 4000 samples of naturally occurring affective expressions from 28 students (Graesser et al., 2006). This large data set is sufficient to establish the necessary degree of generalizability for the affect detectors.

*Solutions to Problems Concerning Affect Measurement in Learning Environments*

Compared to previous research on affect detection, a larger number of affective states ( $N = 7$ ) were monitored, and the analyses attempted to automatically distinguish between a subset ( $N = 6$ ) of these. This significantly raises the complexity of the affect detection systems because a six-way classification problem is inherently more complicated than a two- or three-way problem.

There are also differences in the affective states that are addressed in this research. In contrast to the basic emotions that Ekman intensely investigated (e.g., sadness, happiness, anger, fear, disgust, surprise), researchers have identified a different distribution of emotions that prevail during complex learning. These include boredom (Miserandino, 1996), confusion (Graesser & Olde, 2003; Kort et al., 2001), frustration (Kort et al., 2001; Patrick, Skinner, & Connell, 1993), delight (Fredrickson & Branigan, 2005; Silvia & Abele, 2002), and flow (Csikszentmihalyi, 1990). It should be noted that some researchers may view some of these emotions (i.e., affect states) as cognitive states, whereas other researchers would classify them as either emotions or affect states (Barrett, 2006; Meyer & Turner, 2006; Stein & Hernandez, 2007). The position adopted in this dissertation agrees with the latter group because these states are accompanied by enhanced physiological arousal (compared with neutral) and affect-cognition amalgamations are particularly relevant to complex learning. This research represents one of the first attempts to automatically measure these learning-centered affective states in an ecologically-valid fashion.

## **Research Overview**

### *Research Goals*

The research described here is organized around the following goals that are incremental steps toward the larger goal of developing functional affect detection systems. Although Goals 2, 3, and 4 contribute to the computational merits of this research, Goal 1 is a necessary prerequisite. As discussed earlier, feature identification, feature selection, and classifier selection are the three critical phases required to develop automated affect detection systems. The accuracy of the affect detector is directly related to the design decisions permeating in each of these phases. Therefore, the major goals and contributions of this research are related to each of these phases.

*Goal 1: Empirical Data Collection.* This goal pertains to the collection of an ecologically valid data set that integrates sensory data with human-provided affect categories. This data set will be used to train and validate the affect classifiers developed in this research. It will also be used to establish a baseline on the accuracy by which human judges detect emotions. Performance of automated affect detection systems (Goals 2, 3, and 4) will be compared to this human baseline.

*Goal 2. Automated Affect Detection from Conversational Cues.* This goal involves the development of a real-time automated system that classifies affective states based on conversational cues that are generated during natural language tutorial dialogues. A key subgoal is to develop a feature selection algorithm that selects the most predictive,

generalizable set of features in situations where there is ambiguity in the ground-truth affective labels.

*Goal 3. Automated Affect Detection from Gross Body Language.* This goal involves the development of a real-time automated system to classify affective states by monitoring the gross body language of a person while performing a learning task. An important subgoal is to extract novel features from a stream of bodily motions. Affect classifiers that monitor these features are expected to be serious contenders to more common methods such as facial feature tracking and monitoring speech contours.

*Goal 4. Automated Affect Detection from a Combination of Conversational Cues, Gross Body Language, and Facial Features.* This goal pertains to the development of a real-time multimodal system that classifies affective states by combining conversational cues, gross body language, and facial features. An important subgoal is to develop and evaluate different strategies to combine the two sensors and determine if a combination of the two sensors resonates with superadditivity or redundancy as defined above.

#### *Accomplishments and Pending Work*

*Goal 1: Empirical Data Collection.* Goal 1 was addressed by conducting a study where 28 learners completed a tutorial session with AutoTutor, an intelligent tutoring system that helps students learn by holding a conversation in natural language (Graesser & D'Mello, in preparation; Graesser et al., 2006). Videos of the participant's face and computer screen were recorded during the tutorial sessions. Posture patterns and conversational cues were also automatically recorded for offline analyses. After

completing the tutorial session the affective states of the learner were judged by the learner, a peer, and two trained judges.

The data collected in this study was used in two ways. First, it was used as training and validation data for the automated affect detectors. This was accomplished by constructing seven data sets that temporally integrated the affective judgments with the conversational cues and posture patterns of each learner. The first four datasets corresponded to the judgments of the learner, a peer, and two trained judges, while the remaining three data sets combined judgments of two or three judges. A data set that combined judgments of all four judges was limited in size and not considered in the analyses.

Second, the data collected in this study was used to establish a baseline for affect detection accuracy. Inter-rater reliability scores were computed for the different pairs of judges in order to determine the accuracy by which humans measure affect. This information was used to gauge the accuracy of the automated affect detection systems.

*Goal 2. Automated Affect Detection from Conversational Cues.* Affect detectors that automatically classified learners' affective states in real time by monitoring lexical and semantic features from a tutorial dialogue between AutoTutor and a learner were developed (D'Mello, Craig, A. et al., 2005; D'Mello et al., 2006; D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008; D'Mello & Graesser, 2006; Graesser et al., 2008). The use of lexical and semantic features constitutes the feature selection phase. Feature selection was performed by a novel feature selection method that operates in situations where there is ambiguity in the ground-truth categories of a classification task.

The method was used to select features that were (a) most diagnostic of the affective states, and (b) expected to generalize above and beyond individual differences.

Machine learning experiments for classifier selection indicated that the classifiers were moderately successful in discriminating the affective states of boredom, confusion, flow, frustration, and neutral, yielding a peak accuracy of 42% with neutral (chance = 20%) and 54% without neutral (chance = 25%). Individual detections of boredom, confusion, flow, and frustration, when contrasted with neutral affect, had maximum accuracies of 69%, 68%, 71%, and 78%, respectively (chance = 50%). The classifiers that were trained on the emotion judgments of the trained judges and combined models outperformed those based on judgments of the novices (i.e., the self and peer). Follow-up classification analyses that assessed the degree to which machine-generated affect labels correlated with affect judgments provided by humans revealed that human-machine agreement was on par with novice judges (self and peer) but quantitatively lower than trained judges.

*Goal 3. Automated Affect Detection from Gross Body Language.* This dissertation explores the reliability of detecting learners' affect by monitoring their gross body language (body position and arousal) (D'Mello & Graesser, 2009; D'Mello, Picard et al., 2007; D'Mello, Chipman, & Graesser, 2007). An automated body pressure measurement system was used to capture the pressure exerted by the learner on the seat and back of a chair during the tutoring session. Two mechanisms to monitor affective states from the pressure maps were developed. The first method focused on monitoring the average pressure exerted, along with the magnitude and direction of changes in the pressure

during emotional experiences. The second method monitored the spatial and temporal properties of naturally occurring pockets of pressure.

Machine learning experiments for classifier selection yielded affect detection accuracies of 73%, 72%, 70%, 83%, and 74%, respectively (chance = 50%), in detecting boredom, confusion, delight, flow, and frustration, from neutral. Accuracies involving discriminations between two, three, four, and five affective states (excluding neutral) were 71%, 55%, 46%, and 40% with chance rates being 50%, 33%, 25%, and 20%, respectively.

*Goal 4. Automated Affect Detection from a Combination of Channels.* This goal was realized by developing and evaluating two methods for multimodal sensory fusion. These include feature-level fusion which involves grouping features from the various sensors before classification, and decision-level fusion where fusion occurs after classification. An additive algorithm in which a multi-channel feature vector was created by appending features from each individual channel was considered for feature-level fusion. A spreading activation network with projecting and lateral links was used to model decision-level fusion (Rumelhart, McClelland, & Group, 1986).

Classification results that compared the individual channels supported a channel × judgment type interaction, where the face was the most diagnostic channel for voluntary affect judgments, and conversational cues were superior for the mandatory judgments. The analyses also indicated that the accuracy of the multi-channel (face, dialogue, and posture) model was statistically higher (albeit marginally) than the best single-channel model for the mandatory but not voluntary judgments. However, multi-channel models

reduced the discrepancy (i.e., variance in the precision of the different emotions) of the single-channel models for both mandatory and voluntary judgments. The results also indicated that the combination of channels yielded superadditive effects for some emotions, but additive, redundant, and inhibitory effects for others. Finally, the analyses indicated that five weighting schemes for decision-level fusion yielded accuracy scores that were on par with each other and with feature-level fusion.

### *Research Novelty*

There are a number of factors that contribute to the novelty of this research. This section provides a brief discussion of the novel aspects of this dissertation within the context of each of the research goals listed above. Extended discussions of the novelty of this research that are contextually situated in literature reviews appear in Chapters 3, 4, and 5.

*Goal 1: Empirical Data Collection.* The data set used to train and validate the classifiers represents one of the largest naturalistic data sets that is currently used in automated affect detection. In addition to a large number of affect instances (approximately 3000), there are four affect judges (self judgments, peer judgments, and two trained judges) and three sensory channels (conversational cues, gross body language, and facial features). The richness and validity of the data set can be appreciated by virtue of the fact that it has been used in over 20 publications (see <http://emotion.autotutor.org>).

*Goal 2. Automated Affect Detection from Conversational Cues.* The research addressed in this subgoal is novel for three reasons. First, it is one of the pioneering

attempts to detect affect from a large set of conversational channels. Second, the algorithms use a novel multi-level regression-based feature selection algorithm. The third source of novelty emerges from the fact that the classifiers incorporate multiple measures for ground-truth affect categories.

*Goal 3. Automated Affect Detection from Gross Body Language.* There was only one system (Mota & Picard, 2003) that attempted to automatically detect affective states from gross body language when this research was being developed. Their algorithm detected three affect states by monitoring posture patterns. This research significantly extends these ideas by developing two algorithms that monitor a larger number ( $N = 6$ ) of states.

*Goal 4. Automated Affect Detection from a Combination of Conversational Cues, Gross Body Language, and Facial Features.* Multimodal systems for affect detection and general user modeling have been widely discussed but rarely implemented (Jaimes & Sebe, 2007). At the time of writing there are only a handful of systems that attempt to detect affect by combining sensors. Most of these systems focus on fusing audio-visual features alone. Furthermore, these systems primarily rely on fusing facial features with acoustic-prosodic features of speech. The current approach that involves a combination of conversational cues, gross body language, and facial features has never been attempted.

#### *Research Contributions*

This is an interdisciplinary research project that makes contributions to the fields of computer science, artificial intelligence, psychology, and education. This research

contributes to computer science and artificial intelligence by developing new algorithms for affect classification from novel sensors. Although the application of these algorithms has been related to affect detection in learning environments, they can be applied to similar problems in related domains. This research enhances the field of HCI by the development of robust automatic affect detectors that can be used in any affect-sensitive interface. The research contributes to the field of psychology by exploring relationships between conversational cues, gross body language, and affective experience. Contributions to education include an exploration of the complex interplay between affect and learning as well as the development of affect-sensitive learning environments.

## **Dissertation Organization**

This dissertation is organized into four chapters that address each of the goals listed above. Chapter 2 provides the background information for the automatic affect detection systems by (a) describing the learning environment used in this research, (b) describing the sensors used for affect detection, (c) describing the protocol used to collect training and validating data, (d) providing data on the proportions of the emotions observed, and (e) establishing a baseline accuracy for the automated affect detection systems by investigating the reliability by which humans classify affect. The content of Chapter 2 has been adapted from Graesser and D'Mello (in preparation) and Graesser, McDaniel, Chipman, Witherspoon, D'Mello, and Gholson (2006).

Chapter 3 describes and empirically validates the affect classifier that monitors conversational cues by: (a) providing a literature review on previous attempts to detect affect from discourse features, (b) describing the new feature selection method, (c) listing details on the conversational features that were monitored (feature identification), (d) providing results of statistical analyses that were performed to select the most generalizable set of features (feature selection), and (e) discussing the results of an extensive set of classification analyses that explored overall performance trends, isolated the best classification models, and compared the best classification models to humans (classifier selection). The content of Chapter 3 has been adapted from D'Mello et al. (2008), Graesser, D'Mello et al. (2008), D'Mello and Graesser (2006), D'Mello, Craig, Sullins, and Graesser, 2006, and D'Mello et al. (2005).

Chapter 4 describes and empirically validates the affect classifier that monitors gross body language by: (a) providing a literature review on previous attempts to detect affect from gross body language, (b) describing two algorithms that monitor affect by tracking body language, (c) specifying a hierarchical classification scheme motivated by the Pandemonium model (Selfridge, 1959), and (d) empirically validating the algorithms by an extensive set of classification analyses that explored overall performance trends and isolated the best classification models. The content of this chapter has been adapted from D'Mello and Graesser (in press), D'Mello, Picard, and Graesser (2007), and D'Mello, Chipman, and Graesser (2007).

Chapter 5 describes and empirically validates the multimodal affect classifier that monitors conversational cues, gross body language, and facial features by: (a) discussing

earlier systems that aspire to be multimodal affect detectors, (b) developing mechanisms for feature-level and decision-level sensory fusion, (c) deriving metrics to ascertain if a combination of channels yields superadditive, additive, redundant, or inhibitory effects, (d) comparing the accuracy of multi-channel affect detection to single-channel affect detection and feature-level fusion to decision-level fusion, and (e) examining the structure of the multimodal affect classifiers.

Finally, Chapter 6 synthesizes the major accomplishments, highlights limitations, proposes solutions to alleviate the limitations, and outlines plans for future work. It will also situate this research within the broader context of affect-sensitive interfaces. This chapter includes discussions on: (a) the accuracy of the automated affect detection systems, (b) the possibility of boosting affect classification accuracy by modeling the temporal dynamics of the emotions, (c) threats to scalability and potential solutions, (d) the impact of individual differences on affect detection, and (e) a case study that describes an application of the affect detectors developed in this dissertation.

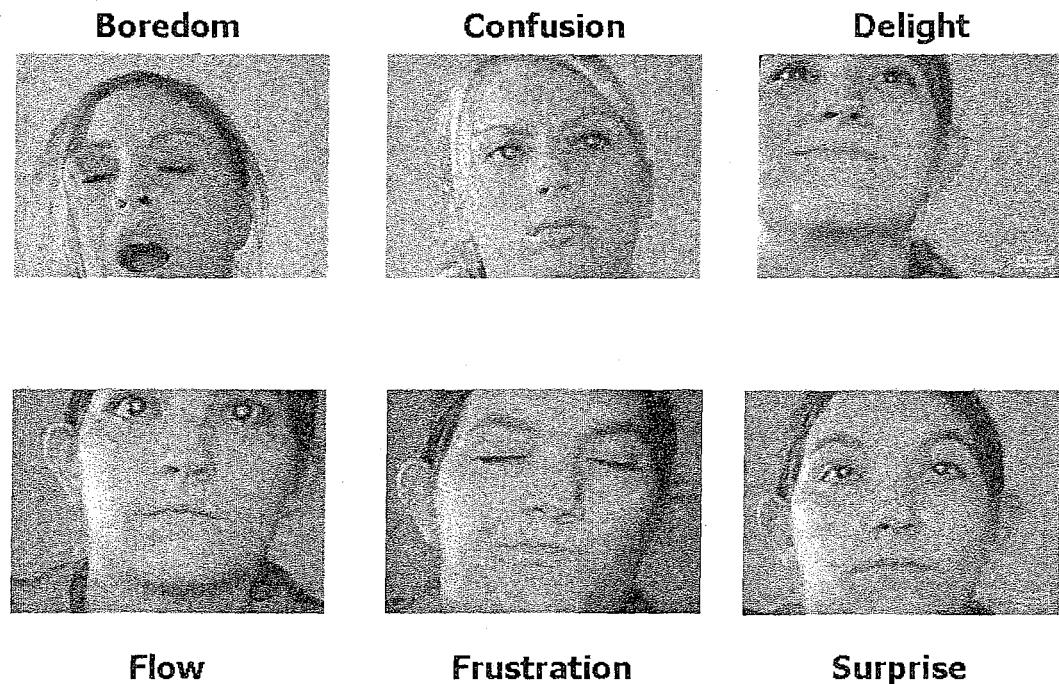
## **Chapter 2: Affect Measurement by Humans**

### **Introduction**

Supervised affect classification systems, such as the ones developed in this research (see Chapters 3, 4, and 5), require data sets that integrate sensory data with ground truth categories of affect. The quality of the data set used to train and validate the classifiers has a major impact on the fidelity of the classifiers. Chapter 1 discussed several limitations of previously used data sets that render the derived classifiers spurious and lacking in generalizability. Potential solutions to some of the problems that reduce the validity and generalizability of these data sets were also discussed. These solutions were implemented in the current study that was designed to collect sensory data as well as affect labels that could be used to train supervised affect classifiers.

In addition to being a source for training and validation data, the current study was also designed to answer a fundamental question pertaining to a baseline for affect detection accuracies. One preliminary step in answering the fundamental question of how affective states are classified is to investigate a simple measurement question: How reliably can emotions be classified by the learners themselves versus peers versus trained judges (experts)? This question was investigated by comparing the extent to which trained judges and untrained peers can accurately identify the affective states of learners.

This chapter begins with a brief description of AutoTutor which was the learning environment used in the study. A detailed description of the methodology and results follows. The focus is on a model of learning-centered emotions obtained from previous research with AutoTutor and other learning environments (Baker et al., *in review*; Craig, Graesser et al., 2004; D'Mello et al., 2006; Lehman, D'Mello et al., 2008; Lehman, Matthews et al., 2008). These learning-centered emotions included boredom, flow/engagement, confusion, frustration, delight, surprise, and neutral. An example of the facial displays accompanying these emotions is presented in Figure 1.



**Figure 1.** Sample facial displays of learning-centered emotions.

## The AutoTutor Learning Environment

AutoTutor is an intelligent tutoring system that helps students learn Newtonian physics, computer literacy, and critical thinking topics through tutorial dialogue in natural language (Chipman, Olney, & Graesser, 2006; Graesser, Chipman, Haynes, & Olney, 2005; Graesser et al., 2003; Graesser, Lu et al., 2004a). The impact of AutoTutor in facilitating the learning of deep conceptual knowledge has been validated in over a dozen experiments on college students for topics in introductory computer literacy (Graesser, Lu et al., 2004b), conceptual physics (VanLehn et al., 2007), and critical thinking (Storey, Kopp, Wiemer, Chipman, & Graesser, in press). Tests of AutoTutor have produced learning gains of .4 to 1.5 sigma (a mean of 0.8), depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor. From the standpoint of the present study, this dissertation assumes that AutoTutor helps learning whereas the direct focus is on the emotions that occur during the tutorial session.

AutoTutor's dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers. For example, the following is an example of a challenging question in computer literacy: When you turn on the computer, how is the operating system first activated and loaded into RAM? These questions require the learner to construct approximately 3–7 sentences in an ideal answer and to exhibit reasoning in natural language. However, when students are asked these challenging questions, their initial answers are typically only 1 or 2 sentences in length. However, 1–2 sentences provide insufficient information to adequately answer the

question so a tutorial dialogue is needed to flesh out a complete answer. AutoTutor engages the student in a mixed-initiative dialogue<sup>1</sup> that draws out more of what the student knows and that assists the student in the construction of an improved answer.

AutoTutor's interface had the five major windows shown in Figure 2. Window 1 (top of screen) is the main question that stays on the computer screen throughout the conversation for the question. Window 2 (left middle-top) is the animated conversational agent that speaks the content of AutoTutor's turns. Window 3 (right middle-top) is either blank or has auxiliary diagrams. Window 4 (middle-bottom) displays the dialogue history of the tutoring session. Finally, Window 5 (bottom) displays the student's answers as they are typed in.

---

<sup>1</sup> Mixed-initiative interaction allows for the direction and control of the interaction to be shifted between participants (Allen, 1999).

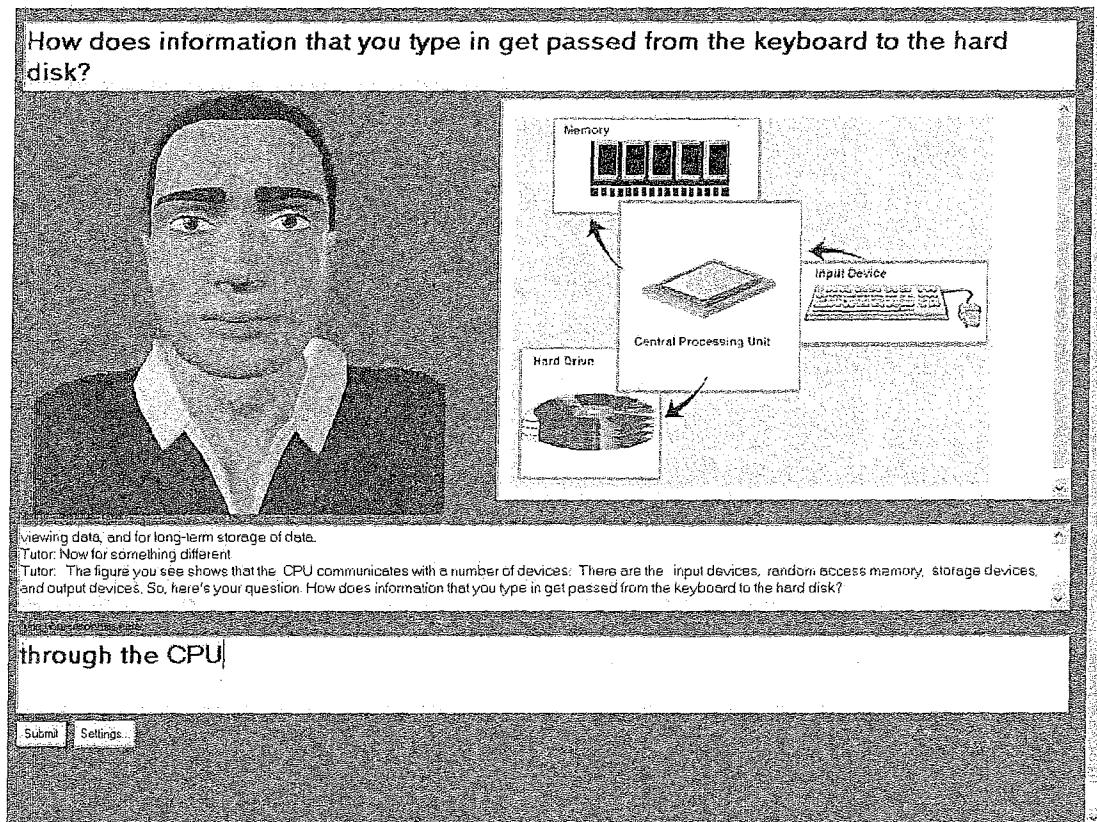


Figure 2. Screenshot of the AutoTutor interface.

Table 1 illustrates an excerpt conversation with AutoTutor that was extracted from an actual tutoring session. This session was with a relatively verbose, knowledgeable student about computer literacy so the conversation is comparatively short.

**Table 1.** Sample conversation between AutoTutor and a student

Turn	Speaker	Text	Dialogue Move
1	AutoTutor	When you turn on the computer, how is the operating system first activated and loaded into RAM?	Main Question
2	Student	When the computer is turned on a file is automatically booted up	Contribution
3	AutoTutor	Anything else?	Pump for information
4	Student	This file contains the operating system	Contribution
5	AutoTutor	What about storage?	Hint
6	Student	The operating system is then read into RAM and activated.	Contribution
7	AutoTutor	Right. The CPU executes instructions in ROM, which helps the operating system to start.	After giving positive feedback, there is an assertion

#### *The Structure of AutoTutor Dialogue*

AutoTutor's dialogue was designed to simulate human tutoring (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001; Graesser, Person, & Magliano, 1995; Shah, Evens,

Michael, & Rovick, 2002). The nature of the dialogue has been described in detail in previous publications (Graesser, Chipman et al., 2005; Graesser, Lu et al., 2004b). Expectations and misconceptions form the underlying conceptual structure of AutoTutor's dialogue. They are the primary pedagogical method of scaffolding good student answers. Both AutoTutor (Graesser, Chipman et al., 2005) and human tutors (Graesser et al., 1995) typically have a list of expectations (anticipated good answers) and a list of anticipated misconceptions associated with each main question. For example, the following four expectations are associated with the question "Why do computers need operating systems? (E1)

The operating system helps load application programs, (E2) The operating system coordinates communications between the software and the peripherals, (E3) The operating system allows communication between the user and the hardware, and (E4) The operating system helps the computer hardware run efficiently.

AutoTutor guides the student in articulating each of the expectations of a problem (or main question) through a five-step dialogue frame that is prevalent in human tutoring (Graesser & Person, 1994; Graesser et al., 1995; VanLehn et al., 2007). The five steps of the dialogue frame are:

1. Tutor asks main question
2. Student gives initial answer
3. Tutor gives short feedback on the quality of the student's answer in #2,
4. Tutor and student collaboratively interact via expectation and misconception tailored dialogue, and

5. Tutor verifies that the student understands (e.g., Do you understand?)

This dialogue frame is implemented over a number of conversational turns. Each turn of AutoTutor in the conversational dialogue has three information slots (i.e., units, constituents). The first slot of most turns is short feedback on the quality of the student's last turn. This feedback is either positive (e.g., "very good", "bravo"), negative (e.g., "not quite", "almost"), or neutral (e.g., "uh huh", "okay").

The second slot advances the coverage of the ideal answer with either prompts for specific words, hints, assertions with correct information, corrections of misconceptions, or answers to student questions. Hints and prompts are carefully selected by AutoTutor to produce content in the answers that fill in missing words, phrases, and propositions. For example, a hint to get the student to articulate expectation E1 might be "What can you say about application programs?"; this hint would ideally elicit the answer "The operating system loads these...." A prompt is a question to get the student to express a specific content word. For example, the prompt question "What does the operating system do with application programs?" would hopefully elicit "it loads them."

The third slot is a cue to the student for the floor to shift from AutoTutor as the speaker to the student. For example, AutoTutor ends each turn with a question or a gesture (rendered by the animated conversational agent) to cue the learner to do the talking. Discourse markers (and also, okay, well) connect the utterances of these three slots of information within a turn. These markers play an important role in the conversational smoothness of AutoTutor's dialogue (Person & Graesser, 2002; Wiemer-Hastings, Graesser, & Harter, 1998)

As the learner expresses information over many turns, the list of expectations is eventually covered and the main question is scored as answered. Complete coverage of the answer requires AutoTutor to have a pool of hints and prompts available to extract all of the content words, phrases, and propositions in each expectation. AutoTutor adaptively selects those hints and prompts that fill missing constituents and thereby achieves pattern completion.

AutoTutor is dynamically adaptive to the learner in other ways than coaching them to articulate expectations. There is the conversational goal of correcting misconceptions that arise in the student's responses. When the student articulates a misconception, AutoTutor acknowledges the error and corrects it. AutoTutor accommodates a mixed-initiative dialogue by attempting to answer the student's questions. The answers to the questions are retrieved from glossaries or from paragraphs in textbooks via intelligent information retrieval. AutoTutor asks a counter-clarification question (e.g., "I don't understand your questions, so could you ask it in another way?") when it does not understand the student's question.

The conversational smoothness and the pedagogical quality of AutoTutor's dialogue moves has been evaluated in a *Bystander Turing test* (Person & Graesser, 2002). A Bystander Turing Test presents a human judge with a transcript containing a dialogue (such as the middle column in Table 1). The task of the judge is to determine whether the last utterance in the transcript was generated by a human or by a computer. Bystanders were unable to discriminate dialogue moves of AutoTutor and dialogue moves of a real human tutor.

### *Interpreting Learners' Contributions*

The three levels of AutoTutor dialogue go a long way in simulating a novice human tutor. AutoTutor can keep the dialogue on track because it is always comparing what the student says to 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, McCarthy, McNamara, & Graesser, 2008) and semantic matching algorithms (Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2008). A two-tiered interpretation scheme is used to understand the learners' utterance. First, a speech act classifier determines whether the incoming utterance best fits a contribution (i.e., directly related to the tutoring content, e.g., "switch to virtual memory") or a frozen expression that signifies a particular discourse function. Frozen expressions are a short response (e.g., "yes", "no"), a metacognitive statement (e.g., "I need help", "I don't know"), a metacommunicative statement (e.g., "please repeat", "could you say that in another way"), and a question (e.g., "what is RAM?"). If the response is classified as a contribution, semantic matching algorithms attempt to compute the conceptual quality of the response. Details on the actual mechanisms that AutoTutor uses to interpret the learners contributions are presented in the subsequent sections. However, at this stage it is important to note that AutoTutor's ability to assess the conceptual quality of learners' answers has been validated in a series of experiments. AutoTutor's ability to interpret the learners' responses correlates moderately with domain knowledge experts, which is an

impressive feat for an artificial tutor with imperfect natural language processing (NLP) mechanisms (Graesser, Penumatsa et al., 2007).

## Methods

### *Participants*

The participants were 28 undergraduate students enrolled in an introductory psychology course at a southern university. The participants received course credit for their participation. Prior coursework in computer literacy was not required. Data from participants who required corrective eyeglasses were eliminated from the study due to restrictions in the facial feature monitoring system used in this experiment.

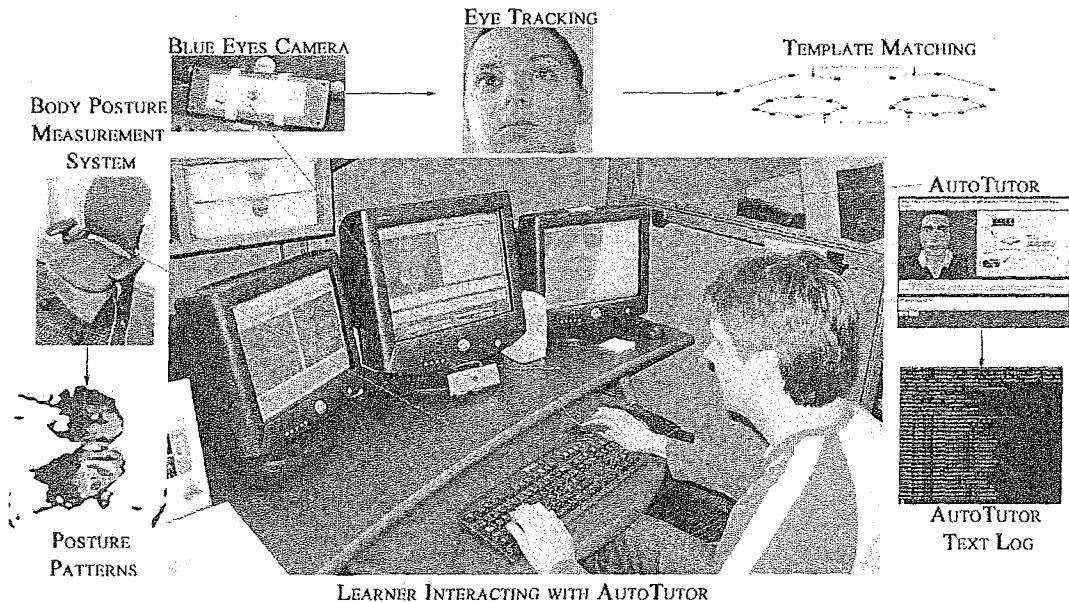
### *Content Covered in the Tutorial Session with AutoTutor*

Participants interacted with AutoTutor version 2.0 for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, Internet, or operating systems (see Graesser, Person, & Harter, for details about AutoTutor content). Each of these topics had 12 questions that required about a paragraph of information (3–7 sentences) in an ideal answer. The questions varied in difficulty and in the presence or absence of an associated diagram. The questions required answers that involved inferences and deep reasoning, such as *why*, *how*, *what-if*, *what if not*, and *how is X similar to Y?*, as illustrated in example questions in the AutoTutor section.

### *Materials and Equipment*

Two computers were used for presentation and data capture. One of the computers handled the tutorial interaction whereas the other computer was used to record a video of the participant's face. The system with AutoTutor had a 2.4 GHz Pentium® IV processor with a 19" Trinitron® monitor and hyper threading capabilities. This computer was also used for the majority of data capture.

Four channels of information were recorded and collected while participants interacted with AutoTutor on topics of computer literacy. First, AutoTutor's log file recorded the entire verbal history of the dialogue and states of the AutoTutor–learner interaction. The interaction state space included states of language, discourse, learning, and coverage of the ideal answer (expectations) at the completion of each turn composed by the student. This information was used to develop the automated affect detector described in Chapter 3. Second, an IBM® Blue Eyes Camera (Morimoto, Koons, Amir, & Flickner, 1998) was used for collecting data on facial movements at a fine-grained resolution. Third, body posture was tracked by the Body Pressure Measurement System (BPMS, version 5.23) developed by Tekscan® Inc (Tekscan., 1997). Fourth, Camtasia Studio™, a screen capturing software package, recorded a video of the participant's entire tutoring session with AutoTutor. The captured video also included an audio stream of the speech generated by the AutoTutor animated conversational agent. Please see Figure 3 for an illustration of the sensing channels while a learner interacts with AutoTutor. Each of these channels is described further below.



**Figure 3. Sensors used during data collection.**

*AutoTutor's Log File.* AutoTutor is a client-server distributed application that maintains a log file of information regarding the interactive session with the student. At each student turn AutoTutor writes its assessment of the student's response along with some internal information to stable storage on hard disks, with suitable backup. Assessment of the student's responses includes information such as: the correctness of an answer (i.e., degree of match to specific expectations and misconceptions), the verbosity of the student contributions, reaction and response times, the length of a specific turn, and a host of other parameters about the conceptual quality of the student's turns.

AutoTutor's Speech Act Classification System (Olney et al., 2003) classifies each contribution within a turn of a student's response into five broad dialog categories: meta-communicative (e.g., What did you say?), metacognitive (I don't follow), shallow

comprehension (okay, that's right), student contributions that reflect deep comprehension, and "other" contributions. Each contribution within a turn was identified by the occurrence of a punctuation mark (? , ! , or .) to segment multiple contributions within a turn; a turn without punctuation was considered a single contribution. After assessing and recording a student's response at each turn, AutoTutor provides short feedback on the contribution (positive, negative, neutral) and makes one or more substantive dialog moves (hint, prompt, assertion, etc.) that advance the conversation. AutoTutor's feedback, next dialogue move, and text responses are also recorded in AutoTutor's log file. The log file is described in more detail in Chapter 3.

*IBM Blue Eyes Camera.* A digital video recording of the faces of the participants was captured with the IBM Blue Eyes camera. This is an infrared camera, designed by IBM and built at the MIT Media Lab that was purposely designed for facial feature analysis (Morimoto et al., 1998). This camera uses alternating infrared pulses and the red-eye effect to create a special hash pattern on the pupil image. This pattern increases the accuracy of the algorithms used for facial feature analysis by providing an "anchor" of known size and shape onto which the algorithms can lock (Kapoor & Picard, 2005). The pupils were used as the anchor in this study. The size of the recorded video was 640 × 480 pixels (width × height) with a frame rate of 30 fps (frames per second).

*Body Pressure Measurement System (BPMS).* The BPMS system, developed by Tekscan™ (1997), consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The pad is paper thin with a rectangular grid of sensing elements that is enclosed in a protective pouch. Each sensing element provides an 8-bit pressure output

in mmHg. The current setup had one sensing pad placed on the seat of a Steelcase™ Leap Chair and another placed on the back of the chair. The BPMS system is discussed in more detail in Chapter 4.

*Camtasia Software.* A screen capturing software package called Camtasia was used to capture the audio and video of the participant's entire tutoring session with AutoTutor. The video consisted of 30 frames per second images ( $1024 \times 768$  pixels) of the AutoTutor interface (see Figure 2). The captured audio included the text to speech messages generated by the AutoTutor agent.

*Knowledge Tests.* Multiple-choice tests were used to assess the student's prior knowledge of computer literacy and also whatever learning gains occurred from the tutorial session. It should be pointed out that the amount of training time and the number of AutoTutor questions covered in this study was much less than previous tutoring sessions with AutoTutor that systematically assessed learning gains (Graesser, Lu et al., 2004b). The goals of the present study were to analyze emotions during learning from AutoTutor rather than assessing learning gains. Given the short amount of training time and the small number of questions covered, impressive learning gains were not expected and therefore did not perform systematic analyses on relationships between learning and emotions.

### *Protocols*

*Retrospective Judgments of Emotion (Affect).* Judgments of emotions (JOEs) were recorded by the participants and by two trained judges. The JOEs were made at 20-

second intervals where the participant or trained judge (described below) viewed a video of the learner's face (captured via the IBM BlueEyes camera) and the tutorial sessions (captured via Camtasia) and filled out a series of tables. Therefore the judge was provided with the facial display of the participant grounded in the context of the interaction. A customized software application to synchronize both videos and display them to the judge with playback controls (start, stop, rewind, and pause) was developed for the judgment process. At the end of each 20-second interval, the video was automatically paused and a table was provided with the emotions (affect states) of interest listed at the top and with a box under each state to be marked if the state was observed just before the video stopped. These points at 20-second intervals were designated as *mandatory* JOEs. The table was designed to allow participants to mark more than one affective state and to indicate whether a reported affective state persisted throughout the 20 second interval. In the instances where more than one affective state was reported, the participants were required to identify a primary mandatory JOE. Additionally, if a state was observed within the 20-second interval, but not at the end of the interval, then it could be identified by recording the exact time in which the observed affective state occurred (start time and end time). These judgments were designated as *voluntary* JOEs. More than one voluntary JOE's could be identified by the participant or trained judge for each 20-second period.

The affective states of interest for this study were boredom, confusion, flow, frustration, delight, surprise, and neutral. *Boredom* was defined as "being weary or restless through lack of interest." *Confusion* was defined as "a noticeable lack of

understanding.” *Flow* was defined as a “state of interest that results from involvement in an activity.” *Frustration* was defined as “dissatisfaction or annoyance.” *Delight* was defined as “a high degree of satisfaction.” *Surprise* was defined as “wonder or amazement, especially from the unexpected.” *Neutral* was defined as “no apparent emotion or feeling.”

#### *Procedure*

Each participant completed the experiment in two sessions that lasted 90–120 minutes each. In session 1, the participants were then given instructions that they would be learning about topics in computer literacy with AutoTutor. The participants were randomly assigned to one of three conditions, corresponding to the computer literacy topic (hardware, software, the Internet) in which they were to be tutored by AutoTutor. They were administered a pretest with the 4AFC multiple choice test to assess their prior knowledge of computer literacy. Participants then interacted with AutoTutor on the assigned topic for 32 minutes. During this time, the BPMS and the Blue Eyes camera recorded a video of the participants’ posture and face respectively. The computer’s display was recorded as well using the Camtasia screen capture software. As soon as the computerized session was completed, participants were administered the posttest on computer literacy.

After interacting with AutoTutor, the video streams from the AutoTutor screen and the participant’s face were synchronized and displayed to the participants. At the end of each 20-second interval, the two video streams were paused (freeze framed) and the

participant was asked to make mandatory JOEs they experienced at that instant. The alternative emotions were boredom, confusion, flow, frustration, delight, surprise, or neutral. As explained in the Retrospective Judgments of Emotion subsection, they were permitted to mark more than one affective state, but would need to designate which one was the primary mandatory JOE. Participants also designated any voluntary JOEs that they had experienced during the 20 seconds in between the previous pause and the current pause. This procedure of participants giving JOEs on their own tutorial sessions constituted the *self* JOEs.

Students participated in a second session of the experiment within a week after the completion of the first session. In the second session, participants provided *peer* JOEs by repeating the same JOE procedure, but this time on another participant's session. Given that AutoTutor tutored participants on three topics related to computer literacy (hardware, operating systems, and the internet), the peer JOE session was always from on the same topic as the participant's topic. For example, a student who was tutored on Operating Systems provided emotion judgments for another participants who was also tutored on Operating Systems.

Two trained judges also provided JOEs on each of the 28 participants' video recordings, using the same procedure and materials. The judges were trained on Ekman's Facial Action Coding System<sup>2</sup> (FACS) (Ekman & Friesen, 1978), and also had

---

<sup>2</sup> FACS specifies how judges are to code specific facial behaviors (i.e., *action unit* or AU), based on the muscles that produce them.

considerable experience interacting with AutoTutor. Hence, their emotion judgments were based on contextual dialogue information as well as the FACS system.

Therefore, each tutorial session had JOEs collected from the self, a peer, and two trained judges. This permitted us to collect inter-judge reliability measures for six pairs of individuals making JOEs: self-peer, self-judge1, self-judge2, peer-judge1, peer-judge2, and judge1-judge2. The distribution of learners' emotions as reported by each judge was also investigated.

## **Results and Discussion**

The results and discussion are segregated into two parts that address two major questions. First, what emotions occur while students learn about computer literacy with AutoTutor? Second, how much agreement in the judgments of emotion (JOEs) exists from the viewpoints of the student (self), a peer, and trained judges? These research questions directly contribute towards the goal of developing automated affect classification systems. The first question is important because it allows us develop automated systems to detect the more prominent affective states. The second question allows us to establish a human baseline for affect detection accuracies. Automated affect detection systems can be compared to this human baseline. An analysis of proportions of JOEs in the various

emotion categories will address the first question, whereas kappa<sup>3</sup> scores (Cohen, 1960) to assess inter-judge reliability will address the second question.

### *What Emotions Occur During Complex Learning?*

The JOEs at both mandatory points at ends of 20-second intervals, as well as at voluntary points in between the mandatory points were analysed. Each of the 28 participants had 96 mandatory JOE observations, so 2688 primary (first-choice) mandatory judgments were expected. However, due to missing judgements and other transcription errors the number of judgements obtained was 2537, 2615, 2678, and 2590 for the self, peer, trained judge 1, and trained judge 2, respectively. A proportion score was computed for each participant that indicated the proportion of approximately 96 primary mandatory judgments that were in each of the seven emotion categories (boredom, confusion, flow, frustration, delight, surprise, and neutral). Participants gave voluntary judgments on in-between points rather frequently, approximately 37% of the observations. Therefore, the proportion scores for the voluntary JOEs were also computed. A separate set of proportion scores for mandatory and voluntary JOEs were collected for the four judges: self, peer, judge1, and judge2. This resulted in a set of 56 proportion scores when considering 2 judgement types, 7 emotions, and 4 judges.

---

<sup>3</sup> The Kappa statistic measures the proportion of agreement between two raters with correction for chance (J. Cohen, 1960).  $kappa = \frac{P_o - P_e}{1 - P_e}$ , where  $P_o$  = proportion of agreement observed and  $P_e$  = proportion of agreement expected by chance.

Table 2 presents means and standard deviations for the 56 proportion scores.

When averaging over the cell means of JOE types and judges, the mean scores showed the following proportions in descending order: Confusion (.29), Neutral (.19), Boredom (.13), Frustration (.12), Flow (.11), Delight (.09), and Surprise (.05). Confusion is also known to be a very prominent affective state in previous studies of AutoTutor when emotions were judged by observers (Craig, Graesser et al., 2004) and by learners who verbally expressed their emotions via an emote aloud protocol (D'Mello et al., 2006). Confusion is a signal that the learner is challenged and is in thought, which results in significant learning gains (Craig, Graesser et al., 2004; Graesser, Chipman, King, McDaniel, & D'Mello, 2007).

Nevertheless, the distribution of these proportions was hardly uniform across judges and types of JOE. A repeated measures ANOVA was performed on the proportion scores in Table 2, with three factors (type, emotion and judge). The proportion scores are constrained to add to 1.0 within type and judge, so it is not meaningful to consider the main effects of type and judge, and the type  $\times$  judge interaction. However, the main effect of emotion and the remaining interactions are not constrained and therefore justifiable. The main effect of emotion was significant,  $F(6, 162) = 38.00, MSe = .038, p < .001$ , partial  $\eta^2 = .585$ , as was also the interactions between emotion and judgment type,  $F(6, 162) = 68.94, MSe = .031, p < .001$ , partial  $\eta^2 = .719$ , emotion  $\times$  judge interaction,  $F(18, 486) = 5.18, MSe = .024, p < .001$ , partial  $\eta^2 = .161$ , and the three-way judgment type  $\times$  emotion  $\times$  judge interaction,  $F(18, 486) = 3.40, MSe = .024, p < .001$ , partial  $\eta^2 = .112$ . Quite clearly, most of the variance is explained by the main effect of differences in

emotions and by the emotion  $\times$  type interaction. Therefore, follow up analyses of simple main effects between emotions within mandatory observations versus voluntary observations were performed.

Consider first the mandatory observations. Bonferroni posthoc tests revealed the following ordering among means: Neutral (.37) > Confusion (.21) = Flow (.19) = Boredom (.17) > Delight (.01) = Surprise (.01). Frustration (.04) fit in as greater than delight and surprise, the same as flow and boredom, and less than neutral and confusion. Therefore, it was confusion and the more subtle emotions (neutral, flow, and boredom) that were more prominent at mandatory points. Such points are the most unbiased representative sample of the emotions that occur during learning. Both confusion and emotions with subtle facial responses (boredom, neutral, and flow) occur 94% of the time. The obvious implication of this result is that the emotions manifested by learners are most often confusion or staid, quite devoid of the expressively rich emotions of delight, surprise, and frustration.

It is, of course, possible to dissect the relative proportion scores among emotions when segregating particular judges and JOE types, given that there was a significant three-way interaction. A difference of .14 reflects a significant difference between means for those readers who wish to dissect the proportion scores in more detail. However, the above trends account for most of the variance.

**Table 2. Descriptive statistics for proportions of emotions observed.**

Type	Judge	Affective State													
		Boredom		Confusion		Delight		Flow		Frustration		Neutral		Surprise	
M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Man	Self	.16	.15	.15	.14	.02	.04	.23	.20	.10	.11	.33	.29	.02	.03
	Peer	.20	.17	.16	.12	.01	.01	.22	.24	.04	.04	.36	.28	.01	.02
	Judge1	.10	.09	.27	.14	.01	.02	.19	.15	.03	.03	.39	.10	.00	.01
	Judge2	.21	.11	.26	.10	.01	.01	.11	.11	.01	.01	.41	.08	.00	.00
Vol	Self	.18	.22	.30	.27	.05	.09	.05	.09	.22	.27	.03	.07	.10	.21
	Peer	.13	.21	.34	.30	.07	.11	.03	.08	.19	.21	.03	.07	.09	.14
	Judge1	.05	.11	.36	.22	.33	.23	.02	.07	.15	.22	.01	.03	.08	.20
	Judge2	.03	.06	.48	.22	.22	.22	.01	.02	.20	.20	.00	.00	.06	.13
Ex. Vol	Judge1	.11	.09	.36	.17	.14	.13	.05	.08	.20	.16	.14	.11	.00	.00
	Judge2	.23	.13	.36	.14	.10	.10	.06	.05	.09	.07	.16	.13	.00	.00

*Notes.* Man – Mandatory judgments, Vol = Voluntary judgments, Ex. Vol = Exhaustive voluntary judgments

A follow up analysis on proportion scores by having the two trained judges re-analyze all voluntary observations that were identified as such by one or more judges (self, peer, judge1, and judge2) was performed. 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). This pattern seemed to be an amalgamation of the patterns for mandatory and voluntary observations.

In summary, confusion was the most prevalent affective state during learning with AutoTutor. The comparative subtle states of flow and boredom were also quite representative of the learning experience whereas delight and surprise were rare and frustration occasionally occurred. The flamboyant emotions, such as confusion, delight, frustration, and surprise, were more prevalent at the voluntary observations, presumably because these emotions are manifested in obvious facial expressions or dialogue patterns.

#### *Inter-judge Agreement on JOEs*

This section reports inter-judge agreement scores for JOEs using Cohen's kappa as a measure of inter-judge reliability. There were four judges (self, peer, trained judge1, and trained judge2) so there were six possible pairs of judges (see Table 3). The reliability scores were based on the first-choice affect states. The observations segregate the mandatory and voluntary JOEs, corresponding to the data reported in the previous section and Table 2. Cohen's kappa scores were computed separately for each of the 28 learners. Statistical analyses were performed on these kappa scores when comparing agreement of the six pairs of judges.

Table 3 presents means and standard deviations of kappa scores for the 84 cells that contrasts judgment type, emotions, and judges. The scores in Table 3 revealed that the trained judges had the highest agreement, the self-peer pair had near zero agreement,

and the other pairs of judges were in between. When averaging over the cell means of emotions and judgment type, the mean kappa scores were .09, .21, .19, .24, .26, and .51, for self-peer, self-judge1, self-judge2, peer-judge1, peer-judge2, and judge1-judge2, respectively. These results support the conclusion that peers are not particularly good at detecting learner emotions. Another conclusion is that training on Ekman's facial action coding system and tutorial dialogue can enhance the reliability and accuracy of judgments of affective states. However, these means are not perfectly accurate because the number of observations was quite low for some of the emotions and there were 0 observations for some cells in the voluntary type. Therefore, separate inferential statistics on the mandatory and voluntary observations will be reported.

**Table 3. Descriptive statistics for kappas between raters for judging affect.**

Type	Judge	Affective States															
		Overall		Bor		Con		Del		Flo		Fru		Neu		Sur	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD		
Man	SF-PR	.06	.07	.07	.13	.10	.15	.08	.18	.03	.09	.03	.08	.03	.09	-.01	.01
	SF-J1	.11	.10	.11	.14	.16	.14	.12	.21	.08	.12	.06	.13	.08	.12	-.01	.01
	SF-J2	.13	.13	.13	.16	.19	.21	.07	.16	.11	.17	.01	.05	.09	.16	.00	.00
	PR-J1	.11	.10	.18	.15	.14	.13	.10	.23	.08	.12	.07	.15	.07	.09	.02	.08
	PR-J2	.15	.09	.26	.16	.16	.15	.11	.23	.09	.12	.08	.18	.10	.12	.08	.29
	J1-J2	.31	.08	.25	.17	.36	.14	.16	.25	.30	.19	.10	.23	.27	.10	.00	.00
Vol	SF-PR	.12	.27	.09	.24	.40	.41	.44	.50	-	-	.08	.25	-	-	-.01	.33
	SF-J1	.31	.36	.51	.41	.61	.42	.27	.56	-	-	.28	.34	-	-	.08	.20
	SF-J2	.24	.30	.24	.41	.57	.39	.32	.53	-	-	.09	.29	-	-	.13	.32
	PR-J1	.36	.37	-	-	.51	.43	.41	.39	-	-	.39	.48	-	-	.00	.08
	PR-J2	.37	.33	-	-	.60	.34	.36	.38	-	-	.29	.42	-	-	.22	.38
	J1-J2	.71	.35	.33	.58	.76	.35	.79	.31	-	-	.52	.43	-	-	.60	.46
Ex. Vol	J1-J2	.49	.22	.44	.32	.59	.21	.58	.30	.31	.37	.37	.32	.32	.31	.26	.42

*Notes.* SF = Self, PR = Peer, J1 = Trained Judge 1, J2 = Trained Judge 2. Man = Mandatory judgments, Vol = Voluntary judgments, Ex. Vol = Exhaustive voluntary judgments. Bor = Boredom, Con = Confusion, Del = Delight, Flo = Flow, Fru = Frustration, Neu = Neutral, Sur = Surprise.

According to Table 2, the mandatory observations showed extremely low proportion scores (i.e., .02 or lower) for the emotions of delight and surprise. Therefore, kappa scores were analyzed only for the emotions of confusion, boredom, frustration, flow, and neutral. A 6 (judge) by 5 (emotion) repeated measures ANOVA<sup>4</sup> was conducted on the kappa scores for these mandatory JOEs. There were statistically significant effects for judge,  $F(5, 135) = 25.01, MSe = .027, p < .001$ , partial  $\eta^2 = .481$ , emotion,  $F(4, 108) = 11.84, MSe = .035, p < .001$ , partial  $\eta^2 = .305$ , and the judge  $\times$  emotion interaction,  $F(20, 540) = 4.04, MSe = .012, p < .001$ , Partial  $\eta^2 = .130$ . The variance explained by the interaction term was significant, but accounted for a modest percentage of the variance. The pattern generally showed that judge1-judge2 had the highest inter-judge reliability scores, the self-peer pair the lowest, with other pairs in between. The mean scores were .05, .10, .10, .11, .14, and .26 for self-peer, self-judge1, self-judge2, peer-judge1, peer-judge2, and judge1-judge2, respectively. Bonferroni posthoc tests indicated the judge1-judge2 kappa scores were significantly higher than the kappa scores for the other five pairs of judges, the self-peer agreement was significantly lower than peer-judge1, and peer-judge2, and the middle group of four pairs were not significantly different from each other. The main effect kappa scores for the different emotions showed the following pattern: boredom (.17), confusion (.18), flow (.11), frustration (.06), and neutral (.11). The statistical tests on the main effect means showed the following trends, although there were a couple of violations in transitivity of

---

<sup>4</sup> Please see Appendix A for detailed results on the kappa scores.

statistical comparisons at the  $p < .05$  level: frustration < neutral = flow < boredom = confusion. Simple main effects within judges did not consistently show this pattern, however.

According to Table 2, the voluntary observations showed a sufficiently large number of observations for confusion, delight, and frustration, but not the other categories of emotions. Therefore, kappa score scores were analyzed only for these three emotions. A 6 (judge) by 3 (emotion) repeated measures ANOVA was conducted on the kappa scores for these mandatory JOEs. There were statistically significant effects for judge,  $F(5, 135) = 20.01$ ,  $MSe = .092$ ,  $p < .001$ , partial  $\eta^2 = .426$ , emotion,  $F(2, 54) = 48.82$ ,  $MSe = .093$ ,  $p < .001$ , partial  $\eta^2 = .644$ , and the judge  $\times$  emotion interaction,  $F(10, 270) = 9.30$ ,  $MSe = .042$ ,  $p < .001$ , partial  $\eta^2 = .256$ . Once again, the main effect means showed the judge1-judge2 kappas to be the highest, the self-peer kappas to be the lowest, and the other four pairs of judges to be in between: Self-Peer (.25), Self-Judge1 (.39), Self-Judge2 (.33), Peer-Judge1 (.44), Peer-Judge2 (.42), and Judge1-Judge2 (.69). Bonferroni posttests on the main effect for emotions showed the following significant differences: Confusion (.58) > delight (.43) > frustration (.25). Simple main effects within judges and within emotions are consistent with these trends, although not consistently. It is quite apparent that the kappa scores are much higher for these voluntary judgments than the mandatory judgments, presumably because they are accompanied by salient facial expressions or dialogue patterns. The high scores for the judge1-judge2 pairs are compatible with the conclusion that training on facial action coding schemes and tutorial dialogue helps improve inter-judge reliability in JOEs. Finally, the

comparatively low JOEs for the self-peer pair suggests that peers are not necessarily good judges of emotions.

## **General Discussion**

The results of this study support a number of conclusions about emotion detection. First, trained judges who are experienced in coding facial actions and tutorial dialogue provide affective judgments that are more reliable and that match the learner's self reports better than the judgments of untrained peers. Second, the judgments by peers have very little correspondence to the self reports of learners. Peers apparently are not good judges of the emotions of learners. Third, an emotion labeling task is more difficult if judges are asked to make emotion judgments at regularly polled timestamps, rather than being able to stop a video display to make spontaneous judgments. The states at regular timestamps are much less salient so there is minimal information for judges to base their judgments, compared with those points when affective states are voluntarily spotted. 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.

The overall low kappa scores between the various judges highlight the difficulty in measuring a complex construct (conceptual quantity) such as emotion. Furthermore, the kappas for the 2 trained judges are on par with data reported by other researchers who have assessed the reliability of emotion detection by humans (Ang et al., 2002; Grimm,

Mower, Narayanan, & Kroschel, 2006; Litman & Forbes-Riley, 2004; Shafran, Riley, & Mohri, 2003). However, statisticians have sometimes claimed that kappa scores ranging from 0.4–0.6 are typically considered to be fair, 0.6–0.75 are good, and scores greater than 0.75 are excellent (Robson, 1993). On the basis of this categorization, the kappa scores obtained from the study would range from poor to fair. However, such claims of statisticians address the reliability of multiple judges or sensors when the researcher is asserting that the decisions are clear-cut and decidable. But the present goal is very different. Instead, the goal is to use the kappa score as an unbiased metric of the reliability of making affect decisions, knowing full well that such judgments are fuzzy, vague, ill-defined, and possibly indeterminate. Therefore, the subsequent analyses consider all four judges as criteria for determining learners' affect. This allows us to examine patterns that generalize above and beyond individual differences among the affect judges. Moreover, there are conditions when the kappa scores are nearly excellent (i.e., trained judges on the more salient emotions associated with voluntary judgments), so it is not a hopeless goal to identify and track learner emotions in naturalistic sessions.

It is unclear what exactly should be the gold standard for deciding what emotions a learner is truly having. Should it be the learner or the expert? There is some uncertainty in the answer to this question, but it is conceivable that some emotions may best be classified by learners and others by experts. Perhaps a composite score that considers both viewpoints would be most defensible.

But there is another possibility as well. It could be the case that instrumentation (i.e., automated affect classifiers) might provide the most reliable affect classifications for

some of the emotions. This hypothesis is the major motivation of this thesis and will be explored in the next three chapters.

## **Chapter 3: Automatic Affect Detection from Conversational Cues**

### **Introduction**

An emotionally sensitive learning environment, whether human or computer, requires some degree of accuracy in classifying the learner's affective states. The emotion classifier need not be perfect but must have some modicum of accuracy. This chapter focuses on detecting affect from discourse features obtained from a natural language mixed-initiative dialogue interaction. Although dialogue has traditionally been a relatively unexplored channel for affect detection, it is a reasonable information source to explore because dialogue information is abundant in virtually all conversations and is inexpensive to collect.

A growing body of research has investigated emotions in human–human dialogues (Alm & Sproat, 2005; Forbes-Riley & Litman, 2004) and human–computer dialogues (Litman & Forbes-Riley, 2004; Porayska-Pomsta, Mavrikis, & Pain, 2008), although the literature on automated affect detection from the latter is somewhat sparse. Dialogue (i.e., discourse) features have typically been used in conjunction with acoustic-prosodic and lexical features obtained through an interaction with spoken dialogue systems. The lexical features usually are restricted to either human or automatic

transcriptions (with speech recognition engines) of user utterances. The acoustic-prosodic features are composed of vocal cues that typically include speech rate, intonation, and volume. A number of research groups have reported that appending an acoustic-prosodic and lexical feature vector with dialogue features results in a 1–4% improvement in classification accuracy (Ang et al., 2002; Lee & Narayanan, 2005; Liscombe et al., 2005; Litman & Forbes-Riley, 2004). A classic example involving this use of dialogue features is work investigating dialogue and emotions conducted on the program ITSPOKE (Litman & Silliman, 2004). ITSPOKE integrates a spoken language component into the Why2-Atlas tutoring system (VanLehn et al., 2002). With ITSPOKE, Litman and Forbes-Riley (2004) analyzed spoken student dialogue turns on the basis of lexical and acoustic features, with codings of negative, neutral or positive affect. They were able to reach high levels of accuracy in detecting the affect categories.

In a similar vein, Ang et al. (2002) reported that the inclusion of discourse features, such as the current turn within a session and the associated dialogue acts of the current turn, resulted in a 4% improvement in performance over lexical and prosodic features. Their research involved detecting annoyance and frustration within the context of a travel reservation system. Similarly, Liscombe, Riccardi, and Hakkani-Tür (2005) reported that the use of dialogue features caused a 1.2% improvement over the use of acoustic-prosodic and lexical features in discriminating between positive and negative emotions. Their results were obtained by analyzing a large database of 5,690 spoken utterances obtained from user interactions with the *How May I Help You* spoken dialogue system (Gorin, Riccardi, & Wright, 1996). An additional 2.8% improvement in accuracy

was obtained by the inclusion of contextual features spanning two previous turns.

Another important example of the use of dialogue for affect detection is provided by Lee and Narayanan (2004). They reported that the use of dialogue features of user utterances obtained from a call center produced a 3% increase in accuracy over prosodic and lexical features in discriminating between negative and non-negative emotions.

More recently, Porayska-Pomsta, Mavrikis, and Pain (2008) used features from tutor-student dialogues collected from mathematics tutorial sessions to predict levels of student confidence, interest, and effort. Other innovative uses of dialogue have emerged from research on the identification of problematic points in human-computer interactions (Batliner, Fischer, Huber, Spilker, & Noth, 2003; Carberry, Schroeder, & Lambert, 2002; Walker, Langkilde-Geary, Hastie, & Gorin, 2002). For example, Carberry, Lambert, and Schroeder (2002) proposed an algorithm to recognize doubt by examining linguistic and contextual features of dialogue in conjunction with world knowledge. Batliner et al. (2003) reported that discourse information resulted in a 1.2% improvement in classification accuracy over lexical and prosodic features alone.

Three major differences between the approach to affect detection adopted in this dissertation and some of the earlier research involving the use of dialogue to detect affect can be identified. The first difference is that this dissertation explores a larger array of discourse variables. The assumption is that dialogue can be a serious competitor to more popular measures of user affect, such as facial and acoustic-prosodic features. The second difference is that previous efforts investigating dialogue were limited to a small set of affective states, such as neutral, negative, and positive (Litman & Forbes-Riley, 2004),

negative versus positive/non-negative (Lee & Narayanan, 2005; Liscombe et al., 2005), or annoyance versus frustration (Ang et al., 2002). These contrasts may be suitable for some domains, but they are not sufficient to encompass a realistic gamut of learning (Conati, 2002; Conati & Maclare, in press). Consequently, this research involves the detection of a larger set of affective states within the arena of complex-learning. The relevant emotions (i.e., affective states) include boredom, confusion, delight, flow, frustration, neutral, and surprise. The third difference between this research and other efforts is the method of establishing ground-truth categories of affect. A number of researchers have relied on a single operational measure when inferring a learner's emotion, such as self reports (De Vicente & Pain, 2002; Klein et al., 2002; Matsubara & Nagamachi, 1996) or ratings by independent judges (Liscombe et al., 2005; Litman & Forbes-Riley, 2004; Mota & Picard, 2003). In contrast, this dissertation proposes the combination of several different measures of a learner's affect. The measures of emotion incorporate judgments made by the learner, a peer, and two trained judges, as elaborated in Chapter 2.

The chapter begins with a description of some of the computational challenges followed by a feature selection algorithm to address these challenges. The set of AutoTutor dialogue features used as predictors of the learners' affective states is described next. The Results section begins with some preliminary predictions regarding the perceived difficulty in automatically detecting affect from dialogue. A series of statistical analyses evaluate the hypothesis that dialogue features can significantly predict the learner's affect. Two feature selection techniques are subsequently investigated as

preprocessing techniques for the machine learning algorithms. The machine learning experiments attempt to assess the reliability of automatically detecting the learner's affect from AutoTutor's dialogue. The conclusion discusses the methods and their associated strengths, limitations, and ideas for future improvements.

## **Computational Challenges and Solutions**

As discussed in the introductory chapter, feature identification, feature selection, and classifier selection are the three major phases towards building an effective affect detection system. Since this chapter focuses on affect detection from features of AutoTutor's dialogue, the feature identification process is trivial. The focus is on the features ( $F$ ) that are stored in AutoTutor's mixed-initiative dialogue (described in the next section). The classifier selection phase, though being nontrivial, is quite straightforward. A series of standard classifiers are compared and the one that consistently provides the best performance is selected. However, the feature selection phase is quite complex. In particular one first needs to select a subset  $G$  of the features ( $G \subseteq F$ ) that are the most diagnostic of the learners' affective states. The classifier will then attempt to discriminate among the emotions  $E$  on the basis of  $G$ . The feature selection phase can be quite straightforward in situations in which the ground truth categories are well defined. However, the situation is severely complicated in the affective domain because there is a degree of ambiguity in the ground truth affect labels. In fact, as the interrater reliability scores in Chapter 2 illustrated, the ambiguity is quite

large. Humans were not very effective in detecting the emotions and multiple measures of the learners' emotions were obtained on the basis of the different judges perspectives. Therefore, a significant challenge entails the development of appropriate feature selection mechanisms that can provide the most representative set of features despite a large degree of ambiguity on the phenomenon being modeled.

Here a mechanism by which the ambiguity in the affective categories can be resolved in the feature selection phase is proposed. Let us begin by considering a simplification of the  $o$ -way classification problem by considering a binary (two-way) classification problem. This instance of the problem involves detecting the presence of emotion  $E$  using  $m$  features  $F = [f_1, f_2, \dots, f_m]$ . The proposed approach operates in three phases. First each affect judge is considered independently and a preliminary predictive model is constructed. This model determines the degree to which the complete feature set  $(F)$  predicts  $E$ . Since the ground truth label for  $E$  is provided by affect judges  $J = [J_1, J_2, \dots, J_n]$ , then  $E_j$  denotes emotion  $E$  measured by one particular affect judge. Therefore, a model that predicts  $E_j$  from a set of  $m$  features,  $F = \{f_1, f_2, \dots, f_m\}$ , is first constructed.

The exact nature of the model can vary. For example, a linear regression model, or a logistic regression model, or a linear discriminant model can be considered. What is important, however, is the coefficients of the model  $\beta_j = [\beta_{1j}, \beta_{2j}, \dots, \beta_{mj}]$ . For example, if a linear regression model is constructed then the  $\beta_j$  vector represents the standardized or unstandardized coefficients. In this situation the coefficient vector denotes the degree

to which a change in each predictor affects a change in the predicted category (i.e., emotion  $E_j$ ). So for example, if the  $\beta_{ij}$  represents the  $i^{th}$  standardized coefficient in predicting  $E_j$ , a one standard deviation unit increase in the  $i^{th}$  predictor would lead to  $\beta_{ij}$  standard deviation increase in the value of  $E_j$ .

The feature selection algorithm operates by estimating the  $\beta_j$  parameter set for predicting  $E_j$  as indicated by each affect judge. After  $\beta_j$  is estimated, a filtering process is initiated. This process selects elements from  $\beta_j$  that make substantial contributions in predicting  $E_j$ . A variety of filtering schemes can be utilized. One involves selecting the statistically significant predictors. Another involves performing a sensitivity analysis on the feature set and obtaining the most sensitive predictors. Or a tolerance analysis can be performed. Although the details of the filtering mechanism employed are not critical, what is important is that the feature set must be reduced into a vector of the most critical predictors. This estimation plus filtering process is repeated for the affect ratings of the different judges.

Hypothetical output of the feature selection algorithm after completing the preliminary estimation and filtering phases is presented in Table 4. Each cell represents whether a particular predictor significantly predicts the affect judgments of each judge. Pluses (+) indicate that a particular feature is a positive predictor of the emotion. Minuses (-) indicate that a particular feature is a negative predictor of the emotion. Empty cells represent predictors that were discarded in the filtering process. This algorithm was run on a simulated feature set of four predictors ( $f_1, f_2, f_3, f_4$ ) with affect ratings being

provided by three judges ( $j_1, j_2, j_3$ ). For example, consider the first row of Table 4. It appears that  $f_1$  was a significant predictor of  $E$ , irrespective of the judge providing the ground truth category in predicting that emotion (i.e., all cells are +). On the other hand, row 2 indicates that  $f_2$  is not a very reliable predictor because it was only reliable in the model where affect judgments were provided by  $j_1$ .  $f_3$  presents an interesting situation. This feature was a positive predictor for the  $j_1$  model and a negative predictor for the  $j_2$  model. This makes it an unreliable predictor that would not be included in the final model.

The third and final step of the algorithm involves selecting the most representative set of predictors from feature set  $F$ . This can be accomplished by simple heuristics. For example, one strategy would be to select a predictor where at least half the judges agree on magnitude and direction. With this strategy, the final set of features ( $G$ ) for the simulated case (See Table 4) would include  $f_1$  and  $f_4$ .

This simple case of performing binary discriminations can be extended to more complex cases of  $o$ -way discriminations by considering each emotion separately and discriminating it from every other emotion (grouped together). Separate feature sets are then estimated for each model.

**Table 4. Simulated output of feature selection algorithm**

Features	Coefficients		
	$j_1$	$j_2$	$j_3$
$f_1$	+	+	+
$f_2$	+		
$f_3$	+	-	
$f_4$	-	-	

The algorithm can be refined even further by adding an additional step in constructing the preliminary predictive models. As described in the introduction, individual differences in affect experience and expression have a profound impact on the generalizability of the features selected. Ideally, a feature set that best discriminates between the affective states and generalizes to new users is desired. This can be accomplished by constructing the preliminary predictive models in two phases. First, dummy coded participant variables are entered as predictors and a model is constructed (Level 1 model). The features are then added as predictors and another model is constructed (Level 2 model). Features that are statistical significant predictors in the Level 2 model are expected to generalize above and beyond individual differences because variability due to individual differences has been partialled out in the Level 1

models. This strategy known as “*disaggregated analysis with dummy-coded groups*” (P. Cohen, Cohen, West, & Aiken, 2002) is a common strategy used in constructing regression models. It can be used in the feature selection algorithm as an additional refinement in situations where individual differences is an important concern as is in the case of affect detection.

## **Features of AutoTutor’s Mixed-Initiative Dialogue**

A session with AutoTutor is comprised of a set of subtopics (main questions) that cover specific areas of the main topics (hardware, internet, and operating systems). Each subtopic has an associated set of expectations, potential dialogue moves to elicit expectations (e.g., hints, prompts, assertions), misconceptions, corrections of misconceptions, and other slots in the curriculum script that need not be addressed here. The expectations are ideally covered by a series of turns in AutoTutor’s conversation with the student in an attempt to help the student construct an answer to the current main question (subtopic). When an acceptable answer with the appropriate details is gleaned from the student’s responses (usually after 30–100 turns), AutoTutor moves on to the next subtopic. 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. Temporal information, such as the student’s reaction time and response time, is also recorded. Table 5 provides an overview of relevant information channels that are available in AutoTutor’s log files of the interaction history.

**Table 5. Description of the information mined from AutoTutor's log files.**

Channel	Sub channel	Description
Temporal Information	Real Time	Time in seconds since the beginning of the session
	Subtopic No.	The current subtopic (question) in this session
	Turn No.	The no. of the conversation turn within a subtopic
	Response Time	Time between question and answer submission
Response Information	No. of words	The number of words in the student's response
	No. of chars	The number of characters in the student's response
	Speech Act	Speech Act category of the student's response
Answer Quality Assessment	Local Good	Similarity of student's response to an expectation
	Delta Local Good	The change in the Local Good Score
	Global Good	Similarity of response history to expectations
	Delta Global Good	The change in the Global Good Score
	Local Bad	Similarity of student's response to a bad answer
	Delta Local Bad	The change in the Local Bad Score
	Global Bad	Similarity of response history to bad answers
	Delta Global Bad	The change in the Global Bad Score
Tutor Directness	Pump	Minimal information provided. e.g. "What else"
	Hint	Provides a hint to the student to fill in proposition
	Prompt	Prompts student to fill in a missing content word
	Correction	Corrects the student's misconception
	Assertion	Asserts information about an expectation
	Summary	Provides a summary of the answer
Tutor Feedback	Positive	Provides feedback terms such as: "good job",
	Neutral Positive	Provides feedback terms such as: "yeah", "right"
	Neutral	Provides feedback terms such as: "uh huh",
	Neutral Negative	Provides feedback terms such as: "kind of"
	Negative	Provides feedback terms such as: "wrong", "no"

*Note.* The various sub channels for tutor directness and tutor feedback channels are ordered onto two individual scales. Therefore, the number of dialogue predictors is taken to be 17 and not 26.

### *Temporal Information*

Temporal information can be viewed as a combination of global and local temporal markers that span the period of interaction. *Real time* measures the time of a dialogue event in the tutoring session, and is measured in milliseconds but rounded to seconds for ease of interpretation. The *subtopic number* indicates the number of main questions answered. It provides a global measure of sequential position within the entire tutorial session. For example, for a one-hour session covering three subtopics, the third subtopic would indicate that the student is approximately in the 40–60 minute time span. The *turn number*, on the other hand, provides a local temporal measure. It is the  $n^{\text{th}}$  turn of the student in the current question (subtopic). Finally, the *student response time* is the elapsed time (in milliseconds converted to seconds for easy interpretation) between the verbal presentation of the question by AutoTutor and the student submitting an answer.

### *Response Information*

AutoTutor uses LSA for the majority of its assessments of the student's responses to a question, as will be discussed below. Another measure that was considered is the *verbosity* of the student's responses. The verbosity is measured by the *number of words* and the *number of characters* in the student's response. Another measure of the student's response to AutoTutor is based on a classification a speech act classification of the response (Olney et al., 2003). The system classifies each response into one of a number of categories; those of interest in this research involve topic-unrelated *frozen expressions* (e.g., I don't know, What did you say?) and topic-related *contributions* (scored as a 1).

### *Answer Quality Assessments*

AutoTutor relies on Latent Semantic Analysis (LSA) (Graesser, Penumatsa et al., 2007; Landauer & Dumais, 1997; Landauer et al., 2008) as its primary computation of the quality of student responses in student turns. The local assessments for a given turn  $N$  measure the student's response for that turn on the basis of its similarity to good answers (expectations) and bad answers (misconceptions and bugs). The *local good score* is the highest match score between the content of student turn  $N$  and the set of expectations representing good answers. The local *bad score* is the highest match to the set of bad answers. A high local good score reflects progress in answering the main question, whereas a high local bad score reflects resonance with misconceptions. The *delta local good score* and the *delta local bad score* measure changes in the local good score and the local bad score, respectively, compared with student turn  $N - 1$ .

The four global parameters (see Table 5) perform the same assessments as the local parameters with the exception that the text used for the LSA match is an aggregation of all of the student's turns (1 through  $N$ ) for a given subtopic. With this scheme, a student's past responses to a subtopic are considered in AutoTutor's assessment of the student's current response.

### *Tutor Directness*

At the end of each student turn, AutoTutor incorporates the various LSA assessments when choosing its next pedagogically appropriate dialogue move. When AutoTutor tries to get a single expectation ( $E$ ) covered (e.g., "The hard disc is a storage medium"), this

goal is posted and is achieved by AutoTutor presenting a series of different dialogue moves across turns until the expectation  $E$  is expressed by the student or as a last resort by the tutor. It first gives a *pump* (What else?), then a *hint* (What about the hard disk?), then a *prompt* for a specific important word (The hard disk is a medium of what?), and then simply *asserts* the information (The hard disc is a medium for storage). After all of the expectations for the problem are covered, a *summary* is provided by AutoTutor. Given this mechanism of encouraging the student to cover the expectations, the dialogue moves chosen can be ordered on a *directness* scale (ranging from  $-1$  to  $1$ ) on the basis of the amount of information AutoTutor supplies to the learner. The ordering is  $pump < hint < prompt < assertion < summary$ . A pump conveys the minimum amount of information (on the part of AutoTutor) whereas a summary conveys the maximum amount of explicit information.

#### *Tutor Feedback*

AutoTutor's short feedback (positive, neutral, negative) is manifested in its verbal content, intonation, and a host of other non-verbal conversational cues. Table 5 shows examples of AutoTutor's responses, characterized by the type of feedback being provided. Similar to the directness scale constructed above, AutoTutor's feedback was mapped onto a scale ranging from  $-1$  (negative feedback) to  $1$  (positive feedback).

## **Results and Discussion**

The present study evaluated whether conversational dialogue features are a viable channel for affect detection. The data used for the analyses was from the multiple annotator study in which 28 participants interacted with AutoTutor on topics in computer literacy (see Chapter 2). When aggregated across each 32–38 minute session for each of the 28 participants, 1470 student–tutor interaction turns and 2967, 3012, 3816, and 3723 emotion judgments were obtained for the self, peer, trained judge 1, and trained judge 2, respectively. A dialogue feature vector based on the features listed in Table 5 was then extracted for each student–tutor interaction turn. The feature vector was then associated with an emotion category on the basis of the human judges' affect ratings. More specifically, the emotion judgment that immediately followed a dialogue move (within a 15-second interval) was bound to that dialogue move. This data collection procedure yielded four ground truth models of the learner's affect, so four labeled data sets were constructed.

Affect judgment reliabilities between the human judges presented above revealed that the highest agreement was obtained between the trained judges ( $\kappa = .36$ ). However, it is still not firmly established whether the trained judges or the self judgments are closer to the ground truth. This issue was addressed by combining affect judgments from the four judges in order to obtain a better approximation of the learner's emotion. In particular, one data set was constructed on the basis of judgments in which both trained judges agreed. Another was constructed for judgments in which any two (or more) judges

agreed. Similarly, a third data set was constructed for the affect judgments in which three (or more) judges agreed. A fourth additional data set was constructed for judgments in which all four judges agreed, but was eliminated from the subsequent analyses because of a very small sample size ( $N = 66$ ). The frequencies of the emotions in each data set are listed in Table 6.

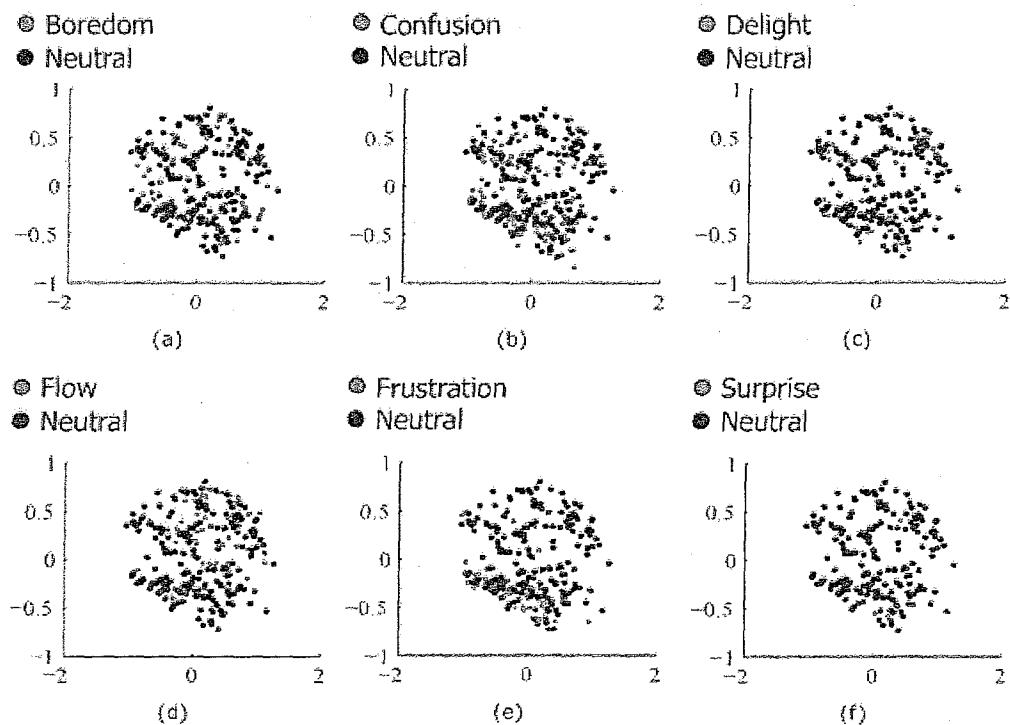
**Table 6. Frequency of affective states in each data set.**

Judge	Affective States								Sum
	Boredom	Confusion	Delight	Flow	Frustration	Neutral	Surprise		
Self	164	172	52	176	167	265	28		1024
Peer	189	172	25	177	97	344	36		1040
Judge 1	104	258	82	190	139	321	21		1115
Judge 2	242	238	67	102	58	390	12		1109
Trained Judges	81	150	61	67	62	196	6		623
Two Agree	144	167	42	105	84	326	6		874
Three Agree	64	90	22	49	30	154	1		410

### *Data Visualization and Preliminary Predictions*

Principal component analyses were applied to each of the seven data sets to reduce the feature vectors to two dimensions for visualization purposes. Figure 4 and Figure 5 show the two-dimensional plots of the dialogue feature vectors obtained from the integrated affective model in which both of the trained judges agreed. Each plot depicts the dialogue-feature vector with respect to the first principal component (*X*-axis) against the second principal component (*Y*-axis). The plots can be useful in making preliminary predictions regarding the potential difficulty that a computer may face in classifying affect from dialogue. However, the predictions should be interpreted with some caution because each feature vector was reduced to only two dimensions. Moreover, these plots are based on the data set in which the two trained judges agreed. This constitutes only one of the seven data models.

Figure 4 depicts six sub-plots in which the dialogue features of each affective state are contrasted with the state of neutral. A number of predictions relating to the perceived difficulty in automatically detecting each affective state from neutral can be made on the basis of these plots. In particular, a degree of difficulty in discriminating between boredom (Figure 4a), confusion (Figure 4b), and delight (Figure 4c) from neutral can be expected because these emotions do not seem to be linearly separable from neutral (i.e., a line cannot divide the delight instances from the neutral instances). The affective states of flow (Figure 4d) and frustration (Figure 4e) seem to be segregated from neutral, with the dialogue feature vectors for flow predominantly occupying the upper part of the space in contrast to those for frustration, which appear in the lower part of the space.

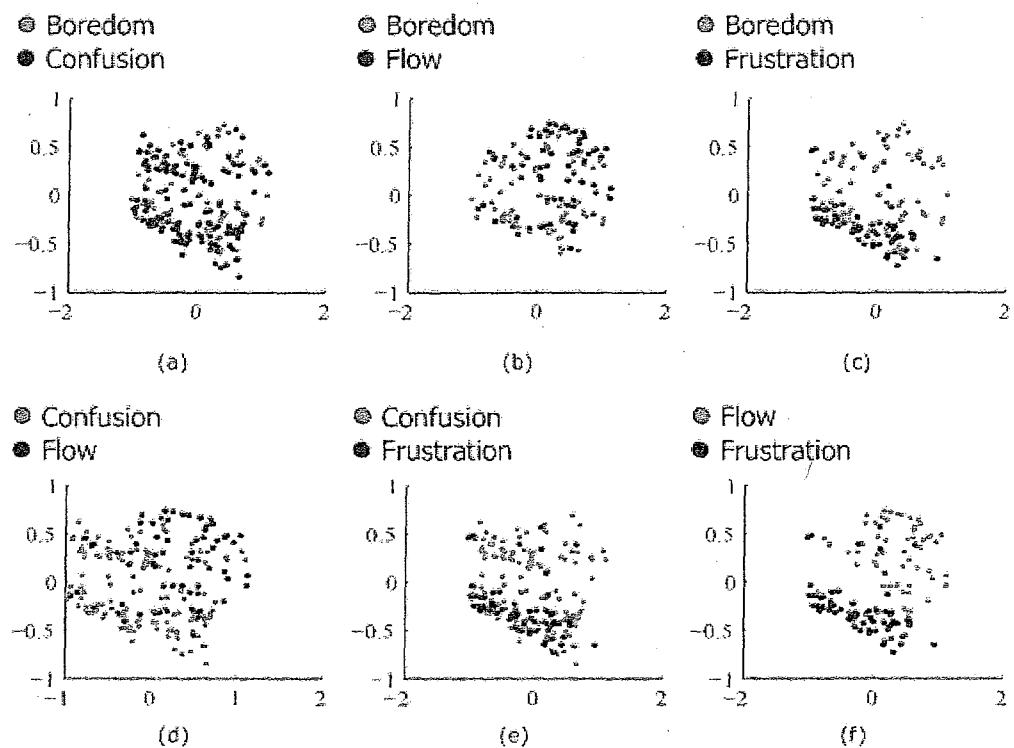


**Figure 4.** Two-dimensional representations of each emotion versus neutral.

The  $X$ - and  $Y$ -axes in each plot represent the first and second principal components, respectively. The data set displayed was based on emotions judgments in which both trained judges agreed.

Similar predictions regarding the discrimination between the various affective states can be made from the plots that offer views of dual affective states for the six combinations of boredom, confusion, flow, and frustration (Figure 5). In particular one can expect some difficulty in discriminating boredom from confusion and flow but little difficulty in discriminating it from frustration. It also appears that one could predict a reduction in accuracy in detecting confusion from both flow. The classifiers should not have any significant problems in discriminating between flow and frustration because

these two affective states seem to be linearly separable. Plots contrasting the affective states of delight and surprise are not presented in this chapter because the statistical analyses (presented next) indicated that these emotions could not be predicted from the dialogue features.



**Figure 5. Two-dimensional representations of various combinations of the affective states.**

The  $X$  and  $Y$ -axes in each plot represent the first and second principal components, respectively. The data set displayed was based on emotions judgments where both trained judges agreed.

### *Illustration of Feature Selection Algorithm*

This section illustrates the use of the feature selection algorithm described above.

Multiple regression analyses were conducted to determine the extent to which the seven affective states of interest could be predicted from the various dialogue features. For each of the seven data sets (self, peer, trained judge1, trained judge2, trained judges agree, any two judges agree, and any three judges agree), seven multiple regression models were constructed, one for each of the affective states, yielding 49 models in all. The criterion variable for each multiple regression analysis was the affective state (1 or 0 if present or absent respectively) whereas the predictor variables were the set of dialogue features. It is widely acknowledged that strongly correlated predictor variables tend to cause instability in multiple regression models. Therefore, the analyses first identified and subsequently eliminated the collinear dialogue features. A correlation threshold (Pearson's  $r > .7$ ) was adopted as the criterion to remove collinear predictor variables. Specifically, when two variables were identified as collinear predictors, the one with a stronger overall correlation with the affective states was preserved. This approach reduced the 17 features to 11 features by discarding real time, local bad score, global bad score, delta local bad score, delta global bad score, and number of words.

In order to partial out variability among participants, the multiple regression analyses were conducted in two steps. In step 1, the predictors included the participants' pretest scores and dummy coded variables to differentiate participants. In step 2, the group of predictors was the 11 different conversation features after six potential predictors were excluded as a result of the collinearity analysis. Step 1 was entered first,

with the residual variance passed on to step 2. In this fashion, any of the variability associated with the participants' characteristics can be partialled out and unique variance that could be ascribed to particular conversation features can be assessed.

*Analysis of Regression Models.* Statistically significant overall relationships (at the  $p < 0.05$  level) were discovered for boredom, confusion, flow, frustration, and neutral, but not for delight and surprise. In the cases of delight and surprise, step 1 was usually significant but not step 2. Such a result implies that the various conversation features were unable to add a significant improvement in classifying affect states above and beyond participant characteristics. Table 7 presents the goodness of fit and significance parameters of the multiple regression models.

A number of conclusions can be drawn from the characteristics of the regression models presented in Table 7. For the affective state of boredom, when aggregated across all 7 models, the features explained about 16.9% of the predictable variance, with 7.1% of the variance being accounted for by the step 2 conversation features alone (see last row of Table 7). Similarly, on average, 14.4% of the variance of affect classification was explained for confusion, with 7.2% obtained from the conversation features. For flow and frustration, the conversational features accounted for 9.2% and 5.9% of the total variances of 17% and 13.2%, respectively. Additionally, the data sets based on face value judgments of the self and peer failed to converge on a statistically significant model for flow. The flow emotion is difficult to detect from the affect ratings made by the novice judges.

**Table 7. Summaries of the multiple regression models for emotions in each data set.**

Rating Type	Model	df1,df2	Affective States									
			Boredom		Confusion		Flow		Frustration		Neutral	
			$R^2_{adj}$	F	$R^2_{adj}$	F	$R^2_{adj}$	F	$R^2_{adj}$	F	$R^2_{adj}$	F
Self	PC	27,996	.134	6.86	.123	6.33	-	-	.129	6.59	.298	17.12
	PC+DF	11,985	.162	4.00	.171	6.26	-	-	.161	4.46	.315	3.18
Peer	PC	27,1012	.162	8.44	.097	5.14	-	-	.085	4.57	.275	15.63
	PC+DF	11,1001	.208	6.39	.107	1.98	-	-	.116	4.26	.287	2.44
Trained	PC	27,1087	.072	4.18	.032	2.37	.098	5.50	.025	2.04	.013	1.54
Judge1	PC+DF	11,1076	.122	6.70	.082	6.33	.194	12.80	.107	10.15	.029	2.67
Trained	PC	27,1081	.048	3.06	.036	2.51	.046	2.98	.054	3.34	.075	4.31
Judge2	PC+DF	11,1070	.128	10.03	.140	12.91	.135	11.09	.103	6.38	.105	4.32
Trained	PC	27,595	.062	2.53	.040	1.96	.082	3.05	.039	1.95	.092	3.34
Judges	PC+DF	11,584	.159	7.22	.123	6.13	.178	7.37	.138	7.19	.117	2.51
Two Judges	PC	27,846	.082	3.88	.064	3.21	.080	3.83	.063	3.19	.057	2.97
	PC+DF	11,835	.160	8.17	.159	9.67	.147	7.03	.138	7.70	.097	4.36
Three Judges	PC	27,382	.128	3.22	.114	2.96	.084	2.39	.120	3.07	.161	3.92
	PC+DF	11,371	.245	6.40	.229	6.16	.196	5.85	.164	2.84	.186	2.04
Mean $R^2_{adj}$ .	PC		0.098		0.072		0.078		0.074		0.139	
	DF		0.071		0.072		0.092		0.059		0.024	
	PC+DF		0.169		0.144		0.170		0.132		0.162	

*Notes.* PC: participant characteristics. DF: dialogue features. All models significant at the  $p < .05$  level.

*The Relationship between Dialogue Features and Learner's Affect.* A number of generalizations regarding the relationship between dialogue and affective states can be gleaned on the basis of the numerical direction (i.e., signs, + -) of the statistically significant coefficients of the multiple regression models (see Table 8). A number of relationships surfaced when one considers the significant predictors of the affective states where at least two judges agreed. In particular, boredom occurs later in the session (high subtopic number), after multiple attempts to answer the main question (high turn number), and when AutoTutor gives more direct dialogue moves (high directness). Alternatively, confusion occurs earlier in the session (low subtopic number), within the first few attempts to answer a question (low turn number), with slower responses (long response time), shorter responses (fewer characters), low quality answers (low local good LSA scores), with frozen expressions (negatively coded speech acts), when the tutor is less direct in providing information, and when the tutor provides negative feedback. The analyses indicated that flow occurs within the first few attempts to answer a question (low turn number), with quicker, longer, proficient responses (low response time, more characters, and high local good LSA score respectively), and is accompanied by positive feedback from the tutor. Frustration was prevalent later in the temporal span of a session (high subtopic number), with longer response times, with good answers towards the immediate question (high local good score), but poor answers towards the broader topic (low global good score), and negative tutor feedback.

**Table 8. Significant predictors for the multiple regression models.**

Dialogue Features	Affective States				
	Boredom	Confusion	Flow	Frustration	Neutral
	S P J1 J2 JA 2 3				
Subtopic No.	+++ + + ++	- - - - -	--	++	-
Turn No.	+++ + + ++	- --	- -	+	
Response Time		+ + ++	- - -	++	- - -
No. Characters		- - - +	+ + + ++		
Global Good				- - -	
Del. Glbl. Good					
Local Good	-	- - -	+	++ + +	
Del. Local Good	-	- - - -			++ + ++
Speech Act	-	- - - -			
Directness	++ +	- -			
Feedback	+	- -	++ + +	- - - -	++ + + +

*Notes.* S: Self Judgments, P: Peer Judgments, J1: Trained Judge1, J2: Trained Judge2, JA: Both trained judges agree, 2: Any two judges agree, 3: Any three judges agree + or - indicates that the feature is a positive or negative predictor in the multiple regression model at  $p < .05$  significance level. Empty cells indicate that a feature was not a statistically significant predictor for the respective emotion.

The relationships between the various dialogue features and the affective states described above are generally intuitive and in the expected directions. Of particular interest are the features that predict the affective state of frustration. In the case of frustration, the learner tends to have not been doing well on the topic in general, as indicated by a negative global LSA score, but they have taken a longer time to answer the question (increased response time) and have given a good answer to the immediate question (higher LSA local good score). However, AutoTutor's internal model of the learner's interaction has erroneously classified the student as being a poor learner and responds with increased negative feedback, which in turn increases frustration. It should be noted that in this case it would appear that the learner generally made an effort, as indicated by the increased response time and higher local good score, but did not receive the positive response expected. This would suggest that this type of frustration could be alleviated by dispensing more positive feedback in cases where a learner who has generally performed poorly takes the time to give a good response.

*Segregating Participant Characteristics from Dialogue Features.* A critical hurdle that accompanies applications in the field of user modeling is whether observed patterns generalize across different participants (i.e., are they universal or person-specific?) (Picard, 1997). In situations where relationships between patterns do not generalize, it is important to introduce some degree of normalization for each participant.

The multiple regression models described above were constructed in two steps, such that the dialogue features were excluded from the first step and were reintroduced in the second step. While this procedure quantitatively separates any explained variance into

two groups, it fails to confirm whether any of the relationships observed in Table 8 would reflect the commonality of variance between participant characteristics and dialogue features. That is, which of these two sources of variance should get credit when there is commonality of variance between the two?

One way to answer this question involved reversing the order in which the two sets of predictors are entered into the two-step multiple regression analyses. If step 1 regression models are constructed on the basis of the 11 dialogue features only and the 29 participant characteristics variables are reserved for the step 2 model alone, one can determine whether differences among college students are significant above and beyond the dialogue features. Specifically, if a particular feature was a statistically significant predictor of an affective state in step 1 of the regression analyses (dialogue features only) as well as in step 2 (dialogue features plus participant characteristics) one could reliably conclude that any covariation observed between this feature and the affective state is valid and not caused by the participants' characteristics. On the other hand if a dialogue feature was a significant predictor in step 1 but not in step 2, then one would be forced to conclude that the relationship between the predictor and the affective states was a characteristic of individual learners and could not generalize beyond individual participants.

In accordance with this reverse regression procedure, the multiple regression analyses were repeated with the order of the dialogue features and participant characteristics reversed (dialogue features for model 1, dialogue features + participant characteristics for model 2). An examination of the significance and direction (+ -) of the

statistically significant predictors across both steps of the regression models did not reveal any serious contradictions to what was reported in Table 8.

The relationships between the various dialogue features and the affective states described above are generally intuitive and in the expected directions. Of particular interest are the features that predict the affective state of frustration. In the case of frustration, the learner tends to have not been doing well on the topic in general, as indicated by a negative global LSA score, but they have taken a longer time to answer the question (increased response time) and have given a good answer to the immediate question (higher LSA local good score). However, AutoTutor's internal model of the learner's interaction has erroneously classified the student as being a poor learner and responds with increased negative feedback, which in turn increases frustration. It should be noted that in this case it would appear that the learner generally made an effort, as indicated by the increased response time and higher local good score, but did not receive the positive response expected. This would suggest that this type of frustration could be alleviated by dispensing more positive feedback in cases where a learner who has generally performed poorly takes the time to give a good response. Therefore, it is possible to conclude that the set of dialogue features does covary with the affective states above and beyond the participant characteristics.

#### *Dimensionality Reduction*

Dimensionality reduction is an important phase in machine learning experiments. In addition to potentially increasing classification accuracy by eliminating unrelated

features, computational advantages are gained in terms of execution time. Therefore, the data were preprocessed before attempting to classify affect from dialogue. Two methods of dimensionality reduction, one based on feature selection and the other based on feature extraction, were pursued. The “feature selection first” method consisted of a supervised selection of features on the basis of the collinearity analyses and the multiple regression analyses described above. In particular the statistically significant standardized coefficients (after the elimination of 6 highly collinear features) listed in Table 8 of the multiple regression analyses were used as the features for the classifiers.

The second dimensionality reduction technique involved extracting features by principal component analyses (PCA) and linear discriminant analysis (LDA). The PRAAT environment (Boersma & Weenink, 2006) was used to accomplish the requisite computation. The combination of these methods have been widely used and well validated as robust dimensionality reduction techniques, particularly in extracting features from speech.

The analyses proceeded by first applying PCA to all seven datasets, each containing the complete set of seventeen features, and then dynamically reducing the dimensionality on the basis of the number of eigenvectors that accounted for 97% of the variance (typically 12–14). LDA was then applied to the decorrelated features to project them onto a lower dimensional space on the basis of the number of discriminant functions developed (number of classes – 1). Machine learning experiments were conducted, with the classifiers being trained on features extracted by the combined use of PCA and LDA as well as separately utilizing PCA and LDA. These strategies yielded classification

accuracies that were slightly better than chance and are not discussed in the subsequent section. This may be due to the fact that the set of predictors is quite small ( $N = 17$ ) and hence extracting a subset may have reduced the predictive power.

#### *Classifying Affective States from Conversation Features*

Seventeen standard classification techniques were applied in an attempt to detect the various affective states based on dialogue features. The motivation behind using a relatively larger set of classifiers was to determine which classifier yields the best performance. It also would be interesting to determine whether classifiers from any particular category (trees, rules, etc.) out performs the others.

The Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2005) was used to comparatively evaluate the performance of various standard classification techniques in detecting affect from dialogue. The classification algorithms tested were selected from a list of categories including Bayesian classifiers (Naive Bayes and Naive Bayes Updatable), functions (Logistic Regression, Multilayer Perceptron, and Support Vector Machines), instance based techniques (Nearest Neighbor, K\*, Locally Weighted Learning), meta classification schemes (AdaBoost, Bagging Predictors, Additive Logistic Regression), trees (C4.5 Decision Trees, Logistic Model Trees, REP Tree), and rules (Decision Tables, Nearest Neighbor Generalization, PART).

The classification process proceeded in three phases. In the first stage, the reliabilities of classifiers in discriminating between boredom, confusion, flow, frustration, and neutral were assessed. In the second phase, neutral was eliminated and this reduced

the scope to boredom, confusion, flow, and frustration. This is a challenging task because the standardized coefficients of the regression models revealed that the diagnosticity of some of the dialogues features with respect to these four affective states was quite low. In the third phase of the classification analyses, the accuracies of detecting each of the four affective states from the base state of neutral were computed. Specifically, the classification algorithms were compared in their ability to differentiate boredom, confusion, flow, or frustration from neutral. There were challenges in this analysis as well. Although the multiple regression models' analyses provided a significant model for detecting the affective state of neutral, most of the variance associated with neutral was accounted for by the participants' pretest scores and overall emotion ratings; that is, the dialogue features were not very proficient predictors of neutral affect. The multiple regression analyses also failed to converge upon statistically significant models for delight and surprise, so these emotions were excluded from the classification analyses.

A uniform baseline for the different emotions was established by randomly sampling an equal number of observations from each affective state category. This sampling process was repeated for ten iterations and all reported reliability statistics were averaged across these ten iterations. For example, consider the task of detecting confusion from neutral with affect labels provided by the self. In this case, an equal number of confusion and neutral samples would be randomly selected, thus creating a data set with equal prior probabilities of both these emotions. Each randomly sampled

data set was evaluated on the seventeen classification algorithms and reliability statistics were obtained using  $k$ -fold cross-validation<sup>1</sup> ( $k = 10$ ).

A three factor repeated measures analysis of variance (ANOVA) was performed in order to comparatively evaluate the performance of the classifiers in detecting affect from the dialogue features.<sup>2</sup> The first factor (*judge*) was the judge or combination of judges that provided the affect judgments. This factor had seven levels: self, peer, trained judge 1, trained judge 2, trained judges agree, any two judges agree, and any three judges agree. The second factor involved the *emotions* classified and was composed of six levels: collectively discriminating between boredom, confusion, flow, frustration, and neutral (level 1, chance = 20%), discriminating between boredom, confusion, flow, and frustration (without neutral, level 2, chance = 25%), and individually detecting boredom, confusion, flow, and frustration from neutral (levels 3, 4, 5, and 6 respectively, chance = 50%). The third factor in the ANOVA was the classification scheme (called *classifier*) divided across six levels for Bayesian classifiers, functions, instance based learners, meta classifiers, rules, and trees. The unit of analysis for the  $7 \times 5 \times 6$  ANOVA was a single iteration of a single classifier. The kappa score was utilized as the metric to evaluate performance of each classifier because this metric partials out random guessing. The ANOVA indicated that there were significant differences in kappa scores across all three

---

<sup>1</sup> In  $k$ -fold cross-validation the data set ( $N$ ) is divided into  $k$  subsets of approximately equal size ( $N/k$ ). The classifier is trained on  $(k - 1)$  of the subsets and evaluated on the remaining subset. Accuracy statistics are measured. The process is repeated  $k$  times. The overall accuracy is the average of the  $k$  training iterations.

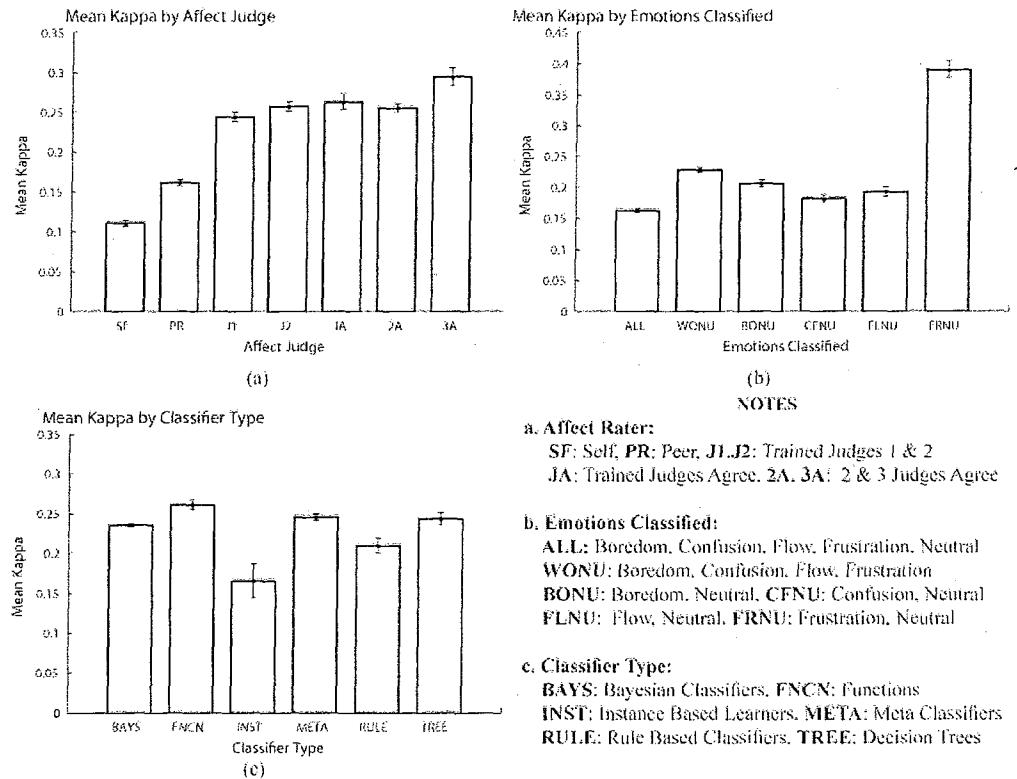
<sup>2</sup> Please see Appendix B for detailed results of the classification analyses.

factors, as well as for various interactions between the factors. On the basis of the ANOVA comparisons between the various levels of the three factors (rater, emotion, and classifier) are reported. Figure 6 graphically depicts the mean kappa score obtained from the emotion classification for each level of each factor of the ANOVA.

*Comparisons across Affect Judges.* The results of the ANOVA indicated that a statistically significant effect was obtained for the judge  $F(6,174) = 492.09$ ,  $MSe = .009$ ,  $p < .001$  (partial  $\eta^2 = .944$ ). Bonferroni post hoc tests revealed that classifiers evaluated on data where at least three judges agreed ( $M_{3A} = .295$ ,  $p < .01$ ) yielded the best performance. However, this finding should be interpreted with caution since this data set probably consists of the most obvious cases, namely when three or more judges were able to agree on an affective state. It was also the smallest data set with only 410 records. The post hoc tests indicated that there were no significant differences in kappa scores for classifiers based on combined affect judgments where two or more judges agreed, the two trained judges agreed, and the judgments of trained judge 2 ( $M_{2A} = .256$ ,  $M_{JA} = .263$ , and  $M_{J2} = .258$ ). Classifiers trained on data with affect labels provided by trained judge 1 were lower ( $M_{J1} = .245$ ) than these three scores. The lowest kappa scores were found in classifiers trained on affective judgments of the self ( $M_{SF} = .111$ ,  $p < .001$ ). Classifiers trained on affective judgments provided by the peer were significantly lower than all others with the exception of the self judgments ( $M_{PR} = .162$ ,  $p < .001$ ).

Mean kappa scores of .137, .252, and .258 are obtained if one aggregates the seven factors into three groups as novice judges (self and peer), trained judges (1 and 2), and combined models (trained judges agree, at least two, and three judges agree).

Therefore, one can conclude that reliability scores obtained by classifiers based on affect categories provided by trained judges and combined models were approximately the same and higher than those obtained by judgments provided by novice judges.



**Figure 6. Mean kappa across: (a) Affect Judge; (b) Emotions Classified; (c) Classifier Type.**

*Comparisons Across Emotions Classified.* The ANOVA revealed statistically significant differences in kappa scores among the emotions classified  $F(5,145) = 638.41$ ,  $MSe = .013$ ,  $p < .001$  (partial  $\eta^2 = .957$ ). Bonferroni post hoc tests indicated that the classifiers were most successful in detecting frustration from neutral ( $M_{FRNU} = .39$ ,  $p <$

.001). The classifiers had more success in collectively discriminating between boredom, confusion, flow, and frustration ( $M_{WONU} = .229$ ) than individually detecting boredom, confusion, and flow from neutral ( $M_{BONU} = .207$ ,  $M_{CFNU} = .182$ ,  $M_{FLNU} = .193$ ). Kappa scores obtained from efforts to detect boredom from neutral ( $M_{BONU} = .207$ ) were significantly higher than similar efforts in detecting confusion from neutral ( $M_{CFNU} = .182$ ).

The least robust results were obtained for the five affective states discrimination (boredom, confusion, flow, frustration, and neutral). In this case the mean kappa score of .163 ( $M_{ALL}$ ), was significantly lower ( $p < .001$ ) than all other combinations of emotions. Discriminating a larger number of affective states is challenging, particularly when the states are collected in an ecologically valid setting (i.e., no actors were used to express emotions and no emotions were intentionally induced). Additionally, these results are on par with kappa scores associated with human judges (e.g., self-peer = .08, self-judge1 = .14, self-judge2 = .16).

*Comparisons across Classifier Schemes.* The results of the ANOVA indicated that there were statistically significant differences in the kappa scores across the various classifier schemes  $F(5,145) = 44.83$ ,  $MSe = .03$ ,  $p < .001$  (partial  $\eta^2 = .607$ ). Bonferroni post hoc tests revealed that the kappa scores of the functions, meta, and tree based classifiers ( $M_{FNCFN} = .261$ ,  $M_{META} = .246$ ,  $M_{TREE} = .244$ ) were similar quantitatively and significantly ( $p < .01$ ) higher than the other categories. Kappa scores for the instance-based learning methods ( $M_{INST} = .166$ ) were significantly lower than the other five classifier categories ( $p < .001$  level). The post hoc tests also indicated that Bayesian

classifiers ( $M_{BAYS} = .236$ ) outperformed rule based classification schemes ( $M_{RULE} = .21$ ,  $p < .001$ ).

*Optimal Classification Accuracy.* The use of multiple assessments of the learner's affect ( $N = 7$ ) and a large number of classifiers ( $N = 17$ ) served an exploratory goal regarding the possibility of detecting affect from dialogue. However, in order to achieve the goal of developing a real time emotion classification system, the focus will be shifted to the data set and classifier that yielded the best performance. Table 9 presents the maximum classification accuracies obtained across all 17 classifiers for each of the 7 data sets in collectively discriminating between the various affective states, as well as in individually detecting each state from neutral.

**Table 9. Comparison of various classification techniques to detect learner's affect.**

Affective States	Base Rate	Max. Classification Accuracy (%)							
		Self	Peer	Judge		Judge	Judges	2	3
				1	2	Agree	Agree	Agree	Agree
<b>Boredom, Confusion,</b>									
Flow, Frustration,	20	29.5	31.4	38.3	42.4	42.4	41.4	42.2	
<b>Neutral</b>									
<b>Boredom, Confusion,</b>									
Flow, Frustration	25	35.1	38.8	50.4	50.5	54.0	52.5	52.7	
<b>Boredom, Neutral</b>									
Confusion, Neutral	50	58.9	59.4	61.2	62.6	60.9	65.4	67.8	
Flow, Neutral		52.9	56.0	66.8	70.0	70.5	63.9	67.3	
Frustration, Neutral		64.1	69.2	73.5	76.7	76.6	76.0	77.7	

Table 10 lists the F-measure scores for the affective states obtained from the most successful classifiers. The results indicate that when the classifiers attempted to collectively distinguish boredom, confusion, flow, and frustration, the reliabilities in detecting frustration and flow were similar and higher than those obtained for boredom and confusion.

When one considers the analyses that distinguished each affective state from neutral, the affective state of frustration was most easily distinguished from neutral. The F-measure<sup>3</sup> for the affective states of boredom, confusion, and flow were similar.

**Table 10. F-Measure for emotions.**

Affective States	F-Measure						
	Self	Peer	Judge 1	Judge 2	Judges Agree	2 Agree	3 Agree
Boredom	.32	.38	.39	.29	.39	.39	.43
Confusion	.35	.21	.33	.43	.36	.42	.42
Flow	.13	.27	.47	.51	.52	.49	.51
Frustration	.35	.41	.48	.56	.56	.50	.47
Neutral	.26	.23	.16	.25	.23	.19	.23
Boredom	.35	.44	.45	.37	.47	.48	.52
Confusion	.43	.31	.44	.44	.44	.48	.49
Flow	.33	.41	.59	.59	.63	.60	.62
Frustration	.41	.49	.58	.59	.63	.60	.59
Boredom	.57	.68	.64	.63	.68	.64	.71
Neutral	.67	.61	.65	.62	.66	.67	.71
Confusion	.63	.64	.69	.66	.66	.67	.68
Neutral	.57	.59	.59	.63	.62	.66	.72
Flow	.59	.64	.68	.69	.70	.64	.72
Neutral	.54	.55	.66	.71	.71	.67	.70
Frustration	.66	.71	.75	.79	.79	.78	.80
Neutral	.63	.68	.72	.75	.75	.73	.75

<sup>3</sup> The F-measure integrates precision and recall. The precision for class C is the proportion of samples that truly belong to class C among all the samples that were classified as class C. The recall score provides a measure of the accuracy in detecting a particular class.

*Comparisons of Computer Generated Categories with Human Assessments.* It is generally accepted that humans face some degree of difficulty in judging affect. The difficulty is apparent when reliabilities associated with assessments of emotions are compared with other psychological entities. For example, methodologists in the social and behavioral sciences have sometimes claimed that kappa scores ranging from 0.4–0.6 are considered to be fair, 0.6–0.75 are good, and scores greater than 0.75 are excellent (Robson, 1993). However, emotions classification does not reach such a high bar of interrater agreement. Interrater reliability scores across a variety of research efforts involving emotion measurements by humans are in general quite low. For example, Litman and Forbes-Riley (2004) report kappa scores of around .4 in detecting positive, negative, and neutral affect. Ang et al. (2002) report a kappa score of .47 in human judgments of frustration and annoyance in human-computer dialogue. Shafran, Riley, and Mohri (2003) report kappa scores ranging from .32 to .42 in coding affect. Recently, Grimm et al. (2006) reported kappa scores of .48 for humans detecting acted emotions. The highest kappa score obtained from the current study was only .36. Therefore, an interesting research question is how well affect categories generated by the various automated classification algorithms compares with human classification of emotions, which has modest reliability at best.

In order to compare the reliability between computer generated emotion categories and human judgment another set of classification analyses was conducted. These analyses focused on assessing reliabilities in collectively discriminating between boredom, confusion, flow, and frustration. Since the previous analyses revealed that a

simple logistic regression yielded the best performance in discriminating between these affective states, the subsequent analyses involve this classifier as representing the automated algorithms. In order to make a legitimate comparison with human judgments of affect, the combined data sets were not included in this analysis. Instead, classifiers were trained and evaluated on the four data sets that were constructed on the basis of individual judgments of the learner, a peer, and two trained judges.

In order to establish a uniform baseline, an equal number of observations was sampled for the affective states of boredom, confusion, flow, and frustration. This process was repeated for ten iterations and accuracy results were averaged across these ten iterations. After this sampling procedure, each of the four data sets (self, peer, two trained judges) were split into two parts: training data and testing data. The training data consisted of the dialogue features of 21 randomly sampled (without replacement) participants. The data for the remaining seven participants served as the testing data. The logistic regression based classifier was then trained on the training data of a single judge and correspondingly evaluated on the testing data associated with the other three judges. For example, the classifier trained on the 21-participant subset of the self data was tested on the 7-participant subsets of data from the peer, and the two trained judges. In this manner four versions of the logistic regression classifier were constructed, each version being trained on the data of the self, the peer, and the two trained judges. Additionally, this entire process was repeated for ten iterations, with each iteration involving the random sampling of a unique set of participants that constituted the training (75%) and testing (25%) data.

The results of the automated classifications are presented in Table 11, where they are contrasted with the judgments of the human coders. It should be noted that the reliability scores for the human coders differ from those reported earlier (i.e., self-peer = .08, self-judge1 = .14, self-judge2 = .16, peer-judge1 = .14, peer-judge2 = .18, and judge1-judge2 = .36). This is because the kappa scores for the human judges listed in Table 11 were only performed on a sample of the affective states of boredom, confusion, flow, and frustration in order to make valid comparisons with the machine generated categories of learner emotions.

**Table 11. Comparisons of computer and human classifications.**

Human Measurements					Computer Measurement				
		Rater 1						Testing	
Rater 2	Self	Peer	Judge 1	Judge 2	Self	Peer	Judge 1	Judge 2	Training
<i>Self</i>	-	.131	.299	.288	-	.100	.257	.229	<i>Self</i>
<i>Peer</i>	-	-	.333	.343	.121	-	.278	.273	<i>Peer</i>
<i>Judge 1</i>	-	-	-	.583	.124	.169	-	.349	<i>Judge 1</i>
<i>Judge 2</i>	-	-	-	-	.102	.117	.266	-	<i>Judge 2</i>

A number of conclusions can be drawn from the results of the human-human (left of Table 11) and computer-human interjudge reliability (right of table) scores presented in Table 11. Classifiers based on affect judgments of the self agreed with other human assessments of affect at a rate proportional to agreement between two humans (first row in Table 11). Human-computer agreement scores obtained when the classifiers were trained on the basis of the peer's affect judgments (second row) were slightly lower than human-human agreements for the two trained judges. Human-computer interjudge reliability scores obtained from classifiers trained on data provided by the two trained judges were significantly lower than the human-human interjudge reliability scores for the trained judges affect judgments ( $\text{judge1-judge2}_{\text{human}} = .583$  whereas  $\text{judge1-judge2}_{\text{computer}} = .349$ ,  $\text{judge2-judge1}_{\text{computer}} = .266$ ). On the basis of these observations, one can conclude that machine generated affect labels proportionally agree with novice judges (self and peer) but are inferior to trained judges.

## General Discussion

The problem of automating affect recognition is extremely challenging, on par with automating speech recognition. This project supports the conclusion that significant information can be obtained from AutoTutor's dialogue features, so dialogue can complement bodily measures for emotion detection. It appears that the classification accuracies obtained in this research on dialogue are not quite on par with the state-of-the-art algorithms that detect affect from facial features and speech contours. However, it

should be noted that over a decade of sustained efforts have been directed towards affect detection from facial expressions and speech. This project is one of very few research investigations that classify affect from dialogue alone, whereas earlier efforts used a small number of dialogue features in conjunction with acoustic-prosodic and lexical features. The results also constitute an improvement in classification accuracy compared to previous efforts that used dialogue features. This improvement can be primarily attributed to the diversity and richness of the current set of features. The features of dialogue in the analyses were specific to AutoTutor, but a similar set of features would presumably be relevant to any intelligent tutoring system, particularly in those that advocate deeper learning. In particular, the significant features that were extracted from AutoTutor's dialogue history logs (e.g., local good score, global good score, directness, etc.) would generalize to generic categories of dialogue features in all virtually all intelligent tutoring systems, such as content coverage, temporal parameters, response verbosity, student ability, tutor directness, and tutor feedback. There is also the possibility that some of the emotional predictors might be diagnostic of emotional states beyond intelligent learning environments. However, further research would be required to highlight these features.

This section highlights contributions of this research towards the field of affect detection and human-computer interaction. Some of the limitations of this research are identified and improvements in affect classification that might mitigate these limitations and extend this line of research are discussed.

### *Research Overview*

This chapter has addressed three major research goals. These included (1) the use of multiple human judges for ground truth affect labels, (2) a feature selection algorithm to select dialogue features that are the most diagnostic of the affective states, and (3) analyses of machine learning experiments that assess the accuracy of detecting affect from AutoTutor dialogue. This subsection summarizes and briefly speculates on the implications of some of the major findings.

*Multiple Measures of Learner Emotions.* Models of affect that combine judgments made by the learner, a peer, and trained judges may represent an advance over traditional techniques that often rely on self reports of affect or ratings by independent judges. The motivation behind these *composite* affect judgments of novice judges (self, peer) and trained judges resides in the indeterminacy of what exactly should be the gold standard for deciding what emotions a learner is truly having. In this study, would it be the learner or the expert on facial actions? A composite score that considers both viewpoints is arguably the most defensible position.

*Feature Selection.* The feature selection analyses resulted in significant dialogue predictors for boredom, confusion, flow, frustration, and neutral but not for delight and surprise. The feature selection method allowed us to select a set of predictors that discriminated between the affective states despite a degree of ambiguity in the ground truth affect labels. The two-step multiple regressions allowed us to statistically partial out variance attributable to individual differences (which was a robust amount of variance) before assessing the unique impact of the conversation features on emotions. After

partialling out individual differences, it was found that the dialogue features were able to explain about 7% of the variance for boredom, confusion, flow, and frustration. It is important to acknowledge that the explained variance is modest, but other researchers also report modest correlations and explained variance, as discussed throughout this chapter. Other researchers have not attempted to segregate the systematic variance that can be explained by individual differences per se, versus intrinsic features of the dialogue. The generalizability of these results to other learners is supported by the significant relationships between the various dialogue features and the affective states that persisted after the removal of variables related to the individual learner characteristics. It is of course conceivable that there are hidden factors or interactions among the predictors that could explain additional variance between dialogue and affect, but tests of that possibility would require additional analyses.

It is also conceivable that the tutor-centered actions have a distinct influence on the affective states of the learner. Tutor-centered actions are moves and utterances of the tutor (feedback and directness) rather than the student (number of words in response, speech act, LSA measures, etc.). This possibility could be explored by segregating the dialogue features listed in Table 5 into such categories as basic session information (subtopic and turn numbers), student-centered actions, and tutor-centered information. The amount of variance explained by each category of predictors could then be assessed by incrementally adding or removing each category in the multiple regressions analyses that predict affect categories. These analyses are planned in the future.

*Automated Detection of Learner's Affect.* Automated affect measurement is a difficult machine learning problem. Nevertheless the results indicate that the characteristics of the dialogue are quite diagnostic in predicting the affect states of learners. On the basis of the natural language dialogue features alone, the results showed that conventional classifiers are moderately successful in discriminating the affective states of boredom, confusion, flow, and frustration from each other, as well as from the baseline state of neutral.

The classification accuracy for collectively discriminating five affective states (including neutral) was significantly greater than the base rate. However, the reliability was lower compared to classifiers that individually detected each emotion from neutral or that collectively considered only the four emotions (excluding neutral). This motivates the use of a hierarchical classification scheme in order to improve accuracy. The hierarchical model would operate by first using a binary classifier to classify an incoming stimulus as a positive or a negative emotion, followed by an additional classification step that provides a finer discrimination as to what the individual positive or negative emotion may be. In a similar vein, a hierarchical classifier motivated by a pandemonium model can be considered (Selfridge, 1959). A collection of affect-neutral classifiers would first determine whether the incoming dialogue pattern resonated with any one or more of the emotions versus a neutral state. If there is resonance with only one emotion, then that emotion is declared as being experienced by the learner. If there is resonance with two or more emotions, then a second level of classification would be initiated in which classifiers collectively attempt to differentiate among boredom, confusion, flow, and

frustration. Rigorous tests of these alternative hierarchical models will be pursued in Chapter 4.

### *Limitations*

*Limited Contextual Information.* One limitation of the data analyses is that each emotion judgment was analyzed in the context of only the immediately preceding turns of the student and tutor. The dialogue features that involved changed scores (delta local good, delta global good, delta local bad, and delta global bad) did encompass the context of one previous turn, but these features did not prove to be predictive of the affective states. Perhaps classification accuracies could be boosted by incorporating a broader scope of contextual information, including patterns of conversation that evolve over a series of turns leading up to an emotional experience. The exclusion of this larger snapshot of context preceding an emotion utterance could possibly account for some of the lower classifier accuracies. Future efforts will be directed towards the analysis of conversation features across a larger temporal span and number of turns.

*Inability to Detect Delight and Surprise.* The dialogue channels were unable to detect the affective states of delight and surprise. Perhaps these affective states are simply not manifested in AutoTutor's conversation features and their detection would require more sophisticated sensors. Delight and surprise are affective states that are generally expressed through animated facial features, so it may be possible to detect these states by means of the Facial Action Coding System (FACS) (Ekman & Friesen, 1978). Ekman's research has associated action units 1 (inner brow raiser), 5 (raised upper eyelids), 26

(jaw drop), and 27 (mouth stretch) with surprise. Although Ekman and Friesen (1978) did not investigate delight, they have associated action units with happiness, an emotion that is presumably similar to delight. In particular, action units 6 (raised lower eyelid), 7 (lid tightener), 12 (lip corner puller), 26 (jaw drop), and 27 (mouth stretch) have been affiliated with happiness. With the assistance of automated facial feature tracking software, it might be possible to detect surprise and happiness (delight), thus compensating for the inability of detecting these affective states from AutoTutor dialogue.

*Reliance on Shallow Assessments of Performance.* A rather subtle limitation to the present results is that the affect detector relied exclusively on AutoTutor's assessments of the learner's contributions and its decisions regarding the type of feedback to give the student. Available research supports the claim that AutoTutor's assessments of the student's contributions highly correlate with human judgments (Graesser, Penumatsa et al., 2007; Graesser et al., 2000; Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999) and that AutoTutor's conversational patterns have a close correspondence with human tutors (Person & Graesser, 2002). However, AutoTutor's assessments are not error free or complete. There could be additions to the LSA scores in the computations of a learner's performance, such as measures of semantic entailment (Rus & Graesser, 2007; Rus et al., 2008) and the cohesion within each turn and across multiple turns. Cohesion refers to the linguistic properties of text that connect ideas conceptually. Perhaps student contributions with high cohesion may be indicative of a degree of understanding and would be diagnostic of the affective state of flow. On the other hand, a student contribution with

low cohesion may be diagnostic of the affective state of confusion. A system called Coh-Metrix provides over 100 measures of various types of cohesion, including referential, spatial, temporal, causal, and structural cohesion (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Louwerse, & Graesser, 2008). Future efforts will be devoted to expanding the set of dialogue features to include measures of cohesion.

## **Chapter 4: Automatic Affect Detection from Gross Body Language**

### **Introduction**

State-of-the-art affect detection systems have overlooked posture as a serious contender when compared to facial expressions and acoustic-proposed features, so an analysis of posture merits more close examination. There apparently are some benefits to using posture as a means to diagnose the affective states of a user (Bull, 1987; Demeijer, 1989; Mehrabian, 1968a, 1968b, 1971, 1972). Human bodies are relatively large and have multiple degrees of freedom, thereby providing them with the capability of assuming a myriad of unique configurations (Bernstein, 1967). These static positions can be concurrently combined and temporarily aligned with a multitude of movements, all of which makes posture a potentially ideal affective communicative channel (Coulson, 2004; Montepare, Koff, Zaitchik, & Albert, 1999). Posture can offer information that is sometimes unavailable from the conventional non-verbal measures such as the face and paralinguistic features of speech. For example, the affective state of a person can be decoded over long distances with posture, whereas recognition at the same distance from facial features is difficult or unreliable (Walk & Walters, 1988). Perhaps the greatest advantage to posture based affect detection is that gross body motions are ordinarily

unconscious, unintentional, and thereby not susceptible to social editing, at least compared with facial expressions, speech intonation, and some gestures. Ekman and Friesen (1969), in their studies of deception, have coined the term *nonverbal leakage* to refer to the increased difficulty faced by liars, who attempt to disguise deceit, through less controlled channels such as the body when compared to facial expressions. Furthermore, although facial expressions were once considered to be the objective gold standard for emotional expression in humans, there is converging evidence that disputes the adequacy of the face in expressing affect [see Barrett (2006) for a comprehensive review]. At the very least, it is reasonable to operate on the assumption that some affective states are best conveyed through the face, while others are manifested through other non-verbal channels.

In the popular scientific literature, Peter Bull in his book *Posture and Gesture* (1987) substantiates the case for role of posture in conveying emotions and attitudes of a user. He finds that participants express interest by leaning forward and drawing back their legs. In contrast, boredom is expressed by leaning back, dropping the head, leaning the head on one side, supporting the head with one hand, and frequently stretching out the legs. Agreement is expressed by leaning from side to side whereas the body communicates disagreement with the head erect, arms folded, and legs crossed around the knee.

Mehrabian attempted to decode the postures that accompany the attitudes of seated communicators. His results indicate that a positive attitude is expressed by leaning forward (similar to interest) and a notable decrease in a backward lean (Mehrabian,

1968a, 1968b; Mehrabian & Friar, 1969). Mehrabian and Friar (1969) also reported that a sideways lean occurs more frequently when the recipient of the message is of a lower status than the sender. Body openness (no crossed leg or folded arm) was consistently associated with a positive attitude whereas the arms' akimbo<sup>1</sup> position has a negative meaning (Mehrabian, 1968b).

More recently, and perhaps more relevant to this research, is a study by Coulson (2004) where participants judged the affective states of anger, disgust, fear, happiness, sadness, and surprise from computer generated static postures of mannequin figures. The figures were viewed from three perspectives, namely the front, the side, and the rear. Participants were able to accurately judge emotions from the static postures with concordance scores being the lowest for disgust and the highest (over 90%) for anger and sadness. Coulson also correlated the six affective states with various bodily movements (twisting the abdomen, bending the chest and head, and elbow, swinging and rotating the shoulders, and the general transfer of weight).

The Bull, Mehrabian, and Coulson studies described above, along with several others (Birdwhistell, 1975; Boone & Cunningham, 1998, 2001; Demeijer, 1989; Furnham, 1999; Montepare et al., 1999; Walk & Homan, 1984; Walk & Walters, 1988; Wallbott, 1998; Walters & Walk, 1986, 1988), make a very significant contribution to the literature on expression of emotion via body language. These findings have played a significant role in guiding the theoretical perspectives and research methods adopted

---

<sup>1</sup> Hands are placed on the hips with the elbows bowed outwards

here. However there are limitations in these prior research projects that motivated us to explore posture in more detail. One limitation is that the majority of these previous studies on posture have concentrated on the “basic” emotions (i.e., anger, fear, sadness, enjoyment, disgust, and surprise) (Ekman, 1992) rather than the learning-centered emotions.

Another limitation of the earlier research on posture and emotions is that human judges were required to manually decode the postures of the participants in order to associate these with their affective states. Notable posture coding schemes that were developed for this endeavor are the Posture Scoring System and the Body Movement Scoring System (Bull, 1987). Aside from the labor-intensive coding required by this methodology, there were additional complexities that arose from the fact that an affect-sensitive interface requires real-time detection of the user’s emotions. Therefore, until fairly recently, the development of an automated, real-time, posture-based, affect detection system remained an important but unattainable vision.

One option towards automated posture analysis is to use cameras and associated computer vision techniques to monitor body position and movement of a user. However, this approach is riddled with the problems that accompany nearly all computer vision-based applications, such as lighting, background conditions, camera angles, and other factors (Mota & Picard, 2003). Fortunately, there is a relatively new sensor that circumvents these challenges. In 1997 Tekscan™ released the Body Pressure Measurement System (BPMS), which consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The system provides pressure maps in real time

that can be analyzed for a variety of applications. For example, Tan and colleagues demonstrated that the BPMS system could be used to detect several static postures (leaning right, right leg crossed, etc.) quite reliably with an average recognition accuracy of 85% (Tan, Lu, & Pentland, 1997; Tan, Slivovsky, & Pentland, 2001).

Mota and Picard (2003) reported the first substantial body of work that used the automated posture analysis via the BPMS system to infer the affective states of a user in a learning environment. They analyzed temporal transitions of posture patterns to classify the interest level of children while they performed a learning task on a computer. A neural network provided real time classification of nine static postures (leaning back, sitting upright, etc.) with an overall accuracy of 87.6%. Their system then recognized interest (high interest, low interest, and taking a break) by analyzing posture sequences over a three-second interval, yielding an overall accuracy of 82.3%.

This chapter explores the possibility of using posture to automatically detect the affective states of college students during a tutoring session with the AutoTutor learning environment (Graesser, Lu et al., 2004a). A larger set of affective states than Mota and Picard (2003) are monitored, specifically the learning-centered affective states: boredom, flow, confusion, frustration, delight, and neutral. Some additional considerations arise because college students were monitored rather than children as in the Mota and Picard (2003) work. Children are much more active than the college students, so the algorithms used to detect affective states may differ. The movements of college students are more subtle, so it is important to pick up fleeting transitions in body pressure. Two different methods to infer affect from body movement were developed. Both of these methods

monitor gross body movements, rather than explicit postures, and hence did not require an additional training phase for static posture detection, as in the Mota and Picard (2003) work.

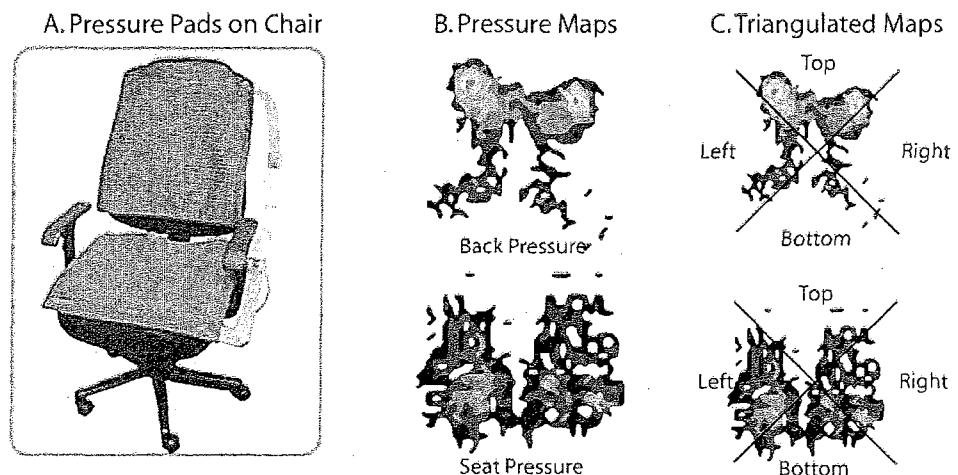
This chapter is organized in three sections. First, the BPMS system is described in some detail as well as two sets of posture features used to develop the affect-detector. Second, the Results section begins with a description of a series of experimental simulations that attempt to measure affect recognition accuracy. A summary of the major findings, limitations of the methodology, potential resolutions, and future work is presented in the General Discussion.

## **Architecture of the Posture Based Affect Detector**

### *The Body Pressure Measurement System (BPMS)*

The BPMS system, developed by Tekscan™ (1997), consists of a thin-film pressure pad (or mat) that can be mounted on a variety of surfaces. The pad is paper thin with a rectangular grid of sensing elements that is enclosed in a protective pouch. Each sensing element provides an 8-bit pressure output in mmHg. The current setup had one sensing pad placed on the seat of a Steelcase™ Leap Chair and another placed on the back of the chair (see Figure 7A).

The output of the BPMS system consists of a  $38 \times 41$  pressure matrix (rows  $\times$  columns) for each pad. Each cell in the matrix monitors the amount of pressure exerted on a single sensing element (see Figure 7B). During the tutoring intervention, at each sampling instance (1/4 second for the present study), matrices corresponding to the pressure exerted on the back and the seat of the chair were recorded for offline analyses.



**Figure 7. Body pressure measurement system.**

(A) The two pressure pads placed on the chair. (B) Pressure maps obtained from the pressure pads on the back and seat. (C) Pressure maps divided into four triangular areas for spatial analyses.

#### *High Level Pressure Features*

This feature set monitored the average pressure exerted on the back and seat of the chair along with the magnitude and direction of changes in pressure over a brief time window. Several features were computed by analyzing the pressure maps of the 28 participants recorded in the study. Six pressure-related features and two features related to the

pressure coverage for the back and the seat were computed, yielding 16 features in all.

Each of the features was computed by examining the pressure map at the time of an emotional episode (called the *current frame* or the frame at time  $t$ ).

Perhaps the most significant pressure related feature was the *average pressure*, which measured the mean pressure exerted in the current frame. This was computed by summing the pressure exerted on each sensing element and dividing the sum by the total number of elements. The average pressure is expressed in **Error! Reference source not found.** where  $R$  is the number of rows in the pressure matrix,  $C$  the number of columns, and  $p_{ij}$  is the pressure of a sensing element in row  $i$  and column  $j$ . For the current study,  $R = 38$  and  $C = 41$ .

$$\mu = \frac{1}{R \times C} \sum_{i=1}^R \sum_{j=1}^C p_{ij} \quad \text{Eq. 1}$$

There was another feature to detect the incidence of sharp forward versus backward leans, which ostensibly occurs when a learner is modulating his or her engagement levels. This feature measured the pressure exerted on the top of the back and seat pads. This was obtained by first dividing the pressure matrix into four triangular regions of equal area (see Figure 7C) and then computing the average pressure for the *top* triangular region. For the seat, this feature measured the force exerted on the frontal portion of the chair (sharp forward lean), while for the back it indicated whether the

shoulder blades of the learner were on the back rest of the chair (heightened backward lean).

The next two features measure the direction of pressure change. These include the *prior change* and *post change*, which measure the difference between the average pressure in the current frame ( $t$ ) and the frame  $J$  seconds earlier ( $t - J$ ) and  $K$  seconds later, ( $t + K$ ), respectively (see Eq. 2 and Eq. 3). For the current analyses,  $J = K = 3$  seconds. A positive prior change value is indicative of an increase in the pressure exerted, while a positive post change value reflects a reduction in pressure.

$$\Delta_{prior} = \mu_t - \mu_{t-J} \quad \text{Eq. 2}$$

$$\Delta_{post} = \mu_t - \mu_{t+K} \quad \text{Eq. 3}$$

The *reference change* (See Eq. 4) measured the difference between the average pressure in the current frame ( $t$ ) and the frame for the last known affective rating ( $r$ ). The motivation behind this measure was to calibrate the impact of the last emotional experience on the current affective state. It should be noted that unlike  $J$  and  $K$ , which were used to compute the prior and post changes, respectively,  $r$  is not a constant time difference. Instead  $r$  varies in time across different affective experiences. For the current analyses,  $r$  was 20 seconds for a majority of the instances since affect judgments were elucidated every 20 seconds. However, since the affect judges voluntarily offered judgments between the 20 second time slots, in several cases,  $r < 20$  seconds.

$$\Delta_{ref} = \mu_t - \mu_{t-r}$$

Eq. 4

Finally, the *average pressure change* ( $a_{pressure}$ ) measured the mean change in the average pressure across a predefined window of length  $N$  (see Eq. 5). The window was typically four seconds, which spanned two seconds before and two seconds after an emotion judgment.

$$a_{pressure} = \frac{1}{N} \sum_{i=1}^N |\mu_i - \mu_{i+1}|$$

Eq. 5

The two coverage features examined the proportion of non-zero sensing units (*average coverage*) on each pad along with the mean change of this feature across a four-second window (*average coverage change*). The computations for average coverage can be depicted as follows. Consider  $x_{ij}$  to be an indicator variable that determines whether the pressure ( $p_{ij}$ ) on sensing element  $ij$  is non-zero. Then:

$$x_{ij} = 1, \quad \text{if } p_{ij} > 0$$

$$x_{ij} = 0, \quad \text{if } p_{ij} = 0$$

The average coverage was the proportion of  $x_{ij}$  values that were non-zero as indicated by Eq. 6. Analogous to Eq. 4, the average coverage change is expressed in Eq. 7 as:

$$c = \frac{1}{R \times C} \sum_{i=1}^R \sum_{j=1}^C x_{ij} \quad \text{Eq. 6}$$

$$a_{coverage} = \frac{1}{N} \sum_{i=1}^N |c_i - c_{i+1}| \quad \text{Eq. 7}$$

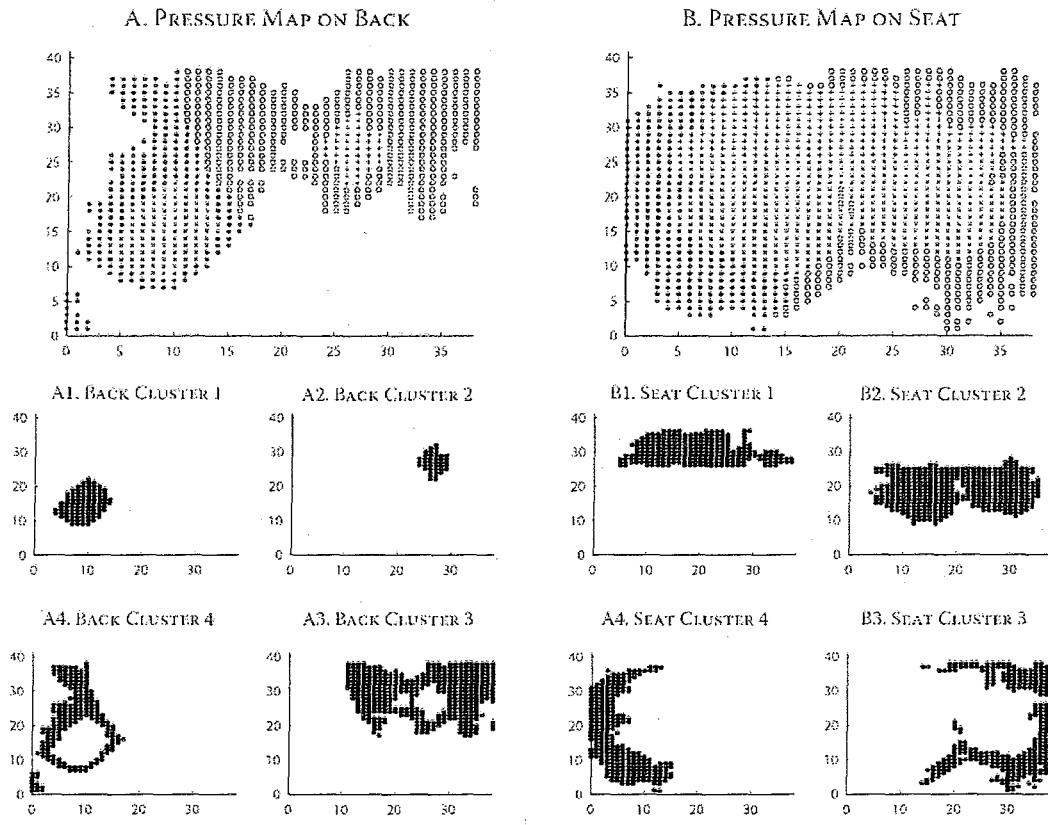
### *Spatial-Temporal Features*

The second set of features used for the posture affect detector involved monitoring the spatial distribution of pressure contours, and the magnitude by which they changed over time. Pressure contours were obtained by clustering the pressure maps for the back and seat of the chair using the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). The inputs to the EM algorithm were pressure values for each of the 1558 ( $38 \times 41$ ) pressure sensing elements from each sensor pad (i.e., back and seat). Each sensing element, along with the corresponding pressure exerted on it, was represented as a three dimensional point. The first two dimensions represented the relative position of the sensing element (i.e., its  $X$  and  $Y$  coordinate) on the sensor map while the third dimension was the pressure exerted on it.

The EM algorithm was configured to output four clusters based on earlier findings by Mota and Picard (2003) and preliminary experimental simulations in which the

number of clusters was varied ( $k = 2, 3, 4, 5, 6$ ). Figure 8 shows an example of the clustering data from the pressure maps with the EM algorithm. Since each data point was three-dimensional (i.e.,  $X$  and  $Y$  position co-ordinates and pressure), each cluster was represented by a 3D mean, a 3D standard deviation, and a prior probability (i.e., proportion of the 1558 data points that are included in the cluster). Consequently, seven features were extracted from each cluster: three for the means, three for the standard deviations, and one for the prior probability. By tracking four clusters on each pad 28 (4  $\times$  7) features were obtained. Additionally, since both the back and the seat are being tracked, there were  $28 \times 2 = 56$  features in all.

The aforementioned 56 features provide a snapshot of the spatial distribution of pressure exerted on the back and the seat of the chair while the learner interacts with AutoTutor. In order to obtain a measurement of arousal the change in each pressure contour (cluster) over a short time span was tracked. In particular, the pressure contours across a four second window were monitored and the means and standard deviations of each of the 56 features were computed. Therefore, effective dimensionality was 112. In this manner the features selected were spatial (distribution of pressure on pad) and temporal (changes in distribution over time).



**Figure 8. Clustering pressure maps on the back and the seat with the EM algorithm.**

The left half of the image is the output from the back while the right half is the output from the seat of the chair. The top quadrants (left and right) show the pressure maps on the back and seat. The eight plots on the bottom depict each individual clusters: four for the back (left) and four for the seat (right). Note that the clusters are based on position ( $X$  and  $Y$  co-ordinates of each sensing element) as well as intensity (pressure exerted on the sensing element).

#### *Hierarchical Classification via an Associated Pandemonium*

Perhaps the simplest method to develop an affect classifier is to experiment with several standard classifiers (neural networks, Naïve Bayes classifiers, etc.) and select the one that yields the highest performance in collectively discriminating between the affective states

of boredom, confusion, delight, flow, frustration, and neutral (surprise was not included in the affect detector since its occurrence was quite rare). However, since affect detection is a very challenging pattern recognition problem, on par with automatic speech recognition, developing a six-way affect classifier that is sufficiently robust is quite challenging. An alternative approach is to divide the six-way classification problem into several smaller two or three way classification problems.

One way to decompose the six-way classification problem into several smaller problems is to include a collection of affect-neutral classifiers that would first determine whether the incoming posture pattern resonated with any one or more of the emotions (versus a neutral state). If there is resonance with only one emotion, then that emotion would be declared as being experienced by the learner. If there is resonance with two or more emotions, then a conflict resolution module would be launched to decide between the alternatives. This would essentially be a second-level affect classifier. If three or more emotions are detected, then the second level classifier would perform a three-way classification task. In situations where the emotion expression is very murky, four- or five-way distinctions might be required as well.

Figure 9 depicts the manner in which the various classifiers are organized and interact. Central to the model lie five affect-neutral classifiers, each performing an emotion vs. neutral discrimination. To the left one finds ten classifiers that specialize in making binary discriminations among the affective states. On the top there are ten three-way emotion classifiers that are recruited when three or more affective states are detected

by the affect-neutral classification layer. On the right the various possibilities for four-way emotion classifiers are listed. Finally, the one five-way classifier lies at the bottom.

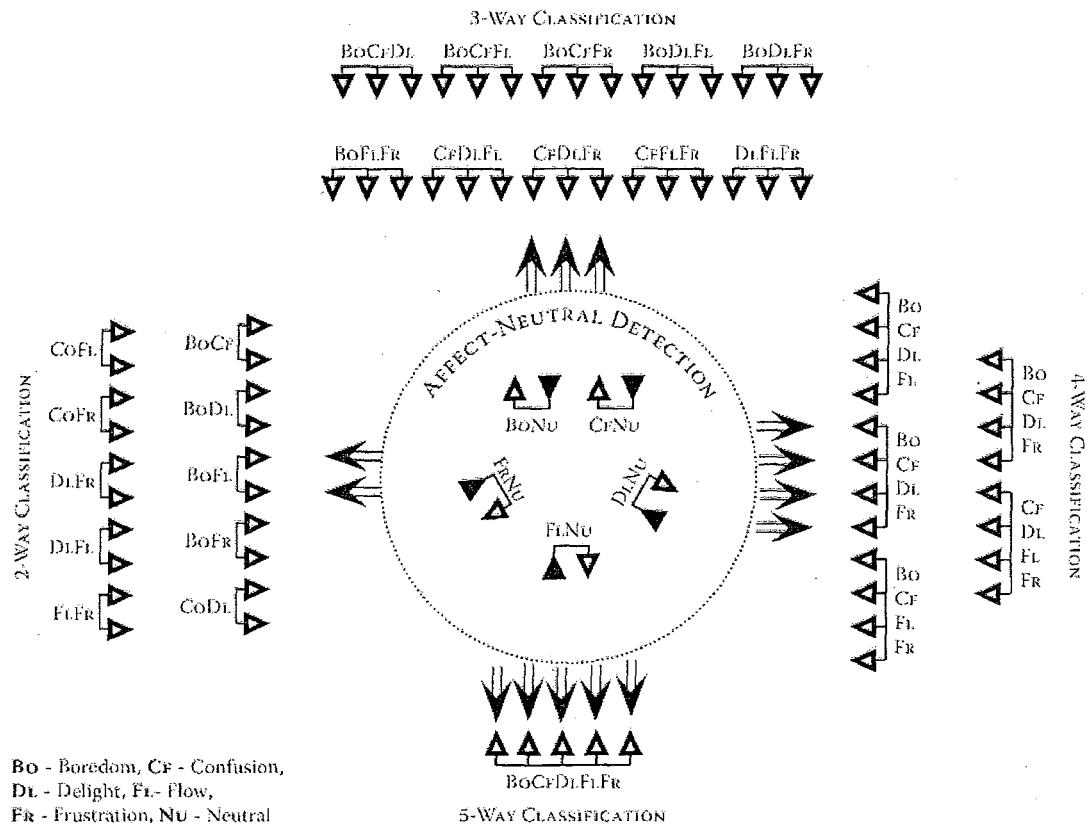


Figure 9. Hierarchical classification via an associated pandemonium.

As an example, consider a situation where the affect-neutral classifiers output [Boredom, Neutral, Neutral, Neutral, Neutral]. In this case the *Boredom – Neutral* classifier has detected boredom, while the other four affect-neutral classifiers have detected neutral. In this situation, the system would declare that the learner is

experiencing the boredom emotion. If instead, the output of the affect-neutral level is [*Boredom*, *Neutral*, *Neutral*, *Flow*, *Neutral*], where the *Boredom – Neutral* classifier detects boredom, the *Flow – Neutral* classifier detects flow, and the other affect-neutral classifiers declare neutral, then the *Boredom – Flow* binary discriminator would be recruited to resolve the conflict (see Figure 9).

Such a classification scheme is strongly motivated by the Pandemonium model (Selfridge, 1959). It is expected that in most cases the level 1 classifier (affect-neutral) or a two-way affect classifier would suffice. When more subtle distinctions are required from ambiguous input, a three-way or higher order classifier may also be necessary. However, four-way or five-way discriminations are expected to be much more rare, as discussed later.

## **Measuring the Accuracy of the Posture Based Affect Detector**

The subsequent analyses include a series of classification experiments to evaluate the reliability of affect detection from gross body language.

### *Experimental Setup*

*Data Set Creation.* The data used for the analyses was from the multiple judge study that was described earlier in which 28 participants interacted with AutoTutor on topics in computer literacy. Posture feature vectors for each method (high-level pressure features and spatial-temporal pressure contours) were extracted from the BPMS data. The feature vector was then associated with an emotion category on the basis of each of the four human judges' affect ratings. More specifically, each emotion judgment was temporally bound to each posture based feature vector. This data collection procedure yielded four ground truth models of the learner's affect (self, peer, two trained judges), so four labeled data sets were constructed. When aggregated across each 32-minute session for each of the 28 participants, 2967, 3012, 3816, and 3723 labeled data points for the self, peer, trained judge 1, and trained judge 2, respectively were obtained.

Affect judgment reliabilities between the human judges presented above revealed that the highest agreement was obtained between the trained judges ( $\kappa = .36$ ). However, it is still not firmly established whether the trained judges or the self judgments are closer to ground truth. This issue was addressed by combining affect judgments from the trained judges in order to obtain a better approximation of the learner's emotion. In

particular, an additional data set was constructed on the basis of judgments in which both trained judges agreed; this sample therefore focused on observations in which there is some confidence about the emotion. The frequencies of the emotions in each data set are listed in Table 12.

**Table 12. Frequency of affective states in each data set.**

<b>Affect Judge</b>	<b>Frequency of Affective States</b>								<b>Sum</b>
	<i>Boredom</i>	<i>Confusion</i>	<i>Delight</i>	<i>Flow</i>	<i>Frustration</i>	<i>Neutral</i>	<i>Surprise</i>		
Self	483	533	94	593	335	849	80		2967
Peer	582	555	50	605	207	942	71		3012
Trained Judge 1	379	1151	184	568	291	1192	46		3811
Trained Judge 2	770	1101	120	357	131	1214	30		3723
Trained Agree	268	701	102	224	97	663	17		2072

*Classification Analyses.* The Waikato Environment for Knowledge Analysis (WEKA) (Witten & Frank, 2005) was used to comparatively evaluate the performance of various standard classification techniques ( $N = 17$ ) in detecting affect from posture. The classification algorithms tested were selected from a list of categories including Bayesian classifiers (Naive Bayes and Naive Bayes Updatable), functions (Logistic Regression and Support Vector Machines), instance based techniques (K-Nearest Neighbor with  $k = 1$  and  $k = 3$ , K\*, Locally Weighted Learning), meta classification schemes (AdaBoost, Bagging Predictors, Additive Logistic Regression), trees (C4.5 Decision Trees, Logistic

Model Trees, REP Tree), and rules (Decision Tables, Nearest Neighbor Generalization, PART).

The classification analyses proceeded in two phases. In phase 1 the higher level pressure features ( $N = 16$ ) were inputs to the classifiers. For phase 2, the spatial-temporal features ( $N = 112$ ) were used to detect the affective states. Each phase was independent of the other since the primary goal here was to evaluate the accuracy of each method. Therefore, the accuracy of each of the 17 classifiers in discriminating the affective states grouped in the five categories of the hierarchy was evaluated for each phase (see Figure 9). There were 31 different classification experiments conducted for each feature set. These included five affect-neutral discriminations, ten two-way discriminations, ten three-way discriminations, five four-way discriminations, and a single five-way discrimination.

A uniform baseline for the different emotions was obtained by randomly sampling an equal number of observations from each affective state category. This sampling process was repeated for ten iterations and all reported reliability statistics were averaged across these ten iterations. For example, consider the task of detecting confusion from neutral with affect labels provided by the self. In this case one would randomly select an equal number of confusion and neutral samples, thus creating a data set with equal prior probabilities of both these emotions. Each randomly sampled data set was evaluated on the 17 classification algorithms and reliability statistics were obtained using  $k$ -fold cross-validation ( $k = 10$ ).

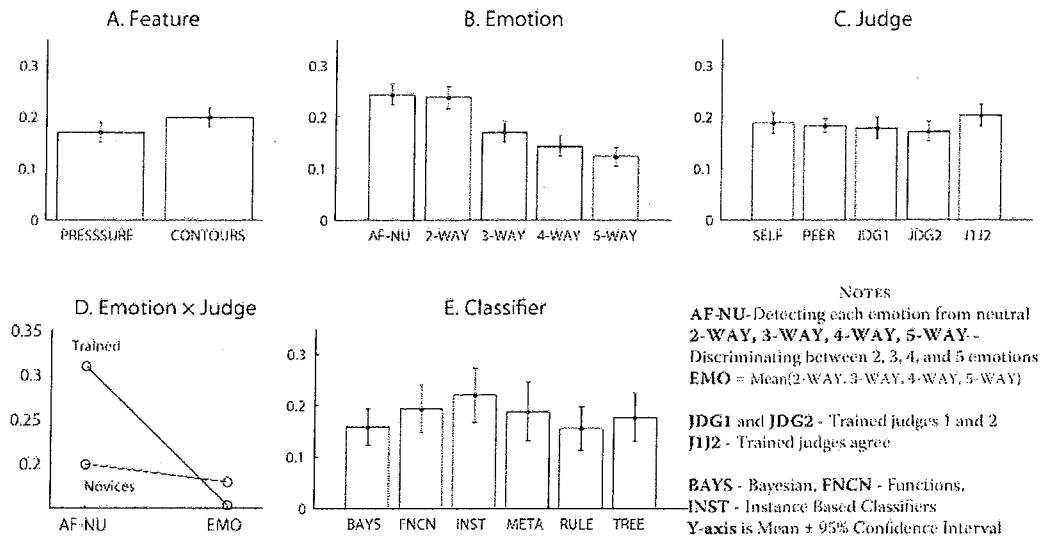
### *Trends in Classification Accuracy*

A three factor repeated measures analysis of variance (ANOVA) was performed in order to comparatively evaluate the performance of the classifiers in detecting affect from the posture features.<sup>2</sup> The first factor (*feature*) was the feature set used as input into the classifier and had two levels: *pressure* and *contours* for the high level pressure features and the spatial-temporal contours respectively. The second factor involved the *emotions* classified, and was composed of five levels: affect-neutral discriminations (chance = 50%), two-way affect discriminations (chance = 50%), three-way affect discriminations (chance = 33%), four-way affect discriminations (chance = 25%), and five-way affect discriminations (chance = 20%). The third factor was the *judge* that provided the affect judgments. This factor also had five levels: self, peer, trained judge 1, trained judge 2, and observations in which trained judges agree. The unit of analysis for the  $2 \times 5 \times 5$  ANOVA was the accuracy obtained by each of the 17 classifiers. The kappa score was utilized as the metric to evaluate performance of each classifier because this metric partials out random guessing.

The ANOVA indicated that there were significant differences in kappa scores across all three factors, as well as for various interactions between the factors. On the basis of the ANOVA comparisons between the various levels of the three factors (feature, emotion, and judge) are reported. Figure 10 graphically depicts the mean kappa score obtained from the emotion classification for each level of each factor of the ANOVA.

---

<sup>2</sup> Please see Appendix C for detailed results of the classification analysis.



**Figure 10. Mean kappa for affect detection.**

(A) Feature type. (B) Emotions classified. (C) Affect judge. (D) Interaction between emotions classified and affect judge. (E) Classification scheme.

*Comparison across Feature Sets.* The results of the ANOVA indicated that there was a statistically significant difference in classification accuracy obtained from each feature set,  $F(1, 16) = 55$ ,  $MSe = .003$ ,  $p < .001$  (partial  $\eta^2 = .775$ ). In particular, the classifiers based on the spatial-temporal contours ( $M_{\text{CONTOUR}} = .20$ ) outperformed those trained on the higher-level pressure features ( $M_{\text{PRESSURE}} = .17$ , see Figure 10A). However, the performance increments attributed to the spatial-temporal contours were marginal (an 18% increase in kappa over high level pressure features). This marginal improvement may be indicative of problems commonly associated with high dimensional feature spaces ( $N = 112$  for spatial-temporal contours) due to cross-correlations among features.

*Comparison across Emotions.* The ANOVA revealed statistically significant differences in kappa scores for the emotions classified,  $F(4, 64) = 269.14$ ,  $MSe = .002$ ,  $p < .001$  (partial  $\eta^2 = .944$ ). Bonferroni post hoc tests revealed that classification accuracy associated with discriminating each emotion from neutral ( $M_{AF-NU} = .243$ ) and two-way classifications ( $M_{2-WAY} = .238$ ) were on par and quantitatively higher than classification accuracy associated with three-way ( $M_{3-WAY} = .177$ ), four-way ( $M_{4-WAY} = .143$ ), and five-way ( $M_{5-WAY} = .123$ ) discriminations (see Figure 10B).

Discriminating a larger number of affective states is challenging, particularly when the states are collected in an ecologically valid setting (i.e., no actors were used to express emotions and no emotions were intentionally induced). As expected, there appears to be a linear relationship between the number of emotions simultaneously being discriminated and the associated classification accuracy score ( $R^2 = .91$ ). It appears that each additional affective state included in the classification model is accompanied by a .04 (kappa) reduction in classification accuracy.

*Comparison Across Affect Judges.* The ANOVA revealed that there were statistically significant differences in kappa scores based on which judge provided the affect ratings used to train and validate the classifiers,  $F(4, 64) = 26.42$ ,  $MSe = .001$ ,  $p < .001$  (partial  $\eta^2 = .623$ ). Bonferroni post hoc tests revealed that classifiers based on affect ratings where the trained judges agreed ( $M_{J1J2} = .203$ ,  $p < .01$ ) yielded the best performance as depicted in Figure 10C.

This finding should be interpreted with some caution since this data set probably consists of some of the more obvious cases, namely since the trained judges were able to agree on an affective state.

Figure 10C indicates that overall classification performance between the self, peer, and two trained judges were on par with each other ( $M_{SELF} = .188$ ,  $M_{PEER} = .183$ ,  $M_{JDG1} = .178$ , and  $M_{JDG2} = .172$ ). However, interesting patterns appear when one considers interactions between the affect judge and the emotions classified (see Figure 10D), which was statistically significant,  $F(16, 256) = 167.76$ ,  $MSe = 0$ ,  $p < .001$  (partial  $\eta^2 = .913$ ). When one considers simple affect-neutral distinctions, it appears that classifiers trained on data sets in which affect judgments were provided by the novice judges (self and peer,  $M_{NOVICES} = .309$ ) were much higher than classifiers based on affect judgments from the trained judges ( $M_{TRAINED} = .199$ ). However, a reverse pattern was discovered for more complex discriminations between the various emotions (obtained by averaging accuracy scored for two-way, three-way, four-way, and five-way classifications). These classifiers were best for the trained judges ( $M_{TRAINED} = .179$ ) compared with the novices ( $M_{NOVICES} = .153$ ). This suggests that the novices were more adept at making affect-neutral distinctions whereas the trained judges are more capable at performing complex emotion discriminations. Perhaps this phenomenon may be explained by the fact that the trained judges had considerable experience interacting with AutoTutor, and they make use of this contextual knowledge coupled with their facial expression training to discriminate between the affective states.

*Comparisons across Classifier Schemes.* A separate repeated measures ANOVA was performed to determine whether there was any significant differences among the various classifier schemes described above. This analysis had two factors: *feature* and *classifier*. Similar to the three-way ANOVA described above, the first factor (*feature*) was the feature set used as input to the classifier and had two levels: *pressure* and *contours*. The second factor of the ANOVA was the classification scheme (called *classifier*) divided across six levels for Bayesian classifiers, functions, instance based learners, meta classifiers, rules, and trees. The unit of analysis for this  $2 \times 6$  ANOVA was kappa scores associated with each of the affective models (affect-neutral, two-, three-, four-, and five-way classifications).

As expected from the previous analyses, there was a statistically significant difference among the two feature sets used with the spatial-temporal contours feature set outperforming the high level pressure feature set. There were also significant differences in the kappa scores across the various classifier schemes  $F(5, 20) = 34.81$ ,  $MSe = .006$ ,  $p < .001$  (partial  $\eta^2 = .807$ , see Figure 10E). Bonferroni post hoc tests revealed that the kappa scores of the instance based classifiers ( $M_{INST} = .22$ ) were significantly higher than the others. Performance of the function-based classifiers, meta classifiers, and trees ( $M_{FNCN} = .193$ ,  $M_{META} = .188$ ,  $M_{TREE} = .155$ ) were similar quantitatively and higher than Bayesian classifiers and rule-based learning schemes ( $M_{BAYS} = .158$ ,  $M_{RULE} = .177$ ).

It is informative to note that the results showed no statistically significant interactions between the feature set and the classification schemes ( $p = .113$ ). This result indicates that the relative performance of the six classification schemes derived above is independent of the feature set (pressure or contours) (see Table 13).

#### *Maximum Classification Accuracy*

The use of multiple assessments of the learner's affect ( $N = 5$ ) and a large number of classifiers ( $N = 17$ ) was useful to investigate the effect of different factors (feature set, affect judge, etc.) on affect detection accuracy. However, in order to achieve the goal of developing a real time emotion classification system, the discussion will shift to the classifier that yielded the best performance. Table 13 presents the maximum classification accuracies obtained across all 17 classifiers across the five data sets in discriminating the various combinations of affective states specified by the hierarchy (see Figure 9).

**Table 13. Maximum classification accuracy in detecting affect.**

Model	Affective States	Classification Accuracy (%)				Baseline (%)
		Max Pressure	Contours	Mean Pressure	Contours	
Affect-Neutral Detection	BO-NU	70.70	73.05	72	74.2	50
	CF-NU	68.50	71.65			
	DL-NU	70.55	70.30			
	FL-NU	78.70	82.45			
	FR-NU	71.65	73.70			
2-Way Affect Classification	BO-CF	67.50	70.15	69.4	71.2	50
	BO-DL	72.70	74.80			
	BO-FL	76.20	78.25			
	BO-FR	68.70	68.00			
	CF-DL	67.80	64.00			
	CF-FL	68.20	71.65			
	CF-FR	64.05	66.00			
	DL-FL	72.40	78.25			
	DL-FR	66.75	69.85			
	FL-FR	70.10	70.50			
3-Way Affect Classification	BO-CF-DL	51.63	52.37	52.5	55.2	33.33
	BO-CF-FR	50.69	51.49			
	BO-CF-FL	54.84	58.80			
	BO-DL-FR	53.03	55.18			
	BO-DL-FL	56.58	64.09			
	CF-DL-FR	48.14	49.15			
	CF-DL-FL	53.77	54.57			
	CF-FL-FR	50.35	52.50			
	BO-FL-FR	53.44	55.31			
	DL-FL-FR	52.03	58.19			
4-Way Affect Classification	BO-CF-DL-FR	41.35	42.33	42	45.9	25
	BO-CF-DL-FL	42.10	47.88			
	BO-CF-FL-FR	41.80	44.88			
	BO-DL-FL-FR	45.03	49.45			
	CF-DL-FL-FR	39.85	45.18			
5-Way Affect Classification	BO-CF-DL-FLFR	36.00	39.20	36	39.2	20

*Notes. Pressure – High level pressure features, Contours – Spatial-Temporal Pressure Contours. BO – Boredom, CF- Confusion, DL – Delight, FL- Flow, FR-Frustration, NU - Neutral*

The results revealed that the accuracy for affect-neutral discrimination and two-way emotion resolutions are reasonable (74% and 71%, respectively), but the accuracy drops when the system attempts to resolve conflicts between three or more emotions (3-way = 55%, 4-way = 46%, 5-way = 39%). Therefore, it appears that the efficacy of the hierarchical classification scheme is inversely proportional to the probability of requiring the higher order emotion classification models to resolve discrepancies that arise during the affect-neutral discrimination phase. Simply put, the hierarchical method for affect detection would be feasible if affect-neutral, two-way, and the occasional three-way classifications would suffice.

There is some evidence that suggests that human disagreements among the affective states usually occur at the affect-neutral stage or the two-way classification stage. An analysis on the source of classification errors made by the human judges (self, peer, judge1, and judge 2) revealed that 63.5% of the time each emotion was confused with neutral (affect-neutral detection). The two-way conflicts occurred 30.4% of the time, while three-way conflicts were much rarer (5.5%). The four-way and five-way discriminations almost never occurred. Taken together, the ideal model for emotion classification would involve (a) the detection of single emotions compared to neutral states, a resonance-based process that fits the Pandemonium model very well, and (b) the resolution of pairs of emotions that have some modicum of activation.

## **General Discussion**

This research was motivated by the belief that the affective states of learners are manifested through their gross body language via configurations of body position and modulations of arousal. Several milestones that suggest that significant information can be obtained from body position and movement have been achieved. Although, gross body language is rarely used for affect detection, the results indicate that the characteristics of the body posture are quite diagnostic of the affect states of learners. On the basis of two sets of body pressure features alone, the results showed that conventional classifiers are moderately successful in discriminating the affective states of boredom, confusion, delight, flow, and frustration from each other, as well as from the baseline state of neutral.

This chapter introduced a two-step affect detection model where affect-neutral classifiers first determined whether the incoming pressure maps resonated with any one or more of the emotions (versus a neutral state). If there is resonance with only one emotion, then that emotion would be declared as being experienced by the learner. In situations where there is resonance with two or more emotions, additional two-, three-, four-, or five-way conflict resolution modules are recruited.

Comprehensive evaluations of this model were not presented in this chapter because the focus was primarily on exploring the potential of a posture based affect-detection. However, initial analyses with this model revealed that classification accuracy scores were notably lower for four- and five-way emotion classifications than for affect-neutral, two-way, and three-way emotion discriminations.

This research provides an alternative to the long standing notion that extols the virtues of the face as the primary modality through which emotion is communicated. An important hope is to have established a foundation for the use of gross body language as a serious contender to traditional measures for emotion detection such as facial feature tracking and monitoring the paralinguistic features of speech. Although the face might reign supreme in the communication of the basic emotions (i.e., anger, fear, sadness, enjoyment, disgust, and surprise, Ekman & Friesen, 1978), the results clearly indicate that there is the possibility of inferring emotions by tracking gross body language. Furthermore, the face can be quite deceptive when learners' attempt to disguise negative emotions such as frustration. But bodily motions are ordinarily unconscious, unintentional, and thereby not susceptible to social editing. These factors make an ideal channel for non-intrusive affect monitoring.

## **Chapter 5: Automated Affect Detection by Combining Sensors**

### **Introduction**

Multimodal systems for affect detection have been widely discussed but rarely implemented (Jaimes & Sebe, 2007). This is mainly due to the inherent challenges with unisensory affect detection, which undoubtedly increase in multisensory environments. Therefore, much of the known research in affect detection has involved the use of a single modality to infer affect. Some notable exceptions involve the use of four physiological signals to detect eight basic emotions (Picard et al., 2001) and various combinations of audio-visual features (Chen, Huang, Miyasato, & Nakatsu, 1998; Dasarathy, 1997; Yoshitomi, K., Kawano, & Kilazoe, 2000). Another popular approach involves a combination of acoustic-prosodic, lexical, and discourse features for affect detection (Ang et al., 2002; Lee & Narayanan, 2005; Liscombe et al., 2005; Litman & Forbes-Riley, 2004).

More recently, Scherer and Ellgring (2007) considered the possibility of combining facial and vocal features to discriminate among 14 emotions (e.g., hot anger, shame, etc. Base rate = 1/14 = 7.14%). Single-channel classification accuracies from 21 facial features and 16 acoustic parameters were 52.2% and 52.5%, respectively (accuracy

rates for gesture and body movements were not provided). A combined 37-channel model yielded accuracy rates of 79%, but the accuracy dropped to 54% when only 10 of the most diagnostic features were included in the model. It is difficult to assess whether the combined models led to enhanced or equivalent classification scores compared to the single-channel models because the number of channels between model types were not matched. Furthermore, the study used context-free, acted, emotional expressions, while the focus is on naturalistic emotional expressions that are grounded in learning contexts.

Of more relevance to the context of learning, Kapoor and Picard (2005) developed a 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). 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.

Chapters 2 and 3 examined the accuracy by which learners' emotions can be detected by monitoring conversational cues and gross body language. This chapter considers the possibility of classifying emotions on the basis of a combination of sensors. Chapter 5 also considers facial features for affect detection. Although an important goal of this dissertation is on developing alternate techniques for affect detection, in contrast to existing methods that primarily rely on the face, there are two reasons to consider facial features. First, the problem space is significantly broadened when three (instead of

two) channels are considered. For example, it might be the case that a combination of dialogue and posture does not resonate with enhanced classification accuracy. But a combination of dialogue + face, or posture + face, or dialogue + posture + face might produce classification accuracies that are superior to what can be obtained by any one channel alone. A data set that includes facial features allows us to explore the potential of these additional models.

Second, the inclusion of facial features permits a direct comparison between the new methods of monitoring dialogue and posture with traditional methods that typically rely on the face. It might be the case that the face always dominates dialogue and posture. Alternatively, the face might be superior in some contexts, but the other channels are more accurate in other situations. Uncovering the conditions under which a particular channel provides the most reliable affect classification advances both basic and applied research goals.

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 chapter considers both feature-level and decision-level fusion algorithms in developing a composite (or multi-channel) affect detector.

The fact that more than one affect sensor will be monitored raises the issue of how the results should be evaluated. 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.

In summary, there were three major goals to this research: (1) to develop mechanisms for feature-level and decision-level fusion and to compare between these two fusion techniques, (2) to compare the efficacy of dialogue and posture to facial feature tracking, and (3) to assess whether the combination of conversational cues, gross body language, and facial features resulted in superadditive, additive, redundant, or inhibitory effects.

## Mechanisms for Feature and Decision-Level Fusion

### *Feature-Level Fusion*

A naïve feature-level fusion algorithm in which a multi-channel feature vector was created by appending features from each individual channel was considered for feature-level fusion. One problem with this approach is that the number of features in the combined model increases as additional channels are considered. Therefore, the number of estimated parameters is quite different in single versus multiple channels. This makes it difficult to defend comparisons of the classification accuracy of the combined model with the single-channel models. Furthermore, data sparsity increases as the number of features increases, so performance is compromised by virtue of the *curse of dimensionality* (Bellmann, 1961).

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:  $f_c = \min(f_1, f_2, \dots, f_m)$ . In a uniform distribution, the number of features that each channel contributes is:  $f_{equal} = \lfloor f_c/m \rfloor$ .

As an example, consider a situation where  $f_{face} = 17$ ,  $f_{dialogue} = 10$ , and  $f_{posture} = 9$ . For the face + dialogue model,  $m = 2$ ,  $f_c = \min(17, 10) = 10$ , and  $f_{equal} = \lfloor 10/2 \rfloor = 5$ . For the combined face + dialogue + posture model,  $m = 3$ ,  $f_c = \min(17, 10, 9) = 9$ , and  $f_{equal} = \lfloor 9/3 \rfloor = 3$ .

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, which incorporates a combination of forward and backward variable entry techniques (Hocking, 1976). Entropy reduction and information gain feature selection techniques can also be considered (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 differed across the different emotions and thereby was capable of discriminating emotions.

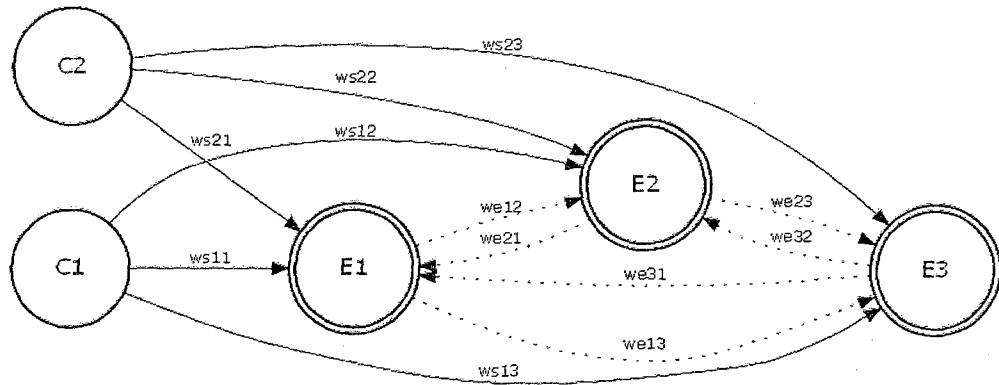
#### *Decision-Level Fusion*

A spreading activation network with *projecting* and *lateral* links was used to model decision-level fusion (Rumelhart et al., 1986). A sample network is presented in Figure

11. This hypothetical network has two sensor nodes,  $S1$  and  $S2$ , and three emotion nodes,  $E1$ ,  $E2$ , and  $E3$ . Each sensor is connected to each emotion by a projecting link (solid lines in Figure 11). The degree to which a particular sensor activates a particular emotion is based on the accuracy by which the sensor has detected the emotion in the past (see  $ws$  weights in Figure 11). So if the BPMS is more accurate at detecting boredom than confusion, it will excite the boredom node more than the confusion node, even if its current estimates on the probability of both emotions are approximately equivalent.

Each emotion is also connected to every other emotion with a *lateral* link (dotted lines in Figure 11). These links are weighted and can be excitatory or inhibitory (see  $we$  weights in Figure 11). Related emotions excite each other while unrelated emotions inhibit each other. For example, confusion would excite frustration but boredom would inhibit engagement.

Each emotion node receives activation from both link types and maintains an *activation value*. At any time, the emotion node with the highest activation value is considered to be the emotion that the learner is currently experiencing.



**Figure 11.** Sample activation spreading network for decision-level fusion.

The decision-level fusion algorithm operates in four phases.

1. *Detection by Sensors.* Each sensor provides an independent estimate of the likelihood that the learner is experiencing an emotion. The likelihood can be represented as a probability value for each emotion.
2. *Activation from Sensors.* Sensors spread activation and emotion nodes aggregate this activation.
3. *Activation from Emotions.* Each emotion spreads the activation received from the sensors to the other emotions, so that some emotions are excited while others are inhibited.
4. *Decision.* The emotion with the highest activation is selected to be the emotion that the learner is currently experiencing.

*Mathematical Model.* Assume that a set of  $m$  sensors are monitoring a set of  $n$  emotions. Such a network would contain  $m + n$  nodes, (i.e., a node for each sensor and a

node for each emotion), and  $(m \times n) + n(n - 1)$  links. Here,  $(m \times n)$  represents the number of projecting links between each sensor and each emotion, whereas  $n(n - 1)$  is the number of lateral links between the emotions (an emotion is connected to every other emotion excluding itself).

Each emotion node has an activation value  $e_j$  that represents the degree to which this emotion is activated at time  $t$ . Each emotion node receives activation from its lateral as well as its projecting links. Let:  $a_{ij}^s$  be the activation from sensor node  $i$  on emotion  $j$ , and  $a_{jk}^e$  be the activation on  $j$  from another emotion node  $k$  ( $j \neq k$ ). Summing up the two sources of activation, the total activation of emotion  $j$  is:

$$e_j = \sum_{i=1}^m a_{ij}^s + \sum_{\substack{k=1 \\ k \neq j}}^n a_{kj}^e \quad \text{Eq. 8}$$

The links between each sensor node and emotion are weighted. Let  $w_{ij}^s$  be the weight between sensor node  $i$  and emotion node  $j$ . If  $s_{ij}$  is the probability by which sensor  $i$  detects emotion  $j$ , then  $a_{ij}^s$  can be specified as:

$$a_{ij}^s = w_{ij}^s \times s_{ij} \quad \text{Eq. 9}$$

The links between the various emotion nodes are also weighted. Let  $w_{kj}^e$  be the weight between emotion node  $j$  and emotion node  $k$  ( $j \neq k$ ). Now if  $e_k$  is the activation of emotion  $k$ , then the lateral activation that emotion  $k$  spreads to emotion  $j$ , or  $a_{jk}^e$  is:

$$a_{kj}^e = w_{kj}^e \times e_k \quad \text{Eq. 10}$$

Substituting Eq. 9 and Eq. 10 in Eq. 8 yields:

$$e_j = \sum_{i=1}^m w_{ij}^s \times s_{ij} + \sum_{\substack{k=1 \\ k \neq j}}^n w_{kj}^e \times e_k \quad \text{Eq. 11}$$

The decision-level framework described above assumes that the sensors have equal sampling rates, in the sense that they activate emotion nodes at approximately equal intervals. Complications occur because this assumption is routinely violated in practical applications. For example, the BPMS sensor requires about 3–4 seconds of data to detect an emotion. On the other hand, data from the dialogue is only available every 20–40 seconds. In this situation, the activation values of the BPMS will inhibit the activation values of the dialogue sensor because the BPMS is being updated more frequently.

This problem can be corrected by introducing a parameter  $r_i$ , which is the sampling rate of sensor  $i$ . Dividing the activation received from each sensor by its

sampling rate adjusts for any biases caused by heterogeneous sampling rates. The corrected values for  $a_{ij}^s$  and  $e_j$  are specified in Eq. 12 and Eq. 13, respectively.

$$a_{ij}^s = \frac{w_{ij}^s \times s_{ij}}{r_i} \quad \text{Eq. 12}$$

$$e_j = \sum_{i=1}^m \frac{w_{ij}^s \times s_{ij}}{r_i} + \sum_{\substack{k=1 \\ k \neq j}}^n w_{kj}^e \times e_k \quad \text{Eq. 13}$$

#### *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. The kappa statistic ( $k$ ) and its associated standard error,  $se(k)$ , can be computed from the confusion matrix that is obtained when a model is validated (J. Cohen, 1960). An important property of the kappa statistic is that a z score can be computed according to Eq. 14, which asymptotically approximates a normal distribution (Fleiss, 1981).

$$z = \frac{k}{se(k)}$$

Eq. 14

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. 15). 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  $k_2$ .

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

Eq. 15

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*. A threshold for superadditivity ( $s_{1+2}$ ) that considers an additive combination of effects is presented in Eq. 16.

$$s_{1+2} = k_1 + k_2 - k_1 \times k_2$$

Eq. 16

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. 16 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:

$$s_{1+2+3} = k_1 + k_2 + k_3 - k_1 k_2 - k_1 k_3 \\ - k_2 k_3 + k_1 k_2 k_3 \quad \text{Eq. 17}$$

It should be noted that the aforementioned criterion for evaluating superadditivity is somewhat stringent. An alternate more lenient metric considers the incremental gain from multisensory fusion over and above the maximum unisensory response (see Eq. 18). This metric is widely used by neuroscientists studying multisensory integration with respect to visual, audio, and tactile senses in humans and animals (Holmes & Spence, 2005). It should also be noted that a degree of stringent superadditivity can also be

computed by substituting  $s_{1+2}$  or  $s_{1+2+3}$  for  $\max(k_1 - k_2)$  in Eq. 18. Both metrics are used in this chapter, although the analyses primarily rely on the more stringent metric.

$$superadditivity = \frac{k_{1+2} - \max(k_1 - k_2)}{\max(k_1 - k_2)}$$
 Eq. 18

## Data Sets

The current analyses included data sets from the *Multiple Judge Study* described in Chapter 2. In this study, judgments for each learner's affective states were provided by the learner themselves (self), an untrained peer, and two trained judges (Graesser et al., 2006). There were two different judgment points for each AutoTutor session. Mandatory judgments were made at regular 20-second intervals, whereas voluntary judgments were obtained at any time between the 20-second points. There are two important differences between these two judgment types. First a different distribution of emotions was elicited for each judgment type. For the mandatory judgments, the most common affective state was neutral (.37), followed by confusion (.21), flow (.19), and boredom (.17); the remaining states of delight, frustration, and surprise totaled .06 of the observations. The voluntary points had a rather different distribution. The most prominent affect state was confusion (.37), followed by frustration (.19), delight (.17), and boredom (.10), whereas the remaining affective states comprised .17 of the observations. So the subtle states of flow and boredom are more frequently observed at the mandatory points when compared

to the more obvious states of frustration and delight which are routinely observed at the voluntary 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 (mandatory points), rather than being able to stop a video display to make spontaneous judgments (see Chapter 2). The states at regular timestamps are much less salient so there is minimal information to base their judgments, compared with those points when affective states are detected by the judge. Furthermore, there is a reason to suspect that the fidelity of the different data channels varies as a function of the judgment type. Since the voluntary judgment points consist of the more obvious cases of emotion expressions, it is quite plausible that the face plays a more salient role for these points. On the other hand, it is quite possible that the context (i.e., dialogue features) is the most reliable channel for the mandatory judgment points. Separate data sets for the mandatory and voluntary 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 mandatory judgment data set consisted of boredom, confusion, flow, and neutral, while boredom, confusion, delight, frustration, and neutral were included in the voluntary data set.

As described in Chapter 2, automated systems were used to monitor gross body language and conversational cues. However, there were some technical challenges associated with the automated detection of facial expressions. As an initial step, two trained judges coded a sample of the observations of emotions on facial action units.

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, instead of considering affect judgments provided by all four judges, there was a focus on points where the two trained judges agreed on the learners' emotions. This approach had the advantage that kappa scores between the trained judges was superior to the self and peer judgments. The increased kappa scores made us more confident in the validity of the emotion judgments.

#### *Data Sampling*

*Mandatory Data Set.* Each trained judge provided 96 mandatory 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. Specifically, an approximately equal number of observations was obtained from each participant and for each of the affective states of boredom, confusion, flow, and neutral. There were 85 instances of boredom, 80 of confusion, 74 of flow, and 78 of neutral.

*Voluntary Data Set.* There were only 1133 points of voluntary observations. A subset of the voluntary points was identified by only a single judge (self, peer, or one of the trained judges). This problem was mitigated by an exhaustive voluntary affect judgment session in which the trained judges repeated the judgment procedure with the

added requirement that they had to provide affect ratings on all 1133 voluntary emotion observations (see Chapter 2). The two trained judges were found to agree 64% of the time ( $N = 720$ ), yielding a kappa score of 0.49. This kappa score is higher than that achieved for the mandatory observations (0.31) but substantially lower when compared to the purely voluntary observations (0.71). However, it is on par with reliability scores reported by other researchers who have assessed identification of emotions by humans (see Chapter 2).

A subset of 407 emotion episodes were randomly sampled from the 720 data points obtained from the exhaustive voluntary affect judgment procedure. An approximate 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.

#### *Extracting Features from Sensors*

Due to the relatively small number of samples in each data set, a subset of features from each sensor was considered. These included features that were the most diagnostic of the learners' emotions (Craig et al., 2008; D'Mello, Craig et al., 2008; D'Mello & Graesser, 2009) and are briefly discussed below.

*Conversational Cues (Dialogue Features).* There were nine dialogue features in all. These features included temporal assessments for each student–tutor turn, such as the *Subtopic Number*, the *Turn Number* within a subtopic, and the student's *Reaction Time* (interval between presentation of the question and the onset of the submission of the

student's answer). Assessments of response verbosity included the *Number Characters* (letters, numbers) and *Speech Act* (that is, whether the student's speech act 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, Penumatsa et al., 2007; Landauer & Dumais, 1997; Landauer et al., 2008). LSA-based measures included a *Local Good Score* (the conceptual similarity between the student's current response and the set of expectations being covered) and a *Global Good Score* (the similarity of a set of student responses to a set of expectations in a good answer). 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). Chapter 3 provides a detailed description of each feature.

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

*Gross Body Language (Posture Features)*. There were four pressure-related features and one feature related to the pressure coverage for both the back and the seat, yielding ten features in all. Each of the features was computed by examining the pressure

map during the emotional episodes for the mandatory and voluntary data sets (called the *current frame*). The pressure-related features include the *Average Pressure*, which measures the average pressure exerted. The *Prior Change* and *Post Change* scores measure the difference between the net pressure in the current frame and the frame three seconds earlier and later, respectively. Finally, the *Average Pressure Change* measures the mean change in the net pressure across a predefined window, typically four seconds, that spans two seconds before and two seconds after an emotion judgment. The coverage feature examined the proportion of non-negative sensing units (*Average Coverage*).

Chapter 4 describes each feature in more detail.

*Facial Action Units (Facial Features)*. Ekman and Friesen (1978) highlighted the expressive aspects of emotions with their Facial Action Coding System. This system specifies how “basic” emotions can be identified on the basis of facial behaviors and the muscles that produce them. Each movement in the face is called an *action unit* (or AU). There are 58 action units in all. These prototypical facial patterns have been used to identify the six basic emotions: happiness, sadness, surprise, disgust, anger, and fear. The development of a system that automatically detects the action units is quite a challenging task, however, because the coding system was originally created for static pictures rather than on changing expressions over time.

As an initial step, two trained raters coded a sample of the observations of emotions on the action units. Since the expression of emotions tends to be very fast and only last for about three seconds (Ekman, 2003), two raters independently scored the three seconds before the emotion expression using the Facial Action Coding System

(Ekman & Friesen, 1978). Therefore, for the three seconds, raters watched the videos and recorded visible AUs and the time of observation. The raters' coding of AUs was used to create an action unit database. Each record in the database consisted of one or more AUs from the same clip based on the time stamp in which they were observed. This allowed for multiple action unit records for the same emotion.

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, D'Mello 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 14. 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 ( $M_{\text{mandatory}} = .624$ ,  $M_{\text{voluntary}} = .733$ ) (Robson, 1993).

**Table 14. Description of action units, incidence of occurrence, and kappa scores.**

<b>Facial Action Unit</b>	<b>Mandatory</b>			<b>Voluntary</b>		
		<i>Proportion</i>	<i>Kappa</i>		<i>Proportion</i>	
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	- <sup>a</sup>	.071	.605
	AU45	Blink	.455	.745	.234	.681
Lower Face	AU12	Lip Corner Puller	.054	.520	.112	.707
	AU14	Dimpler	.064	.394	.041	- <sup>a</sup>
Lip Parting/	AU25	Lips Part	.114	.742	.134	.912
Jaw Opening	AU26	Jaw Drop	.030	.452	.098	.851
Eye Position	AU64	Eyes Down	.165	.736	.049	.833

*Note.* <sup>a</sup> Kappa could not be computed as only one judge observed this action unit.

## Feature-Level Fusion Models

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.

When each channel was considered independently, there were 10 predictors for the *F* and *P* models, and nine for the *D* model. 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 mentioned above, this increase in the size of the feature sets of the multi-channel models causes some difficulty in interpreting the effects of the models. In particular, it is inappropriate to directly compare a multi-channel model with a single-channel model when the multi-channel models have twice or thrice the number of features than the single-channel models. Any effect in favor of the multi-channel models could be attributed to a sheer increase in the size of the features and not to any real additive effects of the different channels.

To alleviate this concern, the feature selection mechanism described above was implemented to ensure that the same number of features was used in the one-channel, two-channel, and three-channel models. 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. There is the obvious question of which features from the single-channel models should be retained in the multi-channel models.

and which should be discarded. The F-ratio from a univariate ANOVA that tested each feature to determine if its mean differed across the different emotions was used for feature selection. 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 15.

As an example consider the process of constructing the *FD* model for mandatory 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<sup>1</sup>.

---

<sup>1</sup> It should be noted that preliminary tests comparing multi-channel 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.

**Table 15. Features included in the various classification models.**

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

*Notes.* x 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.

Linear discriminant analyses (LDA)<sup>2</sup> were conducted on the mandatory and voluntary data sets (Klecka, 1980). The LDA analyses for the mandatory data set performed a four-way discrimination between boredom, confusion, flow, and neutral, while a five-way discrimination (boredom, confusion, delight, frustration, and neutral) was considered for the voluntary data set. There were seven analyses for each dataset, as specified above. 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.

#### *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 multi-channel models. Kappa scores for the mandatory 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<sup>3</sup> greater than zero, but the kappa score for the face was statistically indistinguishable from

---

<sup>2</sup> A quadratic discriminant analysis yielded accuracies similar to linear discriminant functions.

<sup>3</sup>  $p < .001$  unless specified otherwise.

zero ( $p = .07$ ), which is consistent with random guessing. So quite clearly, the face does not provide sufficient cues to discriminate the subtle emotional expressions during these 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 mandatory data set. The posture features were about half as reliable as the dialogue features.

A rather different pattern of results was found for the voluntary 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 abysmal for the mandatory judgments, it was spectacular for the voluntary judgments, at least when compared to the other channels.

These results are significant because they challenge the importance of the face as the primary communicative channel for emotional expressions. The face can be quite diagnostic of emotions when the expressions are accompanied by heightened activity in the face that can be visibly detected by the judges (i.e., voluntary 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 mandatory 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 much touted, 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 (mandatory, voluntary). The next step is to determine whether fusing data from multiple channels yields classification accuracies that are superior to the best individual channel.

### *Multi-Channel Models*

Results of the mandatory models are presented in Figure 12A. Consider first the two-channel models. The kappa scores for these models were statistically indistinguishable ( $\kappa_{FD} = .271$ ,  $\kappa_{FP} = .205$ , and  $\kappa_{DP} = .242$ ). 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 ( $[FD = D] > F^4$  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 ( $[DF = DP = D] > P > F^5$ ).

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 ( $\kappa_F = -.058$ ,  $\kappa_P = .107$ ,  $\kappa_{FP} = .205$ ). Furthermore, the kappa for the

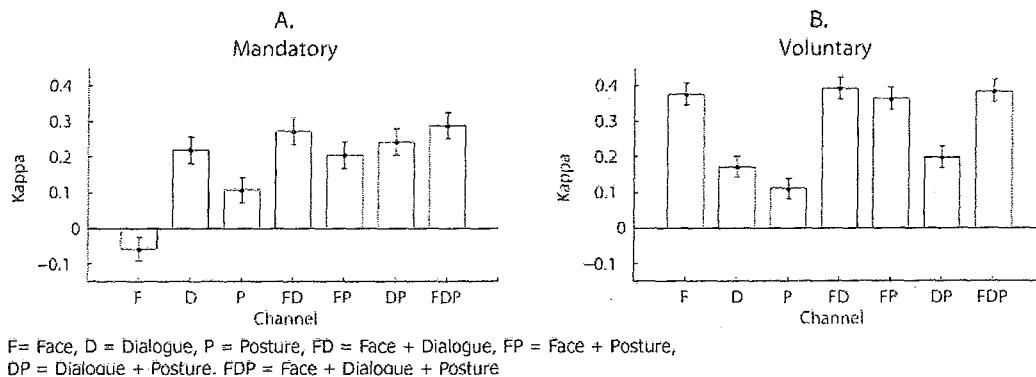
---

<sup>4</sup>  $[A = B] > C$  implies that A and B are statistically equivalent and are significantly greater than C.

<sup>5</sup> DF, DP, and D are statistically equivalent. They are statistically greater than P and F. P is greater than F.

FP model clearly exceeds the superadditivity threshold for two-channel models (see Eq. 16), 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, and marginally greater than the D model ( $p = .068$ ). But does this model resonate with superadditivity? According to Eq. 17, the superadditivity threshold for this three-channel model is .263. So it appears that combining features from the face, dialogue, and posture resonates with effects that surpass an additive combination of the individual channels.



**Figure 12. Overall classification results for feature-level fusion.**

Results of voluntary models are presented in Figure 12B. Recall that the face was the dominant channel for the voluntary 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 ( $\kappa_{FD} = .391$ ,  $\kappa_{FP} = .361$ ) and

significantly higher than the model without the face ( $\kappa_{DP} = .198$ ). The following pattern of results is obtained when 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 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 mandatory models, superadditivity effects were not discovered for the voluntary models.

#### *Discrepancy Reduction*

The results suggest that the composite models have no merits for the voluntary 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.

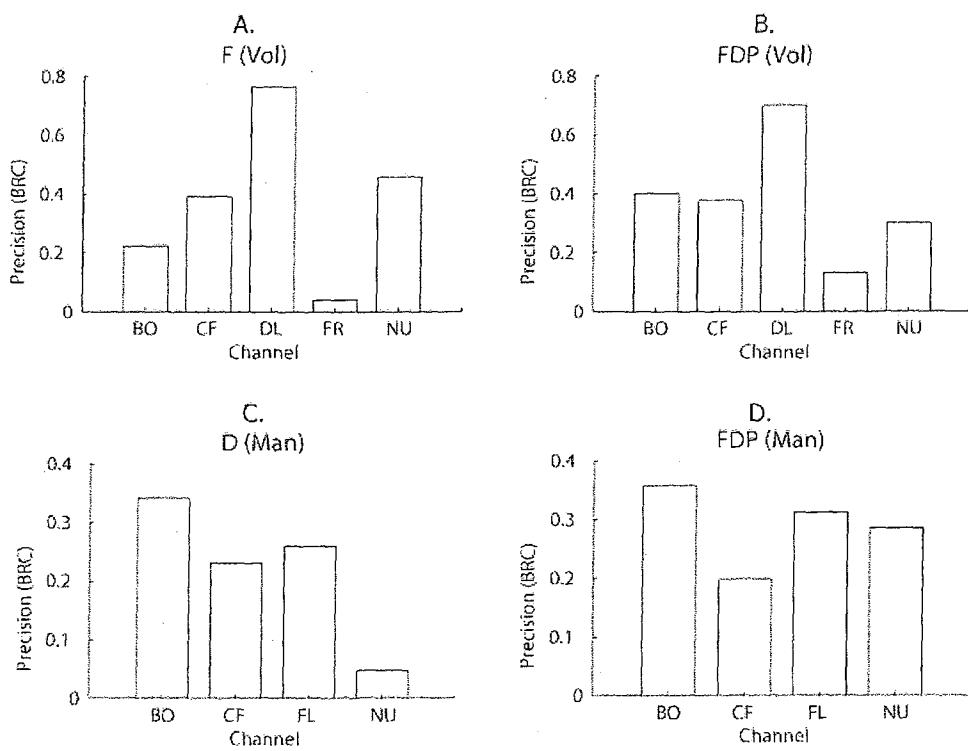
Figure 13 offers a comparative view of the precision scores for the best single model (dialogue for mandatory and face for voluntary) along with the associated three-channel models. The precision scores have been corrected for base rate biases as there

were minor differences in the distributions of the various emotions.<sup>6</sup> According to Figure 13A, the face was very successful in detecting delight for the voluntary judgments, but was abysmal for frustration. In fact the precision for frustration was approximately zero. An interesting phenomenon occurs when the composite model was considered (Figure 13B). 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 mandatory judgments. In this case, the precision of boredom, confusion, and flow were approximately equivalent for the single channel models, but the precision of neutral was much lower (Figure 13C). The precision of neutral substantially improved for the composite model (Figure 13D), at the expense of a small negative shift in the precision of confusion.

---

<sup>6</sup>  $Precision(BRC) = \frac{Observed\ Precision - Expected\ Precision}{1 - Expected\ Precision}$ , BRC = Base rate corrected.



F = Face, D = Dialogue, FDP = Face + Dialogue + Posture  
 BO = Boredom, CF = Confusion, DL = Delight, FL = Flow, FR = Frustration, NU = Neutral

**Figure 13. Single versus multi channel precision scores for individual emotions.**

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 in 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.

An obvious metric to quantify the degree of discrepancy of a model is the population variance of the precision scores for the individual emotions. The variance for the F and FDP voluntary models were .058 and .030, 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 mandatory 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.

#### *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 voluntary judgments) suggests that they 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 16). 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 mandatory and voluntary 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 mandatory judgments and the F, FD, and FDP models for the

voluntary 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.

*Boredom.* It appears that the two-channel FD model had no enhanced effect for mandatory boredom judgments (see Figure 14A). However, the FD model did yield superadditive effects for the voluntary data set. According to Eq. 18, 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 voluntary judgments, compared to the situation when the discourse context is considered.

Whereas the D mandatory model was better than the F voluntary model, the FD and FDP models were approximately equivalent. So the addition of D features to the F voluntary model allows precision scores for voluntary boredom to match the precision of mandatory 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.

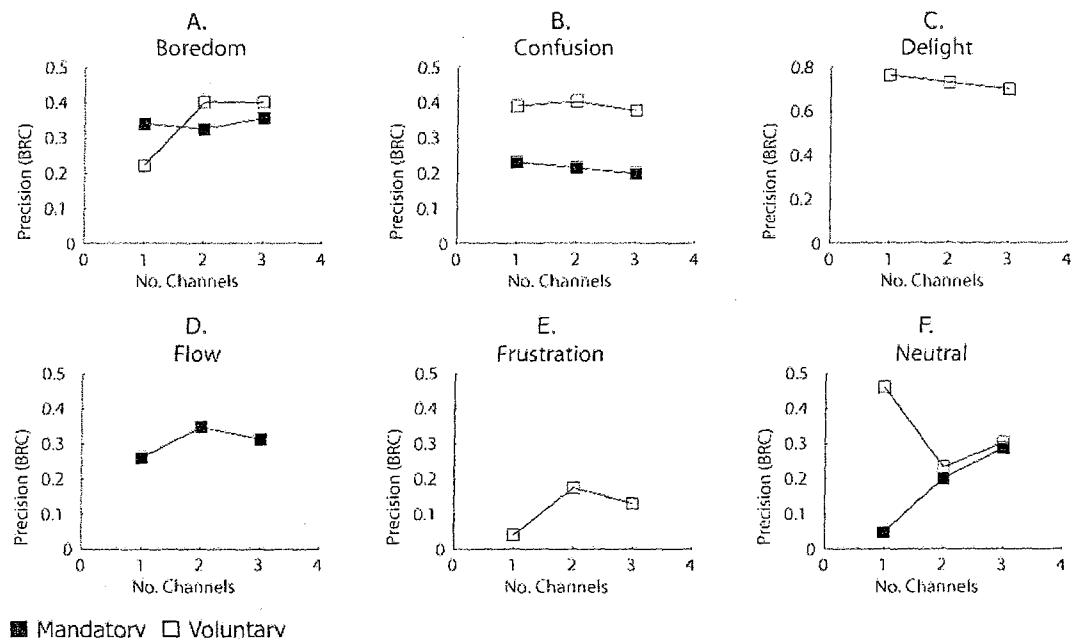
**Table 16. Base rate corrected precision scores for emotions.**

Channel	Boredom	Confusion	Delight	Flow	Frustration	Neutral
<b>Mandatory</b>						
F	-.205	.064	-	.066	-	-.156
D	.341	.231	-	.260	-	.048
P	.132	.064	-	.207	-	.031
FD	.325	.215	-	.348	-	.201
FP	.180	.181	-	.295	-	.167
DP	.357	.248	-	.295	-	.065
FDP	.357	.198	-	.313	-	.286
<b>Voluntary</b>						
F	.223	.390	.762	-	.040	.459
D	.304	.111	.048	-	.565	-.170
P	-.036	.443	.127	-	-.110	.127
FD	.401	.403	.730	-	.175	.232
FP	.272	.390	.746	-	.055	.336
DP	.353	.257	.064	-	.370	-.048
FDP	.401	.377	.699	-	.130	.302

*Confusion.* The face was a good indicator of confusion for the voluntary points (.390) but not the mandatory points (.064). A reverse effect was observed for contextual discourse cues. Here, the dialogue was diagnostic of confusion for the mandatory points (.231) but not the voluntary points (.111). Although it is reasonable to expect that adding these two channels would result in superadditive effects, Figure 14B reveals instead 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 voluntary points.

These results suggest that judges rely on two different criteria in detecting confusion. Although confusion has well defined facial and discourse correlates, judges appear to be consulting each channel independently in judging confusion. They pay attention to the face for the voluntary judgments and concentrate on the dialogues for the mandatory judgments.

*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 (Figure 14C).



**Figure 14. Precision scores for each emotion for the best one, two, and three channel models.**

*Flow.* Flow was only considered in the mandatory data set, so the composite models were 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 14D). So it is the face and the dialogue that collectively signaled heightened engagement akin to a flow experience.

*Frustration.* A more striking pattern was observed for frustration (Figure 14E). 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.

*Neutral.* The composite models had contradictory effects for the mandatory versus voluntary judgments. The precision of neutral and the number of channels were linearly related for the mandatory 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 voluntary 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 a inhibitory effect. So the face by itself is quite 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 voluntary model matched the FDP mandatory model.

#### *Structure of Composite Discriminate 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. While the previous analyses used a leave-one-out cross validation technique to evaluate classification accuracy, the current analysis focuses on discriminant models constructed on the entire data set. Accuracy estimates of non-cross-validated models were obtained from the training data itself, so there is the concern that they might provide an overly optimistic estimation of their efficacy. In fact, substantial differences between accuracy scores from non cross-validated models to cross-validated models have been reported in the literature. For example, Banse and Scherer (1996) obtained a 52.5% accuracy score in a non cross-validated acoustic based emotion discrimination study. The accuracy score was reduced by half (24.5%) when the model was cross-validated (Banse & Scherer, 1996). Similarly, Scherer and Ellgring (2007) reported that an emotion classification accuracy score of 79% was reduced to 49.1% when their discriminant model was cross validated. This model attempted to classify 14 emotions with 38 multimodal predictors (face, speech, and posture). Although this reduction in accuracy appears to be highly problematic, it is important to note that the model stabilized when the number of predictors was reduced to 10.

This positive result is tempered by the fact that the Scherer and Ellgring (2007) model used a data set of acted emotional expressions. Therefore, their results may not generalize to the current data set of naturalistic emotional expressions. An analysis was performed to assess the stability of the FDP models. A model is unstable and overfit when accuracy scores obtained from a non-cross validated evaluation significantly drops when the model is cross validated. Fortunately, the analyses revealed that there were no appreciable differences in the kappa scores for the FDP models. For the mandatory data

set, the kappas for the full and cross-validated FDP models were .347 and .288, respectively ( $p = .113$ ). The kappa scores for the full and cross-validated FDP voluntary models were .431 and .382, respectively ( $p = .108$ ). So there is some confidence in the stability of the composite models. A detailed analysis of how these models classify the learning-centered emotions appears below.

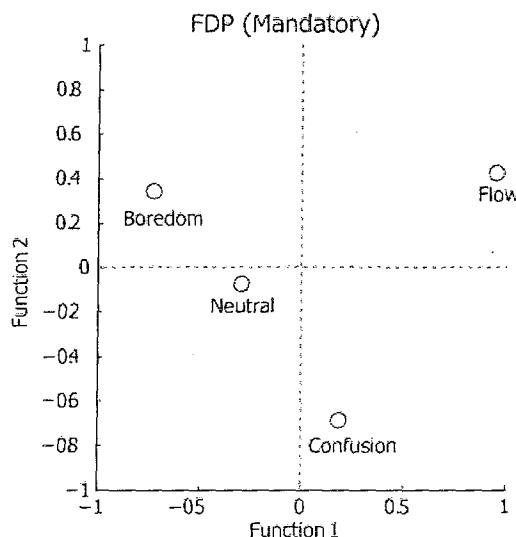
*Mandatory FDP Model.* Three discriminant functions were generated for the mandatory FDP model that attempted to classify four emotions<sup>7</sup>. These functions were all statistically significant,  $\chi^2(27) = 174.6, p < .001$  for Function 1;  $\chi^2(16) = 73.4, p < .001$  for Function 2;  $\chi^2(7) = 17.8, p = .013$  for Function 3. The first two functions were able to account 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 as projected on the first two discriminant functions are depicted in Figure 15. 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

---

<sup>7</sup> The number of functions required to classify  $g$  groups is  $g - 1$  (Klecka, 1980).

that the FDP model has some difficulty in discriminating between these emotions. Similar to boredom, 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. Although confusion is usually considered to be a negative state, it is positively associated with learning (Craig, Graesser et al., 2004; Graesser, Chipman et al., 2007), which explains why it is located on the positive side of the valence dimension.



**Figure 15. Group centroids for mandatory FDP model.**

According to Figure 15, Function 1 distinguishes the negative emotions of boredom and neutral from the more positive emotions of flow and confusion. Function 2 segregates emotions with lower arousal such as boredom, neutral, and 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 17). Each cell in the matrix represents the pooled within-groups correlations between predictors and standardized canonical discriminant functions (Klecka, 1980).

It appears that the number of characters in the learners' responses and the coverage on the seat of the chair positively correlated with Function 1 (Table 17). 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 tutors 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 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 17). 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 (i.e., high tutor directness).

**Table 17. Structure matrix for mandatory FDP model.**

Feature	Discriminant Function		
	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

*Voluntary FDP Model.* Four discriminant functions were generated for the voluntary 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 a 4.4% of the variance, the subsequent discussion focuses on the first three functions.

Figure 16 provides three different views of the distribution of the emotion centroids over the discriminant space. As evident in Figure 16A and Figure 16B, Function 1 segregates frustration and delight from boredom, neutral and confusion. The Structure Matrix listed in Table 18 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 discriminate between the affective states.

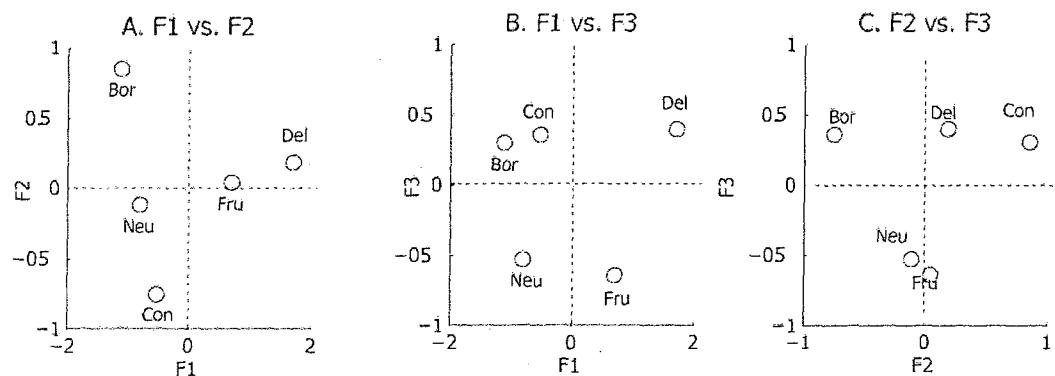


Figure 16. Group centroids for voluntary FDP model.

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 18). Perhaps these patterns are indicative of *fidgeting* as the session progresses.

**Table 18. Structure matrix for voluntary FDP model.**

Feature	Discriminant Function			
	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

Although Function 3 explains less variance than the other two functions, it appears to separate neutral and frustration from boredom, delight, and confusion (Figure 16B and Figure 16C). 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 18).

## Decision-Level Fusion

The purpose of the current analysis was to explore whether decision-level fusion yields superior kappa scores than feature level fusion. In contrast to feature-level fusion, where the sensory channels are fused before classification, fusion occurs after the classification in decision-level fusion. The decision-level fusion algorithm described above (Eq. 8–Eq. 13) was implemented, but with one important exception. According to the decision-level framework, each emotion node has two sources of activation: (1) activation from sensors, (2) activation from other emotions. The current implementation only considered the first source of activation. This is because there is currently no data driven model that specifies weights for the lateral inter-emotion links (i.e.,  $w_{kj}^e$ ). Inter-emotion activation is intended to simulate mixed-emotion effects (i.e., when more than one complimentary emotion is simultaneous experienced, e.g., frustration and confusion but not boredom and engagement). Although it is possible to obtain a set of weights from emotion theories and a variety of heuristics, focusing solely on sensory-driven activation is advantageous because it affords a more controlled comparison to feature-level fusion.

Different methods to weight the activation from the sensors were also considered. According to Eq. 9, a linear function specifies how the weight on a projecting (sensor-emotion) link biases the current activation of an emotion node. If in the current decision cycle, channels C1 and C2, detect emotion E with approximately equal probability, and previous analyses have indicated that C1 is more accurate at detecting E than C2, then the probability of C1 will be amplified and the probability of C2 will be suppressed.

Although the basic algorithm specifies a linear weighting scheme, it is possible to generate different decision models for the same set of sensors and emotions by varying the weighting scheme of the sensor-emotion links. Different weighting functions specify how the amplification and suppression occurs.

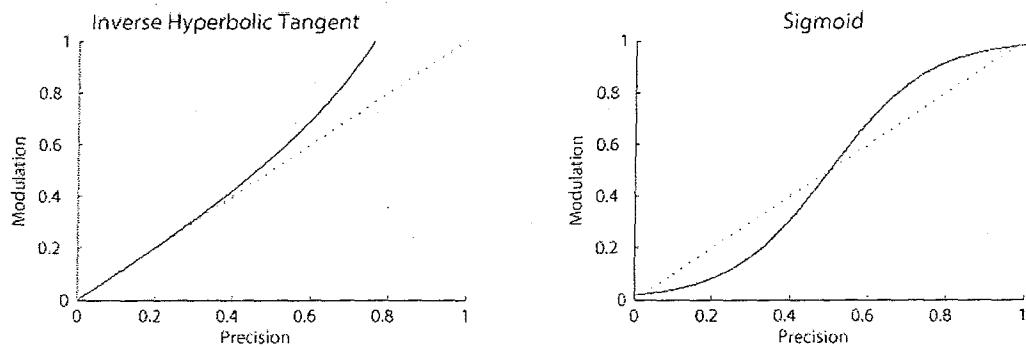
### *Weighting Schemes*

The current analyses compared the efficacy of five weighting schemes. Each function examines the precision by which an emotion was detected in the past and computes a modulation factor. This modulation factor indicates how the channels current estimates should be biased. The first scheme did not perform any weighting ( $w_{ij}^s = 1$ ); so a channel's prior discrimination history had no effect on its current contributions. The unweighted scheme represents the control condition, and one would expect the other four schemes to yield superior performance.

The next three schemes used a linear function (LI), an inverse hyperbolic tangent function (IN), and a sigmoid function (SG). Figure 17 compares the IN and SG functions to the linear function. According to Figure 17A, the IN function provides the same level of modulation as the linear function for the low and medium end of the precision spectrum. However, the modulation rates substantially increases when precision increases and modulation saturates at 1 for the high precision band of 0.8 to 1.

In contrast to the IN function that only provides excitation rates equal or greater to a linear function, the SG function excites as well as inhibits in comparison to the linear function. The modulation scores for this function are lower than linear modulation for

precision scores less than 0.5 but are greater than linear modulation for precision greater than 0.5<sup>8</sup>. Modulation rates of the two functions are approximately equal at the critical precision rate of 0.5.



**Figure 17. Weighting functions used for decision-level fusion.**

The fifth modulation strategy utilized an exclusive weighting function (XW). This weighting scheme is motivated by a *winner take all* strategy. For a particular emotion, the modulation of the sensor with the highest precision is set at 1. The modulation level of the other sensors are set to 0, effectively eliminating them from the decision cycle. It is important to note that the modulation rates are set at the individual emotion level. A single sensor could fully contribute towards one emotion and not contribute at all for another emotion. As a concrete example, assume that the task is to discriminate between E1 and E2 from sensors S1 and S2. If the precision of S1 > S2 for E1 and S1 < S2 for E2,

---

<sup>8</sup> The equation for a sigmoid is:  $y = \frac{1}{1+e^{-(ax+c)}}$ . The current analysis set  $a = 8$  and  $c = -4$ .

then the modulation vectors for S1 and S2 would be [1 0] and [0 1], respectively. In effect S1 fully activates E1 (but not E2) and S2 fully activates E2 (but not E1).

The efficacy by which the five weighting schemes could discriminate between the emotions was independently measured for the mandatory and voluntary data sets. A couple of qualifications need to be stated before presenting the results. First, the precision scores obtained from the discriminant analysis for the individual F, D, and P channels were used as a metric of the fidelity of each channel (see F, D, and P rows in Table 16). Second, with the exception of delight, the precision scores for the other emotions rarely exceeded 0.5. However, an examination of the modulation curves presented in Figure 17 indicates that the entire range of precision scores should be represented for the differential effects of the weighting schemes to be fully realized. Hence, the precision scores were normalized to lie within this range. The modulation scores used in the current analysis are presented in Table 19.

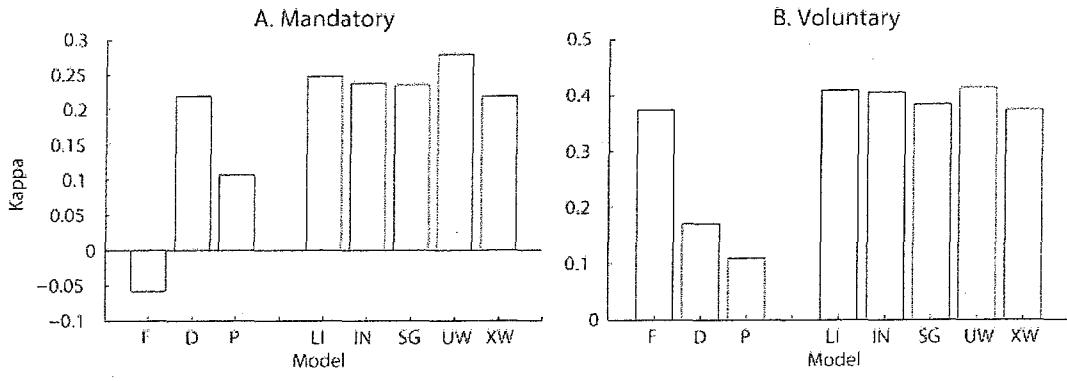
**Table 19. Modulation used in decision-level fusion.**

Function	Channel	Mandatory				Voluntary				
		Bor	Con	Flw	Fru	Bor	Con	Del	Fru	Neu
Linear	F	.250	.478	.458	.2632	.468	.578	.750	.384	.585
	D	.750	.634	.644	.4555	.510	.443	.380	.652	.25
	P	.559	.478	.593	.4395	.333	.604	.421	.308	.4082
ATanH	F	.255	.520	.494	.2695	.507	.660	.973	.405	.67
	D	.973	.748	.764	.4916	.563	.476	.400	.778	.2554
	P	.631	.520	.682	.4716	.346	.699	.449	.318	.4334
Sigmoid	F	.119	.456	.416	.1307	.436	.651	.881	.284	.6637
	D	.881	.745	.759	.4119	.520	.389	.276	.771	.1192
	P	.616	.456	.678	.3813	.208	.696	.347	.177	.3242
Unweighted	F	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	D	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	P	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Exclusive	F	.000	.000	.000	.000	.000	.000	1.00	.000	1.00
	D	1.00	1.00	1.00	1.00	1.00	.000	.000	1.00	.000
	P	.000	.000	.000	.000	.000	1.00	.000	.000	.000

*Note.* Bor = Boredom, Con = Confusion, Del = Delight, Flw = Flow, Fru = Frustration, Neu = Neutral

### *Classification Accuracy*

Each sensor had an independent discriminant analysis in order to obtain its probability distribution over the emotion space (i.e.,  $s_{ij}$  in Eq. 9). These probability estimates were then modulated in accordance to the different weighting scheme. The emotion node with the highest activation was considered to be the winner. The results of the decision-level FDP models are presented in Figure 18. For simplicity, this discussion focuses on the three-channel FDP model instead of the various two-channel combinations. Two important conclusions can be gleaned from these results. First, the accuracy of the feature level models and the decision-level models are statistically equivalent ( $\kappa_{\text{feature}} = .288$ ,  $\kappa_{\text{decision}} = .279$  for mandatory;  $\kappa_{\text{feature}} = .382$ ,  $\kappa_{\text{decision}} = .412$  for voluntary). It is important to note that the mandatory FDP model utilized nine features for classifications. The FDP decision-level model was constructed from 29 features (10 F, 10 D, and 9 P). So although kappa scores for both sensor fusion techniques are on par, the feature level fusion model is preferred by virtue of it having fewer parameters.



F = Face, D = Dialogue, P = Posture, LI = Linear, IN = Inverse Hyperbolic Tangent, SG = Sigmoid  
UW = Unweighted. XW = Exclusive Weighting

**Figure 18. Kappa scores for decision-level FDP models.**

A second finding was that the kappa scores for the five different weighting schemes were statistically equivalent. Although one would expect the unweighted model to be associated with the lowest kappa scores, this was clearly not the case. An examination of the precision scores for the individual emotions provides some insights into why the four weighting schemes did not result in superior classification accuracy over the unweighted model. Consider Figure 19 for the mandatory judgments (A-D). The LI, IN, and SG weighting schemes provided higher kappa scores than the single channel models and the unweighted model for boredom, confusion, and flow. However, these models were abysmal at detecting neutral. What separates neutral from the other states is that the single channel models have low precision for neutral. Therefore, the weighting schemes excite the other emotions but inhibit neutral, resulting in even higher precision scores for boredom, confusion, and flow but negative scores for neutral.

A similar pattern was observed for the voluntary judgments. Here, the precision of the weighted models was noticeably low for frustration and neutral (see Figure 19, E-I). It appears that the weighing schemes fail when one channel is very accurate at detecting an emotion but the other channels are especially poor (e.g., frustration and neutral). However, this pattern is not supported for delight, thereby warranting a deeper analysis of the discrimination patterns for this emotion.

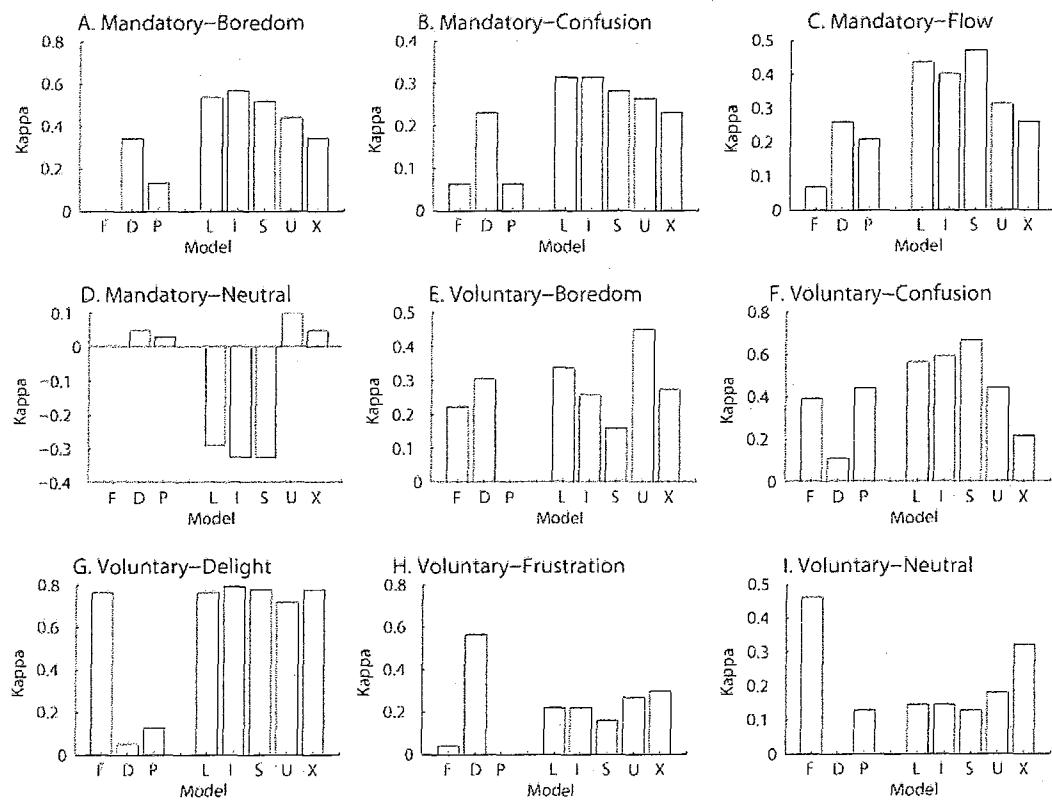


Figure 19. Precision scores for emotions, corrected for base rate.

## **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. These conclusions address the following goals: (1) To compare the accuracy by which the individual channels classify the emotions, (2) To compare single-channel affect detection with multi-channel affect detection, (3) To identify conditions in which a combination of channels yields superadditivity versus situations that indicate redundancy, (4) To analyze the structure of combined classification models, (5) To evaluate five different decision-level fusion techniques, and (6) To compare feature-level fusion with decision-level fusion. The subsequent discussion lists some of the most significant findings, followed with an analysis of some of the limitations and suggestions for alleviating these limitations.

### *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  $\times$  judgment type interaction, where the face was the most diagnostic channel for the voluntary judgments, while the dialogue was superior for the mandatory 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 voluntarily identify an emotion. But of equal importance, are situations where the face is unexpressive, even though the learner is having 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 voluntary judgments) have revealed that a number of action units that span the entire face can detect 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.

#### *Single-Channel versus Multiple-Channel Affect Detection*

The results revealed that the accuracy of multi-channel FDP model was statistically higher (albeit marginally) than the best single-channel dialogue model for the mandatory judgments. This FDP model also reduced the discrepancy (i.e., variance in the precision of the different emotions) of the D model. So a multi-channel model for tracking emotional states at the regularly polled timestamps would be recommended.

What about the more salient emotional states, namely the ones that occur at voluntary 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 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. 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 optimal precision.

Having decided to adopt a multi-channel model, there is the issue of deciding whether all three channels are essential or if a two-channel model would suffice. For the mandatory judgments, the FP model had lower accuracy scores than the FDP model, and hence cannot replace the FDP model. Performance 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 is 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 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). The second reason to select the FD model over the DP model is that its discrepancy score were about three times lower.

Unlike the mandatory judgments where models involving the dialogue were in competition, it is the models that involve the face that compete for the voluntary 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 mandatory judgments. It also had lower discrepancy scores than the FP model 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 mandatory and the voluntary contexts.

#### *Superadditive Effects*

The results indicated that superadditive effects were discovered for the mandatory judgments but not the voluntary 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 voluntary 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 mandatory judgments, where the dialogue was the most reliable channel. In fact, the degree of face and dialogue dominance for the voluntary and mandatory data sets are equivalent. For the voluntary data set, the percent improvement of the face over the next best channel (i.e., dialogue) is 119%. For the mandatory 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 mandatory but not for the voluntary 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 voluntary judgments can be attributed to a simple difference between the mandatory and voluntary judgments. It is not a stretch to assume that judges primarily relied on the face when they provided voluntary 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 voluntary 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 mandatory points. Here, the judges need to carefully monitor the additional channels in making their judgments, thereby yielding superadditive effects.

One remarkable finding about the mandatory 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 flow (.207), but not confusion (.064) and neutral (.031). The addition of facial features resulted in a small improvement in the precision of boredom (.180) and flow (.295), but a drastic improvement in detecting confusion (.295) and neutral (.167). 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.

### *Limitations*

There are three 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. It is important to reiterate that the most important analysis was on gross body language and conversational cues, the two channels that were automated. Hence, concerns with the facial data do not affect the findings for dialogue, posture, and combined dialogue-posture model.

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 10

parameters were included in each model, which eliminated any serious overfitting concerns, and (c) there were a sufficient number of observations to perform the statistical analyses with adequate power.

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.

### *Future Work*

There are two primary avenues of future research. One consists of using automated facial feature coding systems to annotate the videos of the learners' face to see if the observed patterns replicate with automated facial coding. The second research direction involves a deeper examination of the decision-level fusion method. The results indicated that the feature level fusion models could match the performance of the decision-level fusion models with approximately one third of the parameters. The problem with the decision-level fusion models was that weighting schemes overly rewarded emotions with high precision scores and penalized emotions with low precision scores. Additional weighting schemes such as an asymmetric sigmoid can be considered in an attempt to alleviate this problem.

The affect detection models do not include acoustic-prosodic information that has 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, 2007b; Scherer, Johnstone, & Klasmeyer, 2003). However, the multiple judge study has recently been replicated with a new version of AutoTutor that supports speech recognition (D'Mello, King, Entezari, Chipman, & Graesser, 2008). In that study participants verbally expressed their contributions rather than typing them in. 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.

## **Chapter 6: Concluding Discussion on Affect Detectors**

### **Introduction**

This dissertation addressed the possibility of using alternate sensory channels to classify the learning-centered affective states of boredom, confusion, flow, frustration, delight, and neutral (no affect). All four of the goals that were proposed in the introduction have been accomplished. These included: (1) The collection of an ecologically valid data set for training and validating the classifiers, (2) The development of system that classifies affective states based on conversational cues that are generated during natural language tutorial dialogues, (3) The development and evaluation of an affect detector that monitors the gross body language of a person while performing a learning task, and (4) The development of a multimodal system that classifies affective states by combining conversational cues, gross body language, and facial expressions.

The research activities conducted to satisfy each goal have been extensively discussed in the preceding four chapters. Each chapter also discussed the limitations of the methods and provided alternate strategies to alleviate these limitations. Hence, this chapter will focus on some of the broader issues that arise from this research. These include discussions on: (1) The accuracy of the automated affect detection systems, (2)

The possibility of boosting affect classification accuracy by modeling the temporal dynamics of the emotions, (3) Threats to scalability and potential solutions, (4) The impact of individual differences on affect detection, and (5) A case study that describes an application of the affect detectors developed in this dissertation.

### **Accuracy of Automated Affect Detectors**

The challenge of measuring emotions is beset with murky, noisy, and incomplete data, and is compounded with individual differences and contextual influences in experiencing and expressing emotions. Nevertheless the results obtained in this dissertation indicate that the classifiers are moderately successful in discriminating the affective states of boredom, confusion, delight, flow, and frustration from each other, as well as from the baseline state of neutral.

One may object to the use of the term *moderate* to characterize the classification results. However, it is imperative to note that an upper bound on automated classification accuracy of affect has yet to be established. Although human classifications may be considered to be the ultimate upper bound on system performance, human performance is variable and not necessarily the best gold standard. As discussed in Chapter 2, the results of the present research suggest that humans do not achieve a very high degree of concordance in judging emotions. The low inter-judge reliability scores associated with emotion recognition in naturalistic contexts independently replicate findings by a number of researchers (Ang et al., 2002; Forbes-Riley & Litman, 2004; Grimm et al., 2006;

Shafran et al., 2003). A study that directly compared emotion classification performance by humans to machine generated emotion labels is the el Kaliouby and Robinson (2005) work. They reported modest performance when a group of 18 people were asked to classify six (acted) affective states from a set of test videos. Humans had 54.5% accuracy scores, whereas a computer achieved accuracies of 63.5% (el Kaliouby & Robinson, 2005). However, the affect judges in that study were largely software developers. Perhaps higher classification accuracies could be obtained by humans trained in emotional intelligence, as in the case of clinical psychologists or FBI agents.

Statisticians have sometimes claimed, with hedges and qualifications, that kappa scores ranging from 0.4–0.6 are typically considered to be fair, 0.6–0.75 are good, and scores greater than 0.75 are excellent (Robson, 1993). On the basis of this categorization, the kappa scores obtained by the best classifiers would range from poor to fair. However, such claims of statisticians address the reliability of multiple judges or sensors when the researcher is asserting that the decisions are clear-cut and decidable. The present goal is very different. Instead, the goal is to use the kappa score as an unbiased metric of the reliability of making affect decisions, knowing full well that such judgments are fuzzy, ill-defined, and possibly indeterminate. A kappa score greater than 0.6 is expected when judges code some simple human behaviors, such as facial action units, basic gestures, and other visible behavior. However, in this case the human judges and computer algorithms are inferring a complex mental state. Moreover, it is the relative magnitude of these measures among judges, sensors, and conditions that matter, not the absolute magnitude of the scores. The argument put forward in this dissertation is that the lower kappa scores

are meaningful and interpretable as dependent measures (as opposed to checks for reliability of coding), especially since it is unlikely that perfect agreement will ever be achieved and there is no objective gold standard.

Although the moderate accuracy scores achieved by the affect detectors might seem problematic in applied settings, it is important to note that an affect-sensitive system does not need to respond to *all* of the affective experiences of the user. Within the context of affect-sensitive intelligent tutoring systems (ITSs), it is not imperative for an ITS to respond to every emotion it detects. What is important, however, is that if the ITS decides to react to an emotion of a learner, then it should be sufficiently confident that the correct emotion has been detected. Taking inappropriate action, such as incorrectly acknowledging frustration, can have very negative effects on the learner's perception of an ITS's capabilities and presumably learning gains.

Therefore, one strategy for an ITS in situations where an emotion cannot be confidently detected might be to simply ignore the affective element and choose its next action on the basis of the learners' cognitive states alone. Perhaps more attractive alternatives exist as well. For example, the tutor could bias the confidence of its actions as a function of the confidence of the emotion estimate. If the ITS lacks confidence in its assessment of frustration, then an empathetic response may be preferred over the ITS directly acknowledging the frustration and drastically altering its dialogue strategy. Another possibility is to use probabilistic models, such as Dynamic Decision Networks, that can model the noisy data associated with recognizing emotions. Future research will

be devoted to experimenting with these alternative strategies to compensate for imperfect affect recognition.

## **Incorporating Temporal Context in Affect Sensing Methods**

Affective experiences are rarely context free (Aviezer et al., 2008; Barrett, 2006; Russell, 1994, 2003; Stemmler, Heldmann, Pauls, & Scherer, 2001). Hence, affect detectors that consider contextual factors along with bodily channels are expected to outperform systems that rely on bodily expressions alone. The affect sensing system developed with AutoTutor incorporates some contextual information through the features derived from the student-tutor dialogues. The findings confirm the importance of context because the combined face + dialogue model outperformed the model that relied on facial expressions alone. Despite this positive result, there is still room for improvement as the scope of the contextual features that were considered are limited in a significant way. In particular, the focus was on the tutorial context while ignoring two important types of temporal context. These include graded differences in the relative duration of the different emotions, and transitions between emotions. One hypothesis is that the affect detectors can be substantially improved by focusing on these two sources of information.

### *Graded Difference in the Relative Durations of the Affective States*

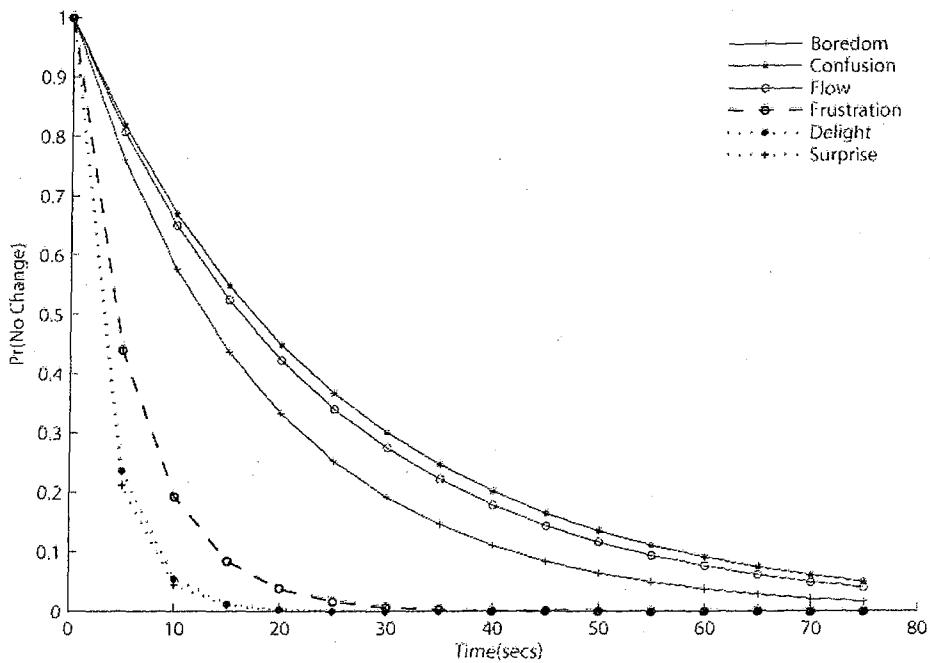
The relative duration of emotion  $E_i$  can be operationally defined as the absolute time between the onset of  $E_i$  and the time when the emotion either transitions into the neutral

state or another emotion  $E_j$  ( $E_i \neq E_j$ ). Modeling the relative duration of an emotion is similar to deriving a function that specifies how the emotion decays over time (i.e., a decay curve). Affect classifiers can consult these decay curves when the bodily measures provide ambiguous results. For example, consider a situation in which an affect classifier detects boredom at time  $t$ , and boredom and confusion are detected with equal likelihood at time  $t + 1$ . If the classifier is fortified with a model that specifies boredom is more likely to persist at  $t + 1$ , then this information can be used to decide between these two competing emotions. In this situation boredom would be excited and confusion would be inhibited.

It is possible to theoretically align the learning-centered emotions on a temporal scale so that affect classifiers can be sensitive to the temporal dynamics of the emotions. The following temporal scale in increasing order of persistence is proposed: (Delight = Surprise) < (Confusion = Frustration) < (Boredom = Flow). These predictions can be best understood from the perspective of goal-appraisal theories of emotion (Mandler, 1976, 1984b, 1999; Stein & Hernandez, 2007; Stein et al., 2008; Stein & Levine, 1991). In general, learners are typically in a *prolonged* state of engagement (flow) or disengagement (boredom) as they attempt to assimilate new information into existing knowledge schemas. When new or discrepant information is detected, attention shifts to discrepant information, the autonomic nervous systems increases in arousal, and the learner experiences a variety of possible emotions, depending on the context, the amount of change, and whether important goals are blocked. In the case of extreme novelty, the event evokes surprise. When the novelty triggers the achievement of a goal, the emotion

is positive, such as delight or even one of those rare *eureka* experiences. Previous research on delight and surprise has indicated that these emotions are typically quite brief (Ekman, 1984, 1992) and it is difficult to envision a learner sustaining states of delight and surprise for more than a few seconds. In contrast, confusion and frustration occur when the novelty triggers an impasse, where students get stuck and important goals are blocked. The learner needs to stop, think, effortfully deliberate, and problem solve. These emotions are expected to persist longer than delight and surprise because of the aforementioned cognitive activities that accompany their experience.

A recently developed set of exponential decay models appear to successfully capture graded differences in the decay rates of the various emotions (see Figure 20) (D'Mello & Graesser, in review). The models supported a tripartite classification of learning-centered emotions along a temporal dimension: persistent emotions (boredom, flow, and confusion), transitory emotions (delight and surprise), and an intermediate emotion (frustration). This pattern somewhat confirms the aforementioned predictions stemming from goal-appraisal theories of emotion (Mandler, 1976, 1984a, 1984b, 1999; Stein et al., 2008; Stein & Levine, 1991), with the exception that confusion was categorized as a persistent rather than an intermediate emotion. The next step is to integrate these models into the affect classifier and evaluate if there are any incremental gains in classification accuracy.



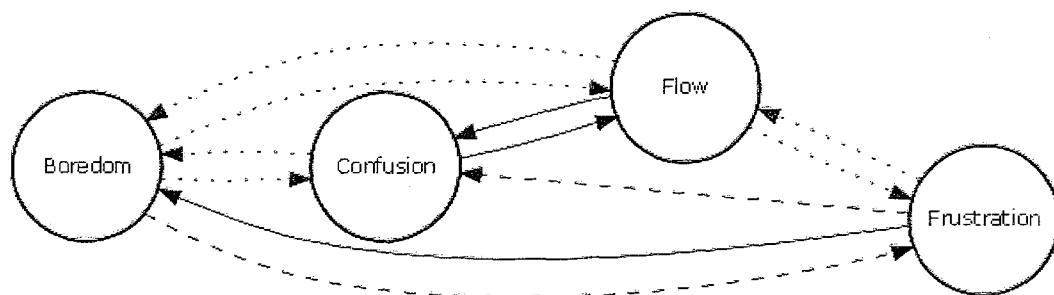
**Figure 20. Exponential decay curves for the learning centered emotions.**

#### *Transitions between the Affective States*

Students periodically change their affective states during the course of a learning session, so the transitions between these different affective states are a critical phenomenon to explore. A model that specifies the transition probabilities between the affective states can be incorporated into affect detectors with the hope that this form of predictive modeling can scaffold sensor-based diagnostic modeling, resulting in enhanced classification accuracies.

There currently is no widely accepted theoretical framework that explicitly addresses the issue of transitions between affective states during complex learning tasks. One cognitive model emphasizes the importance of *cognitive disequilibrium* (Festinger,

1957; Graesser & Olde, 2003; Piaget, 1952) and can be extended to provide some predictions regarding likely affective state transitions. According to this theory, deep comprehension is most likely to occur when learners confront contradictions, anomalous events, obstacles to goals, salient contrasts, perturbations, surprises, equivalent alternatives, and other stimuli or experiences that fail to match expectations (Jonassen, Peck, & Wilson, 1999; Mandler, 1976; Schank, 1986). Individuals in a state of cognitive disequilibrium have a high likelihood of activating conscious and effortful cognitive deliberation, questions, and inquiry that are directed to restore cognitive equilibrium and result in learning gains. Kort, Reilly, and Picard (2001) predicted that the affective states of confusion, and perhaps frustration, are likely to occur during cognitive disequilibrium, while affective states such as boredom and flow would typically occur during cognitive equilibrium. This extended cognitive disequilibrium model would make a number of plausible predictions for the transitions between the states of boredom, confusion, flow, and frustration (see Figure 21).



**Figure 21. Predicted affective transitions.**

Solid links indicate that a transition is expected. Dotted lines indicate that the transition is highly unlikely, whereas dashed links mean that the model makes no explicit prediction.

It is important to note the correlations between affective states and learning in order to fully understand some of these predictions. Boredom is negatively correlated with learning, whereas confusion and flow are positively correlated with learning (Craig, Graesser et al., 2004; Graesser, Chipman et al., 2007). Therefore, a bored learner is not expected to transition into flow or confusion. In contrast, transitions from confusion to flow and vice versa would be expected because of (1) There is a positive correlation between both of these emotions with learning, and (2) An interplay between these affective states has been explicitly predicted by the cognitive disequilibrium model. Students in the state of flow are continuously being challenged within their zones of optimal learning (Metcalfe & Kornell, 2005) and are experiencing two-step episodes alternating between confusion and insight. Transitions from confusion and flow to a state of disengagement (boredom) would also be highly unlikely. On the other hand, it is plausible that frustration may gradually transition into boredom, a crucial point at which the learner simply disengages from the learning process. Frustration is not likely to transition into flow in short learning sessions, whereas this may eventually occur over longer stretches of time. It should be noted that the model does not make predictions regarding the likelihood of Boredom → Frustration or Frustration → Confusion transitions. However, it is conceivable that these transitions might unfold if a learner is in a state of boredom or frustration for a long period of time.

An exploratory analysis tested the fidelity of this hypothesized transition model (D'Mello, Taylor, & Graesser, 2007). The supported predictions include the transitions from the state of boredom into confusion, flow into frustration, and confusion into boredom which occurred significantly *below* chance. The three predictions that had trends in the predicted direction, but were not statistically significant include the unlikely transition from flow into boredom and the likely transitions from flow to confusion and frustration to boredom. While this might be interpreted as evidence to support the extended cognitive disequilibrium model, some of the findings are a big stretch for the model. In particular four predictions made by the model were not supported.

There were also two interesting findings that were not addressed by the model. First, the transition from boredom into frustration occurred significantly *above* chance. Second, the transition from frustration into confusion occurred rarely, was not significant, and had a high degree of variability. This prompts speculation about how individual differences might be especially relevant to this transition. Perhaps some individuals disengage when frustrated, while others view the situation as a challenge and become more energized—and ultimately enter the confusion state while trying to resolve the current misunderstanding.

The next step is to integrate the transition model into the affect detectors and to assess whether this results in an improvement in classification accuracy over the bodily measures alone.

## **Scalability of Affect Detectors**

One disadvantage of the affect detectors is that they require expensive, customized hardware and software, such as the Body Pressure Measurement System (BPMS) and automated facial feature tracking systems. This raises some scalability concerns for those who want to extend this program of research into classrooms. Hence, the prospects of developing systems to classify the emotions on the basis of cost-effective and scalable sensors are being currently explored. Two approaches under consideration are text-based affect sensing and camera-based body position and motion tracking.

### *Text-Based (Language-Based) Affect Detectors*

Text-based affect detectors are advantageous because they are cost-effective and scalable. Furthermore, textual information is abundantly available in any dialogue based tutoring environment. The idea of using textual features to detect affect is not new. A number of research groups have proposed domain-independent, text-based, affect detectors that operate by constructing affective models from large corpora of world knowledge. The models are used to predict the affective tone of segments of text such as movie reviews, product reviews, blogs, instant messaging, and email messages (Gill, French, Gergle, & Oberlander, 2008; Hancock, Curry, Goorha, & Woodworth, 2008; Hancock, Landrigan, & Silver, 2007; Liu, Lieberman, & Selker, 2003; Shaikh, Prendinger, & Ishizuka, 2007, 2008). However, these systems operate under the assumption that affective content is explicitly and literally articulated in the text (e.g., “I have some bad news”, “This movie

is a real drag’’). Although this may be a valid assumption for obviously emotion-rich corpora such as blogs and movie reviews, where people are directly expressing opinions, it is unclear whether learners’ responses to computer tutors resonate with affect rich content. In fact, there is some evidence to the contrary. An examination of 1637 student responses generated from a tutorial session with AutoTutor yielded only a handful of utterances with explicit affective expressions (< 1%). But an in-depth analysis of videos of the tutorial sessions yielded approximately 3000 affective experiences (see Chapter 2). Although students are experiencing affective states while interacting with AutoTutor, their typed responses do not necessarily convey affective content explicitly. Instead their responses mainly consist of domain specific answers to the tutor’s questions even when they are in the midst of rich affective experiences. Therefore, a more systematic textual analysis of tutorial dialogues might be necessary to uncover subtle cues that might be diagnostic of learners’ affective states.

This hypothesis was recently investigated by analyzing cohesion relationships in naturally occurring tutoring dialogues with AutoTutor (D’Mello, Dowell, & Graesser, in press; Graesser & D’Mello, in preparation). Cohesion is an important discourse construct that might provide cues into complex mental phenomenon such as learners’ affective states. Cohesion, a textual construct, is a measurable characteristic of text that is signaled by relationships between textual constituents (Graesser, McNamara et al., 2004; McNamara et al., 2008; McNamara, Ozuru, Graesser, & Louwerse, 2006). It is related to coherence, a psychological construct, that is a characteristic of the text together with the reader’s mental representation of the substantive ideas expressed in the text (Graesser,

McNamara et al., 2004). Therefore, variations in the cohesiveness of the tutorial dialogues should be predictive of the learners' affective experiences. A breakdown in cohesion may be expected to be predictive of affective states such as confusion and frustration, whereas strong cohesive relationships might be indicative of engagement.

Preliminary analyses consisted of automatically computing multiple measures of cohesion (e.g., pronouns, connectives, semantic overlap, causal cohesion, co-reference) using the Coh-Metrix facility for analyzing discourse and language characteristics of text (Graesser, McNamara et al., 2004). Coh-Metrix is a validated computational tool that provides over 100 measures of various types of cohesion. Cohesion measures were selected on a theoretical basis and used as independent variables in multiple regression models that predicted the occurrence of the affective states. The models were found to predict the proportional occurrence of boredom, confusion, flow, and frustration, yielding medium to large effect sizes. These results suggest that tracking cohesion features is indeed a viable method for affect detection.

The next step of this research is to implement real time, cohesion-based, affect detectors. The current set of regression models were constructed at the subject level as the primary goal of the analyses was to explore the possibility of deriving a set of predictors that were diagnostic of the affective states. This goal can be achieved by analyzing cohesion relationships in incremental windows of student and tutor dialogues that are generated as the tutoring session progresses. An analysis to evaluate whether the new lightweight, scalable language-based sensing systems can adequately substitute for

systems that also track gross body language and facial features will be conducted. If not, then body posture and facial expressions are critical.

#### *Camera-Based Body Position and Motion Tracking*

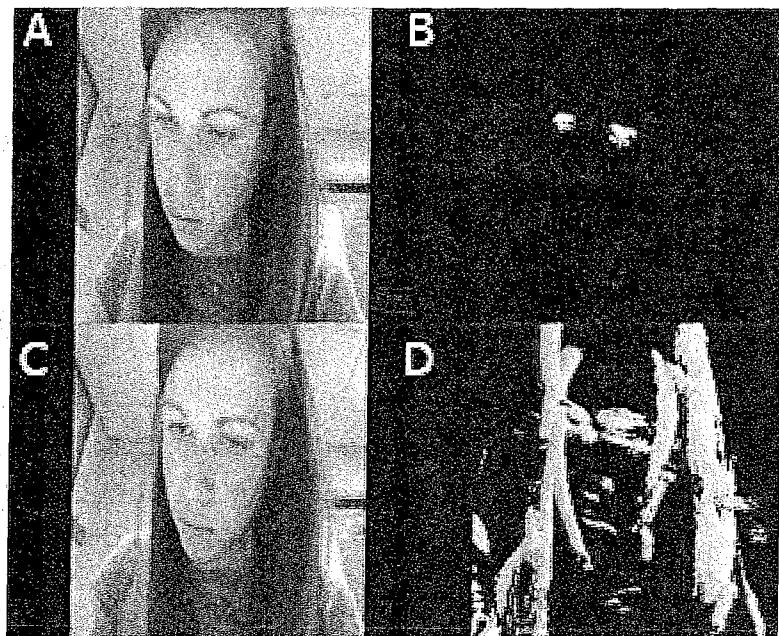
The most expensive sensor currently being used is the Body Pressure Measurement System (BPMS). The advantage of this system is that it affords non-intrusive and high resolution (1558 sensels) sensing of learners' body language when seated in a chair. Its greatest disadvantage is its cost ( $\approx \$10,000$ ), which undoubtedly impacts deployment in *in vivo* settings. Hence, cameras as a low resolution but cost effect-effective alternative to the BPMS system are being considered. One concern with sacrificing resolution to reduce costs is that performance of the affect detector might be adversely affected. However, the analysis that compared classification accuracy of high-level (low-resolution) to low-level (high-resolution) sensing did not yield any substantial differences (see Chapter 5 where high-level pressure features were compared to spatial-temporal contours), suggesting that performance is not compromised with low-resolution sensing.

It is possible to use a video camera to obtain an index of bodily position and arousal, the two primary components of the attentive-arousal framework (Bull, 1987). Body position can be inferred from the distance between a person and the camera. Face detection algorithms usually produce a bounding rectangle that represents a person's face (Viola & Jones, 2001). The area of the rectangle provides a measure of the distance of the person's face from the camera (i.e., as area increases distance to camera increases). If the camera is correctly positioned, for example, a laptop with a built-in camera, then the

distance from the camera can be used to approximate the distance from the target stimuli being projected on the computer screen. Closeness to the screen is a sign of attentiveness and engagement whereas the likelihood of disengagement increases with increasing distances.

Motion tracking algorithms make it possible to estimate how much a person is moving by computing the amount of motion that is present in a sequence of frames from a video camera. A simple motion tracking algorithm would first construct a background model for each pixel. Two actions would be performed when a new frame is encountered. The first involves deciding whether each pixel in the new frame is different (via some thresholding function) from the corresponding pixel in the background model. The second step is to update the background model for each pixel perhaps via a weighted moving average algorithm. The proportion of pixels with motion provides an index of the amount of bodily arousal in each frame.

An illustration of the output of the motion tracking algorithm is presented in Figure 22. Panels A and C represent single frames extracted from a video sequence, while panels B and D show the output of the motion tracking system. In Panel B, with the exception of the eyes, the body is motionless. In contrast, there is significant motion in the face and body that is evident in Panel D. It is important to note that background noise (i.e., the patterns on the walls and ceilings) have been correctly filtered out in both cases.



**Figure 22. Sample frames from motion tracking algorithm.**

A and C are original video frames. B and D are motion frames for A and C, respectively.

In this fashion, both body position and arousal can be automatically tracked via face-detection and motion tracking algorithms, respectively. These systems utilize cost-effective, commercially available web cams that can be readily deployed in *in vivo* settings such as classrooms. The fidelity of these methods as alternatives to the BPMS system is being explored.

## **Individual Differences in Affective Experience and Expression**

It is widely acknowledged that individual differences play an important role in how people experience and express their emotional states. Although it has been argued that there are broad commonalities in the experience and expression of the basic emotions across individuals and cultures, this *natural kind* view of the basic emotions has been challenged in the recent years (see Barrett, 2006 for a comprehensive review of the debate). A striking example of the importance of individual differences in modeling affect can be found in a series of studies by Ohman and his colleagues (Ohman, 2004; Ohman & Soares, 1993, 1994). They studied the effects of conscious and unconscious exposure of fearful stimuli on participants with phobias for snakes versus participants with phobias for spiders, versus those with no such phobias (control). The results revealed that participants' with a snake (but not spider) phobia showed autonomic arousal increases (via skin conductance) when presented with images of snakes, but not spiders and not control stimuli such as buttons or mushrooms. The autonomic arousal of participants' with a phobia for spiders increased when the stimuli were spiders but not snakes and neutral stimuli. This form of selective phobic responses highlights the importance of individual differences in emotional experiences and expressions. It is important to note that contextual factors alone cannot account for this effect because a contextual model that does not consider individual differences would predict that autonomic arousal would increase when exposed to both snakes and spiders, which is clearly not the case.

Given the importance of individual differences in the affective sciences, an affect classifier should account for these differences at some level. There are two primary approaches to addressing individual differences issues in the design of affect detection systems. These include generalizing across individuals and modeling to the individual.

The first approach of generalizing across individuals maintains that a feature is only included in a model if its diagnostic ability generalizes across a sample of participants that are believed to be representative of the population. The feature selection mechanism used in this research was an exemplar of this form of generalization (see Chapter 3). The advantage of this method is that it less laborious, scalable, and can be reasonably accurate. The obvious disadvantage is that it glosses over interesting patterns of individual differences.

Modeling to the individual entails a calibration procedure in which the classifiers are tuned to be sensitive to the subtleties and nuances of each person's affective experiences. So feature  $X$  might be included in the classifier for individual  $I_1$  but not  $I_2$ . Or more commonly, feature  $X$  might be configured differently for  $I_1$  versus  $I_2$ . The calibration procedure usually involves subjecting participants' to stimuli that is expected to elicit particular emotional expressions and observing how each participant responds. Apart from being time consuming and comparatively less scalable (if experimenters are required to perform the calibration), there are two significant disadvantages. The first disadvantage pertains to the fact that there is no guarantee that a participant's expressions to an emotionally eliciting stimuli during the calibration process adequately represents how they might express the same emotion in context. Second, a relatively short

calibration process over a single session might be insufficient to encompass the entire gamut of expressions to a particular emotion. In some situations, it might take weeks to obtain a sufficiently diverse set of exemplars of how a particular emotion is expressed (Picard et al., 2001).

Quite clearly, neither approach is particularly appealing. Perhaps a combined model that utilizes both methods of addressing individual difference concerns will yield the best results. However, most affect detection systems adopt generalizing across individuals as it requires comparatively less effort. Hence, a deeper exploration into the weakness of this method is warranted.

The major problem with this method is that if an emotionally expressive behavior (i.e., facial expression, posture, gesture, etc.) can be controlled by an individual, then there is no guarantee that every individual will exhibit the behavior in the same way. Some may suppress the behavior, while others might exaggerate it. This suggests that the generalization problem might be overcome by focusing on unconscious behaviors that might co-occur with affective experiences.

One viable unconscious bodily behavior that has been recently explored as an extension to this dissertation is the  $1/f$  pattern of bodily motion.  $1/f$ -noise, also known also as pink noise or fractal scaling, occurs in a time series that exhibits both short and long term correlations (Kello, Anderson, Holden, & Van Orden, 2008; Van Orden, Holden, & Turvey, 2003). It is considered by many to be a fundamental property of nonlinear, complex dynamical systems studied across the physical and life sciences (Mandelbrot, 1998). In the cognitive sciences, the presence of this noise in human

behavior has been viewed as evidence that human cognition should be considered a dynamical system (Van Orden et al., 2003). More recently, it has been argued that any reliable measure of cognition will reveal patterns of “intrinsic”  $1/f$  fluctuation (Kello et al., 2008).

One recent analysis considered variations in  $1/f$  patterns, also known as pink noise or fractal scaling, in the gross body language of students during a complex learning task. The variations in  $1/f$  patterns were powerful predictors of individual differences in the experience of mental states such as confusion and frustration (D’Mello, Dale, & Graesser, in review). The results indicated that learners who are in states of blasé comprehension produced body movements that are characteristic of correlated pink noise, as would be expected from self-organizing systems. However, learners who experienced confusion and frustration, states that are diagnostic of cognitive disequilibrium, exhibited fluctuations in body motion that are consistent with a whitening of the signal. This approach shows sufficient promise that future efforts will be devoted towards using the fractal signal in an automated affect detector for body movements.

## **Applications of Affect Detectors**

Classification of learner emotions is an essential step in building a tutoring system that is sensitive to the learner's emotions. The other essential component towards affect-sensitivity is to build mechanisms that empower ITSs to intelligently respond to these emotions, as well as to learners' states of cognition, motivation, social sensitivity, and so on. In essence, how can an affect-sensitive ITS respond to the learner in a fashion that optimizes learning and engagement?

The automated affect detection systems developed in this dissertation were integrated into a new version of AutoTutor that adapts to both the cognitive and affective states of learners. Boredom, confusion, and frustration may be viewed as negative emotions. If these states are handled appropriately, there could be a positive impact on engagement and learning outcomes. Flow, on the other hand, is a highly desirable positive affective state that is beneficial to learning. Although most tutoring environments would want to promote and prolong the state of flow, any intervention on the part of the tutor runs the risk of adversely interfering with the flow experience. Therefore, the current version of the affect-sensitive AutoTutor does not respond to episodes of flow. Instead, it focuses on addressing the affective states of boredom, frustration, and confusion.

At this point in science, there are no empirically proven strategies to address the presence of boredom, frustration, and confusion. An examination of the literature provided some guidance on how best to respond to these three affective states. The focus

has been on two major theoretical perspectives that address the presence of these negative emotions. These two perspectives appeal to attribution theory (Batson, Turk, Shaw, & Klein, 1995; Heider, 1958; Weiner, 1986) and cognitive disequilibrium during learning (Festinger, 1957; Graesser & Olde, 2003; Piaget, 1952). Details on how these theories were used to derive affect-sensitive responses are discussed in recent publications (D'Mello, Jackson et al., 2008; D'Mello, Craig, Fike, & Graesser, in press).

In addition to theoretical considerations, the assistance of experts was enlisted to help create the set of tutor responses. Two experts in pedagogy, with approximately a decade of related experience each, were provided with excerpts from real AutoTutor dialogues (including both the tutor and student dialogue content, screen capture of the learning environment, and video of the student's face). There were approximately 200 excerpts averaging around 20 seconds in length, each of which included an affective response by the student. The experts were instructed to view each of the excerpts and provide an appropriate follow-up response by the tutor.

#### *Strategies to Respond to Learners' Affective States*

A set of production rules that handled and responded to the learner's boredom, confusion, and frustration were created. The rules were an amalgamation of attribution theory, cognitive disequilibrium theory, and the recommendations made by the experts. Although the rules created by the pedagogical experts allowed for any possible action on the part of the tutor, AutoTutor can only implement a portion of those actions. For example, one possibility to alleviate boredom would be to launch an engaging simulation or a

seductive, serious game. However, the current version of the tutor does not support simulations or gaming, so such a strategy is not immediately realizable. Consequently, the current focus is limited to production rules that could be implemented by AutoTutor's current actions which include feedback delivery (positive, negative, neutral), a host of dialogue moves (hints, pumps, prompts, assertions, and summaries), and the facial expressions and speech modulation by AutoTutor's embodied pedagogical agent (EPA).

The production rules were designed to map dynamic assessments of the students' cognitive and affective states with tutor actions to address the presence of the negative emotions. There were five parameters in the student model and five parameters in the tutor model. The parameters in the student model include: (a) the current emotion detected, (b) the confidence level of that emotion classification, (c) the previous emotion detected, (d) a global measure of student ability (dynamically updated throughout the session), and (e) the conceptual quality of the student's immediate response. AutoTutor incorporates this five-dimensional assessment of the student and responds with: (a) feedback for the current answer, (b) an affective statement, (c) the next dialogue move, (d) an emotional display on the face of the EPA, and (e) emotionally modulating the voice produced by AutoTutor's text-to-speech engine.

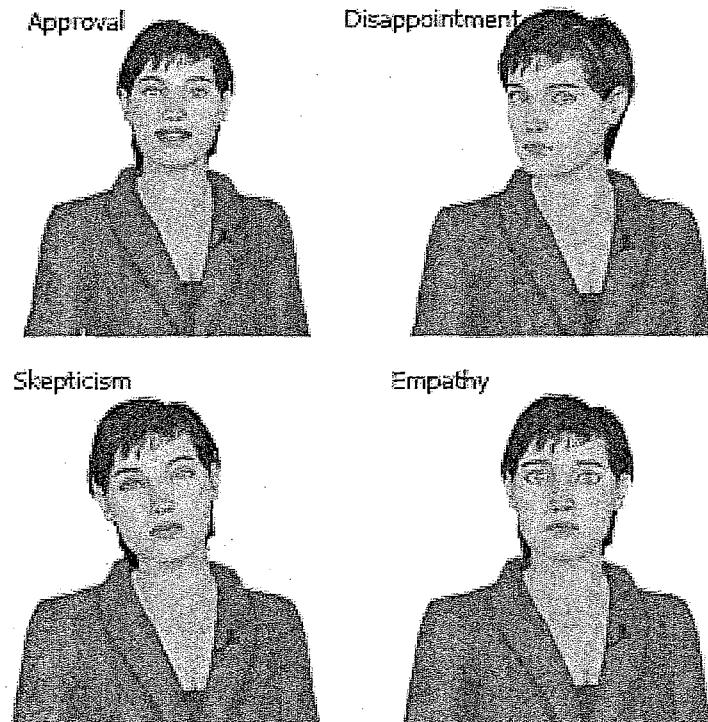
As a complete example, consider a student that has been performing well overall (high global ability), but the most recent contribution is not very good (low current contribution quality). If the current emotion is classified as boredom, with a high probability, and the previous emotion is classified as frustration, then AutoTutor might say the following: "Maybe this *topic* is getting old. I'll help you finish so we can try

something new.” This is a randomly chosen phrase from a list that was designed to indirectly address the student’s boredom and to try to shift the topic a bit before the student becomes disengaged from the learning experience. This rule fires on several different occasions, and each time it is activated AutoTutor will select a dialogue move from a list of associated moves. In this fashion, the rules are context sensitive and are dynamically adaptive to each individual learner.

The subsequent section discusses each of the major components of the affect-sensitive AutoTutor. These include the short feedback, an emotional or motivational expression that is sensitive to the learners’ affective and cognitive states, an emotionally expressive facial display, and emotionally modulated speech.

*Short Feedback.* AutoTutor provides short feedback to each student response. The feedback is based on the semantic match between the response and the anticipated answer. There are five levels of feedback: positive, neutral-positive, neutral, neutral-negative, and negative. Each feedback category has a set of predefined expressions that the tutor randomly selects from. “Good job” and “Well done” are examples of positive feedback, while “That is not right” and “You are on the wrong track” are examples of negative feedback. In addition to articulating the textual content of the feedback, the affective AutoTutor also modulates its facial expressions and speech prosody. Positive feedback is delivered with an *approval* expression (big smile and big nod). Neutral positive feedback receives a *mild approval* expression (small smile and slight nod). Negative feedback is delivered with a *disapproval* expression (slight frown and head

shake), while the tutor makes a *skeptical* face when delivering neutral-negative feedback (see Figure 23). No facial expression accompanies the delivery of neutral feedback.



**Figure 23. Affect synthesis by embodied pedagogical agents.**

(Created by Karl Fike and Sidney D'Mello)

*Emotional Response.* After delivering the feedback, the affective AutoTutor delivers an emotional statement if it senses that the student is bored, confused, or frustrated. A non-emotional discourse marker (e.g., “Moving on”, “Try this one”) is selected if the student is neutral. Two pedagogically distinct variants of the affect-

sensitive AutoTutor are currently being implemented. These include a *Supportive* and a *Shakeup* AutoTutor.

The supportive AutoTutor responds to the learners' affective states via empathetic and motivational responses. These responses always attribute the source of the learners' emotion to the material instead of the learners' themselves. So the supportive AutoTutor might respond to mild boredom with "This stuff can be kind of dull sometimes, so I'm gonna try and help you get through it. Let's go." A more encouraging response is required for severe boredom ("Let's keep going, so we can move on to something more exciting"). An important point to note is that the supportive AutoTutor never attributes the boredom to the student. Instead, it always blames itself or the material.

A response to confusion would include attributing the source of confusion to the material ("Some of this *material* can be confusing. Just keep going and I am sure you will get it") or the tutor itself ("I know *I* do not always convey things clearly. I am always happy to repeat myself if you need it. Try this one"). If the level of confusion is low or mild, then the pattern of responses entails: (a) acknowledging the confusion, (b) attributing it to the material or tutor, and (c) keeping the dialogue moving forward via hints, prompts, etc. In cases of severe confusion, an encouraging statement is included as well.

Similarly, frustration receives responses that attribute the source of the frustration to the material or the tutor coupled with an empathetic or encouraging statement. Examples include: "I may not be perfect, but I'm only human, right? Anyway, let's keep

going and try to finish up this problem.”, and “I know this *material* can be difficult, but I think you can do it, so let’s see if we can get through the rest of this problem.”

The major difference between the breakup AutoTutor and the supportive AutoTutor lies in the source of emotion attribution. While the supportive AutoTutor attributes the learners’ negative emotions to the material or itself, the breakup AutoTutor directly attributes the emotions to the learners. For example, possible breakup responses to confusion are “This material has got *you* confused, but I think you have the right idea. Try this...” and “*You* are not as confused as you might think. I’m actually kind of impressed. Keep it up”.

Another difference between the two versions lies in the conversational style. While the supportive AutoTutor is subdued and formal, the breakup tutor is edgier, flaunts social norms, and is witty. For example, a supportive response to boredom would be “Hang in there a bit longer. Things are about to get interesting.” The breakup counterpart of this response is “Geez this stuff sucks. I’d be bored too, but I gotta teach what they tell me.”

*Emotional Facial Expressions and Emotionally Modulated Speech.* Seven facial expressions were generated for the affective AutoTutor. These include: approval, mild approval, disapproval, empathy, skepticism, mild enthusiasm, and high enthusiasm. The *Short Feedback* section lists some of the conditions upon which these expressions are triggered. The supportive and breakup responses are always paired with the appropriate expression, which can be neutral in some cases.

Example affective displays are illustrated in Figure 23. The facial expressions in each display were informed by Ekman's work on the facial correlates of emotion expression (Ekman, 1984, 1992, 2003; Ekman & Friesen, 1975). For example, empathy is a sense of understanding displayed to the user. This is manifested by an inner eyebrow raise, eyes open, and lips slightly pulled down at the edges (action units 1, 5, 15) (Chovil, 1991). Skepticism is a combination of confusion and curiosity, characterized by a furrowing of the brow, an eye squint, and one outer eyebrow is raised (action units 2, 4, 7) (Craig et al., 2008; McDaniel et al., 2007). These displays were created with the Hapttek™ Software Development Kit.

The facial expressions of emotion displayed by AutoTutor are augmented with emotionally expressive speech synthesized by the agent. The emotional expressivity is obtained by variations in pitch, speech rate, and other prosodic features. Previous research has led us to conceptualize AutoTutor's affective speech on the indices of pitch range, pitch level, and speech rate (Banse & Scherer, 1996; Johnstone & Scherer, 2000; Scherer, 2003; Scherer & Ellgring, 2007b; Scherer et al., 2003).

#### *Evaluating the Affect-Sensitive AutoTutor*

A new version of AutoTutor that aspires to be responsive to learners' affective and cognitive states via supportive and shakeup dialogues has been implemented. The affect-sensitive AutoTutor aspires to keep students engaged, boost self-confidence, and presumably maximize learning by narrowing the communicative gap between the highly emotional human and the emotionally challenged computer. A study that evaluates the

pedagogical effectiveness of the two affect-sensitive versions of AutoTutor when compared to the original tutor is in progress. This original AutoTutor has a conventional set of fuzzy production rules that are sensitive to cognitive states of the learner, but not to the emotional states of the learner. Both versions of the improved AutoTutor are sensitive to the learners' affective states in distinct ways. The obvious prediction is that learning gains and the learner's impressions should be superior for the affect-sensitive versions of AutoTutor. In addition to testing for learning gains, differences in learners' engagement levels while interacting with the different versions of AutoTutor will also be measured. The study will also test if personality differences predict preference for Supportive versus Shakeup AutoTutor.

The pedagogical strategies discussed above involve AutoTutor simply *reacting* to the emotions of the learner. However, this approach might not suffice if learners cycle through their emotions in a context-sensitive fashion. When learners experience the negative emotions of boredom and frustration, they are more likely to stay in these states rather than transition into the more positive states of flow and delight. In contrast, learners in a state of flow tend to remain engaged or alternatively transition into confusion, an affective state that is positively correlated with learning. Therefore, to optimize learning, AutoTutor may need to steer learners into a virtuous cycle of flow and confusion, while simultaneously avoiding the vicious cycle of boredom and frustration. This complex mechanism suggests that it may be important to move beyond the simple reactive strategy of detecting and responding to negative emotions. AutoTutor may also

need to *proactively* anticipate and attempt to prevent the onset of these negative emotions that are detrimental to learning and engagement.

The affect-sensitive AutoTutor represents one out of a handful of related efforts made by a number of researchers who have a similar vision (Aist et al., 2002; Baker, Rodrigo, & Xolocotzin, 2007; Burleson & Picard, 2007; Conati, 2002; Conati & Maclarens, in press; Forbes-Riley & Litman, 2004; Forbes-Riley et al., 2008; Kort et al., 2001; Litman & Forbes-Riley, 2004, 2006; McQuiggan et al., 2008; Picard, 1997; Woolf et al., 2007). Our unified vision is to advance education, intelligent learning environments, and human-computer interfaces by optimally coordinating cognition and emotions. Whether the affect-sensitive AutoTutor positively influences learning and engagement awaits further development and empirical testing.

## References

- Aist, G., Kort, B., Reilly, R., Mostow, J., & Picard, R. (2002). Adding human-provided emotional scaffolding to an automated reading tutor that listens increases student persistence. *Intelligent Tutoring Systems*, 2363, 992-992.
- Allen, J. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems*, 14(5), 14-16.
- Alm, C., & Sproat, R. (2005). *Perceptions of emotions in expressive storytelling*. Paper presented at the 9th biennial conference of the International Speech Communication Association (ISCA), INTERSPEECH.
- Andersen, P. (1999). *Nonverbal Communication: Forms And Functions*. Mountain View, CA: Mayfield Publishing Company.
- Anderson, J., Corbett, A., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, 4(2), 167-207.
- Anderson, J., Douglass, S., & Qin, Y. (2005). How should a theory of learning and cognition inform instruction? In A. Healy (Ed.), *Experimental cognitive psychology and it's applications* (pp. 47-58). Washington, DC.: American Psychological Association.
- Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). *Prosody-based automatic detection of annoyance and frustration in human-computer dialog*.

Paper presented at the International Conference on Spoken Language Processing,  
Denver, CO.

Arnold, J. (1999). *Affect in Language Learning*. Cambridge, UK: Cambridge University  
Press.

Aviezer, H., Hassin, R., Ryan, J., Grady, C., Susskind, J., Anderson, A., et al. (2008).

Angry, disgusted, or afraid? Studies on the malleability of emotion perception.  
*Psychological Science*, 19(7), 724-732.

Bachorowski, J., & Owren, M. (1995). Vocal Expression of Emotion - Acoustic  
Properties of Speech Are Associated with Emotional Intensity and Context.  
*Psychological Science*, 6(4), 219-224.

Baker, R., D'Mello, S., Rodrigo, M., & Graesser, A. (in review). 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*.

Baker, R., Rodrigo, M., & Xolocotzin, U. (2007). The dynamics of affective transitions in  
simulation problem-solving environments. In A. P. R. P. R. W. Paiva (Ed.), *2nd  
International Conference on Affective Computing and Intelligent Interaction* (pp.  
666-677).

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal  
of Personality and Social Psychology*, 70(3), 614-636.

Banziger, T., & Scherer, K. (2007, Sep 12-14). *Using actor portrayals to systematically  
study multimodal emotion expression: The GEMEP corpus*. Paper presented at the

2nd International Conference on Affective Computing and Intelligent Interaction,  
Lisbon, PORTUGAL.

- Barrett, L. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science* 1, 28-58.
- Batliner, A., Fischer, K., Huber, R., Spilker, J., & Noth, E. (2003). How to find trouble in communication. *Speech Communication*, 40(1-2), 117-143.
- Batson, C., Turk, C., Shaw, L., & Klein, T. (1995). Information Function of Empathic Emotion - Learning That We Value the Others Welfare. *Journal of Personality and Social Psychology*, 68(2), 300-313.
- Bellmann, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ.: Princeton University Press.
- Bernstein, N. (1967). *The co-ordination and regulation of movement*. London: Pergamon Press.
- Birdwhistell, R. (1975). Background Considerations To The Study Of The Body As A Medium Of Expression. In J. Benthall & T. Polhemud (Eds.), *The Body As A Medium Of Expression* (pp. 36–58). Bungay, Suffolk: Allen.
- Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6), 4-16.
- Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007, 2005). *How emotion is made and measured*. Paper presented at the Conference on Human Factors in Computing Systems (CHI 2005), Portland, OR.

- Boersma, P., & Weenink, D. (2006). Praat: doing phonetics by computer (Version 4.3.14).
- Boone, R., & Cunningham, J. (1998). Children's decoding of emotion in expressive body movement: The development of cue attunement. *Developmental Psychology, 34*(5), 1007-1016.
- Boone, R., & Cunningham, J. (2001). Children's expression of emotional meaning in music through expressive body movement. *Journal of Nonverbal Behavior, 25*(1), 21-41.
- Bower, G. (1981). Mood and memory. *American Psychologist, 36*, 129-148.
- Bower, G. (1992). How Might Emotions Affect Learning. In S. A. Christianson (Ed.), *The Handbook of Emotion and Memory: Research and Theory* (pp. 3-31). Hillsdale, NJ: Erlbaum.
- Bull, P. (1987). *Posture and Gesture*. Oxford Pergamon Press.
- Burleson, W., & Picard, R. (2007). Evidence for Gender Specific Approaches to the Development of Emotionally Intelligent Learning Companions. *IEEE Intelligent Systems, 22*(4), 62-69.
- Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Carberry, S., Schroeder, L., & Lambert, L. (2002). Toward recognizing and conveying an attitude of doubt via natural language. *Applied Artificial Intelligence, 16*(7-8), 495-517.

- Chen, L., Huang, T., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proceedings. Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 366-371).
- Chi, M., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. (2001). Learning from human tutoring. *Cognitive Science*, 25(4), 471-533.
- Chipman, P., Olney, A., & Graesser, A. (2006). The AutoTutor 3 architecture a software architecture for an expandable, high-availability ITS. In J. C. J. P. V. Filipe (Ed.), *Proceedings of 2nd International Conference on Web Information Systems and Technologies* (pp. 323-332). Setubal, Portugal.
- Chovil, N. (1991). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25, 163-194.
- Clark, A. (1997). *Being There: Putting Brain Body And World Together Again*. Cambridge, MA: MIT Press.
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Cohen, P., Cohen, J., West, S., & Aiken, L. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.): Taylor & Francis, Inc.
- Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, 16(7-8), 555-575.

- Conati, C., & Maclaren, H. (in press). Empirically Building and Evaluating a Probabilistic Model of User Affect. *User Modeling and User-Adapted Interaction*.
- Conati, C., & Zhou, X. (2004). *A Probabilistic Framework for Recognizing and Affecting Emotions*. Paper presented at the AAAI 2004 Spring Symposium on Architectures for Modeling Emotions.
- Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, 28(2), 117-139.
- Craig, S., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive-affective states during learning. *Cognition & Emotion*, 22(5), 777-788.
- Craig, S., D'Mello, S., Witherspoon, A., Sullins, J., & Graesser, A. (2004). Emotions During Learning: The First Step Toward An Affect Sensitive Intelligent Tutoring System. In *Proceedings of the International Conference on eLearning* (pp. 284-288).
- Craig, S., Graesser, A., Sullins, J., & Gholson, J. (2004). Affect and learning: An exploratory look into the role of affect in learning. *Journal of Educational Media*, 29, 241-250.
- Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper and Row.

- D'Mello, S., Craig, S., A., W., J., S., B., M., Gholson, B., et al. (2005). The relationship between affective states and dialog patterns during interactions with AutoTutor. In *Proceedings of the World Conference on E-learning in Corporate, Government, Health Care, and Higher Education* (pp. 2004-2011). Chesapeake, VA: Association for the Advancement of Computing in Education.
- D'Mello, S., Craig, S., Gholson, B., Franklin, S., Picard, R., & Graesser, A. (2005). Integrating Affect Sensors In An Intelligent Tutoring System. In *The Computer In The Affective Loop Workshop At 2005 International Conference On Intelligent User Interfaces* (pp. 7-13).
- D'Mello, S., Craig, S., Sullins, J., & Graesser, A. (2006). Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. *International Journal of Artificial Intelligence in Education*, 16(1), 3-28.
- D'Mello, S., Craig, S., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, 18(1-2), 45-80.
- D'Mello, S., & Graesser, A. (2006, Aug 21-23). *Affect detection from human-computer dialogue with an intelligent tutoring system*. Paper presented at the 6th International Conference on Intelligent Virtual Agents, Marina Del Rey, CA.
- D'Mello, S., & Graesser, A. (2009). Automatic Detection of Learners' Affect from Gross Body Language. 23(2), 123 - 150.

D'Mello, S., Jackson, G., Craig, S., Morgan, B., Chipman, P., White, H., et al. (2008).

*AutoTutor Detects and Responds to Learners Affective and Cognitive States.*

Paper presented at the Workshop on Emotional and Cognitive issues in ITS (WECITS) held in conjunction with Ninth International Conference on Intelligent Tutoring Systems.

D'Mello, S., King, B., Entezari, O., Chipman, P., & Graesser, A. (2008). *The Impact of Automatic Speech Recognition Errors on Learning Gains with AutoTutor*. Paper presented at the Annual meeting of the American Educational Research Association.

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

D'Mello, S., Picard, R., & Graesser, A. (2007). Towards an Affect-Sensitive AutoTutor. *Intelligent Systems, IEEE*, 22(4), 53-61.

D'Mello, S., Taylor, R., & Graesser, A. (2007). Monitoring affective trajectories during complex learning. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 203-208). Austin, TX: Cognitive Science Society.

D'Mello, S., Chipman, P., & Graesser, A. (2007). Posture as a predictor of learner's affective engagement. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 905-991). Austin, TX: Cognitive Science Society.

D'Mello, S., Craig, S., Fike, K., & Graesser, A. (in press). *Responding to Learners' Cognitive-Affective States with Supportive and Shakeup Dialogues*. Paper presented at the Proceedings of 11th International Conference on Human-

Computer Interaction, San Diego, CA.

D'Mello, S., Dale, R., & Graesser, A. (in review). Disequilibrium in the Mind,

Disharmony in the Body.

D'Mello, S., Dowell, N., & Graesser, A. (in press). Cohesion Relationships in Tutorial

Dialogue as Predictors of Affective States. In *Proceedings of 14th International*

*Conference on Artificial Intelligence In Education*.

D'Mello, S., & Graesser, A. (in review). The Half-Life of Emotions.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John

Murray.

Dasarathy, B. (1997). Sensor fusion potential exploitation: Innovative architectures and

illustrative approaches. *Proceedings IEEE*, 85, 24-38.

de Vega, M. (2002). Del Significado Simbólico Al Significado Corpóreo. [From

Symbolic Meaning To Embodied Meaning]. *Estudios De Psicología*, 23, 153-174.

De Vicente, A., & Pain, H. (2002). Informing the detection of the students' motivational

state: An empirical study. In S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.),

*6th International Conference on Intelligent Tutoring Systems* (pp. 933-943). San

Sebastian, Spain.

Demeijer, M. (1989). The Contribution of General Features of Body Movement to the

Attribution of Emotions. *Journal of Nonverbal Behavior*, 13(4), 247-268.

- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1), 1-38.
- deVega, M., Glenberg, A., & Graesser, A. (Eds.). (2008). *Symbols, embodiment, and meaning*. Oxford: Oxford University Press.
- Dodds, P., & Fletcher, J. (2004). Opportunities for new "smart" learning environments enabled by next-generation web capabilities. *Journal of Educational Multimedia and Hypermedia*, 13(4), 391-404.
- Dweck, C. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 61-87). Orlando, FL: Academic Press.
- Ekman, P. (1964). Body Position, Facial Expression And Verbal Behavior During Interviews. *Journal of Abnormal and Social Psychology*, 68(3), 295-301.
- Ekman, P. (1965a). Communication Through Nonverbal Behavior: A Source Of Information About An Interpersonal Relationship. In S. S. Tomkins & C. E. Izard (Eds.), *Affect, Cognition and Personality* (pp. 390-442). New York: Springer.
- Ekman, P. (1965b). Differential Communication of Affect by Head and Body Cues. *Journal of Personality and Social Psychology*, 2(5), 726-735.
- Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 319-344). Hillsdale, NJ: Erlbaum.

Ekman, P. (1992). An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4), 169-200.

Ekman, P. (2002, Nov 16-17). *Darwin, deception, and facial expression*. Paper presented at the Conference on Emotions Inside Out, 130 Years after Darwins the Expression of the Emotions in Man and Animals, New York, NY.

Ekman, P. (2003). *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Henry Holt and Company, LLC.

Ekman, P., & Friesen, W. (1968). Nonverbal Behavior In Psychotherapy Research. In J. Shlien (Ed.), *Research In Psychotherapy Vol. III* (pp. 179-216): American Psychological Association.

Ekman, P., & Friesen, W. (1969). Nonverbal Leakage and Clues to Deception. *Psychiatry*, 32(1), 88-&.

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

Ekman, P., & Friesen, W. (1978). *The Facial Action Coding System: A Technique For The Measurement Of Facial Movement*. Palo Alto: Consulting Psychologists Press.

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

- el Kaliouby, R., & Robinson, P. (2005, Oct 22-24). *Generalization of a vision-based computational model of mind-reading*. Paper presented at the 1st International Conference on Affective Computing and Intelligent Interaction, Beijing, China.
- Elfenbein, H., & Ambady, N. (2002a). Is there an ingroup advantage in emotion recognition? *Psychological Bulletin, 128*, 243-249.
- Elfenbein, H., & Ambady, N. (2002b). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203-235.
- Feldman, R., & Rime', B. (Eds.). (1991). *Fundamentals of nonverbal behavior*. Cambridge, England: Cambridge University Press.
- Fernandez, R., & Picard, R. (2005). *Classical and Novel Discriminant Features for Affect Recognition from Speech*. Paper presented at the 9th European Conference on Speech Communication and Technology.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Fiske, D., & Campbell, D. (1987). Citation-Classic - Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Current Contents/Social & Behavioral Sciences*(14), 14-14.
- Fleiss, J. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: John Wiley & Son.
- Forbes-Riley, K., & Litman, D. (2004). *Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources*. Paper presented at the Human Language

Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL).

Forbes-Riley, K., & Litman, D. (2007). *Investigating Human Tutor Responses to Student Uncertainty for Adaptive System Development*. Paper presented at the Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction.

Forbes-Riley, K., & Litman, D. (2009). *Adapting to Student Uncertainty Improves Tutoring Dialogues*. Paper presented at the Proceedings of the 14th International Conference on Artificial Intelligence in Education.

Forbes-Riley, K., Rotaru, M., & Litman, D. (2008). The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction*, 18(1-2), 11-43.

Fredrickson, B., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition & Emotion*, 19(3), 313-332.

Furnham, A. (1999). *Body Language At Work*. Channel Islands: The Guernsey Press.

Gertner, A., & VanLehn, K. (2000). *Andes: A coached problem solving environment for physics*. Paper presented at the Intelligent Tutoring Systems, Proceedings.

Gill, A., French, R., Gergle, D., & Oberlander, J. (2008). Identifying Emotional Characteristics from Short Blog Texts. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *30th Annual Conference of the Cognitive Science Society* (pp. 2237-2242). Washington, DC: Cognitive Science Society.

- Glenberg, A., Havas, D., Becker, R., & Rinck, M. (in press). Grounding Language in Bodily States: The Case for Emotion. In R. Zwaan & D. Pecher (Eds.), *The Grounding Of Cognition: The Role Of Perception And Action In Memory, Language, And Thinking*. Cambridge: Cambridge University Press.
- Gorin, A., Riccardi, G., & Wright, J. (1996, Sep 30-Oct 01). *How may I help you?* Paper presented at the 3rd IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA-96), Basking Ridge, New Jersey.
- Graesser, A., Chipman, P., Haynes, B., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4), 612-618.
- Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. (2007). Emotions and Learning with AutoTutor. In R. Luckin, K. Koedinger & J. Greer (Eds.), *13th International Conference on Artificial Intelligence in Education* (pp. 569-571): IOS Press.
- Graesser, A., & D'Mello, S. (in preparation). Emotions During Complex Learning.
- Graesser, A., D'Mello, S., Craig, S., A., W., J., S., B., M., et al. (2008). The Relationship between Affective States and Dialog Patterns during Interactions with AutoTutor. *Journal of Interactive Learning Research*, 19(2), 293-312.
- Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (2003, Nov 06). *AutoTutor: A tutor with dialogue in natural language*. Paper presented at the 33rd Annual Meeting of the Society-for-Computers-in-Psychology, Vancouver, CANADA.

Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (2004a).

AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36, 180-193.

Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (2004b).

AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods Instruments & Computers*, 36(2), 180-192.

Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B.

(2006). *Detection of Emotions during learning with AutoTutor*. Paper presented at the Proceedings of the 28th Annual Conference of the Cognitive Science Society, Mahwah, NJ:

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36, 193-202.

Graesser, A., McNamara, D., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist*, 40(4), 225-234.

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

Graesser, A., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through Mixed-initiative Dialogue in Natural Language. In

- T. Landauer, D. McNamara, S. Dennis & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 243-262). Mahwah, NJ: Erlbaum.
- Graesser, A., & Person, N. (1994). Question Asking During Tutoring. *American Education Research Journal*, 31(1), 104-137.
- Graesser, A., Person, N., & Magliano, J. (1995). Collaborative Dialogue Patterns In Naturalistic One-To-One Tutoring. *Applied Cognitive Psychology*, 9(6), 495-522.
- Graesser, A., VanLehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22(4), 39-51.
- Graesser, A., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & Group, T. R. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8, 129-148.
- Grimm, M., Mower, E., Narayanan, S., & Kroschel, K. (2006). *Combining categorical and primitives-based emotion recognition*. Paper presented at the 14th European Signal Processing Conference (EUSIPCO), Florence, Italy.
- Hancock, J., Curry, L., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1), 1-23.
- Hancock, J., Landrigan, C., & Silver, C. (2007). *Expressing emotion in text-based communication*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.
- Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: John Wiley & Sons.

- Hinde, R. (Ed.). (1972). *Non-verbal communication*. Cambridge, England: Cambridge University Press.
- Hocking, R. (1976). Analysis and Selection of Variables in Linear-Regression. *Biometrics*, 32(1), 1-49.
- Holmes, N., & Spence, C. (2005). Multisensory integration: Space, time and superadditivity. *Current Biology*, 15(18), R762-R764.
- Hudlicka, E., & McNeese, M. (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), 1-47.
- Issroff, K., & del Soldato, T. (1996). *Incorporating motivation into computer-supported collaborative learning*. Paper presented at the Proceedings of European conference on artificial intelligence in education.
- Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1-2), 116-134.
- Johnstone, T., & Scherer, K. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland-Jones (Eds.), *Handbook of Emotions* (2nd ed., pp. 220-235). New York: Guilford Press.
- Jonassen, D., Peck, K., & Wilson, B. (1999). *Learning with technology: A constructivist perspective*. Upper Saddle River, NJ: Prentice Hall.
- Kapoor, A., & Picard, R. (2005). *Multimodal affect recognition in learning environments*. Paper presented at the Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore.

- Kello, C., Anderson, G., Holden, J., & Van Orden, G. (2008). The Pervasiveness of 1/f Scaling in Speech Reflects the Metastable Basis of Cognition. *Cognitive Science*, 32(7), 1217-1231.
- Keltner, D., & Ekman, P. (2000). Facial expression of emotion. In R. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (Vol. 2nd ed, pp. 236–264). New York: Guilford.
- Keltner, D., & Haidt, J. (2001). Social functions of emotions. In T. J. Mayne & G. A. Bonanno (Eds.), *Emotions: Current issues and future directions* (pp. 192–213). New York: Guilford Press.
- Kim, Y. (2005). *Empathetic Virtual Peers Enhanced Learner Interest and Self-Efficacy*. Paper presented at the Workshop on Motivation and Affect in Educational Software at the 12th International Conference on Artificial Intelligence in Education, Amsterdam, Netherlands.
- Klecka, W. (1980). *Discriminant Analysis*. Beverly Hills, CA: Sage.
- Klein, J., Moon, Y., & Picard, R. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2), 119-140.
- Knapp, M., & Hall, J. (1997). *Nonverbal communication in human interaction*. New York: Harcourt Brace Jovanovich.
- Koedinger, K., Anderson, J., Hadley, W., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(30-43).

- Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61-78). New York, NY: Cambridge University Press.
- Kort, B., Reilly, R., & Picard, R. (2001). *An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion*. Paper presented at the Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on.
- Landauer, T., & Dumais, S. (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., McNamara, D., Dennis, S., & Kintsch, W. (Eds.). (2008). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Erlbaum.
- Lazarus, R. (1991). *Emotion and adaptation*. New York: Oxford University Press.
- Lee, C., & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2), 293-303.
- Lehman, B., D'Mello, S., & Person, N. (2008). *All Alone with your Emotions: An Analysis of Student Emotions during Effortful Problem Solving Activities*. Paper presented at the Workshop on Emotional and Cognitive issues in ITS at the Ninth International Conference on Intelligent Tutoring Systems.
- Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What are you feeling? Investigating student affective states during expert human tutoring sessions. In B.

- Woolf, A. E., N. R. & L. S. (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 50-59).
- Lepper, M., & Chabay, R. (1988). Socializing the intelligent tutor: Bringing empathy to computer tutors. In H. Mandl & A. Lesgold (Eds.), *Learning Issues for Intelligent Tutoring Systems* (pp. 242-257). Hillsdale, NJ: Erlbaum.
- Lepper, M., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. In J. Aronson (Ed.), *Improving academic achievement: Impact of psychological factors on education* (pp. 135-158). Orlando, FL: Academic Press.
- Linnenbrink, E., & Pintrich, P. (2002). The Role Of Motivational Beliefs In Conceptual Change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 115-135). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Liscombe, J., Riccardi, G., & Hakkani-Tür, D. (2005). *Using Context to Improve Emotion Detection in Spoken Dialog Systems*. Paper presented at the 9th European Conference on Speech Communication and Technology (EUROSPEECH'05).
- Litman, D., & Forbes-Riley, K. (2004). *Predicting student emotions in computer-human tutoring dialogues*. Paper presented at the Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain.

- Litman, D., & Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5), 559-590.
- Litman, D., & Silliman, S. (2004). *ITSPOKE: An intelligent tutoring spoken dialogue system*. Paper presented at the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL) Boston, MA.
- Liu, H., Lieberman, H., & Selker, S. (2003). *A model of textual affect sensing using real-world knowledge*. Paper presented at the Proceedings of the 8th international conference on Intelligent user interfaces, Miami, Florida, USA.
- Mandelbrot, B. (1998). *Multifractals and 1/f Noise: Wild Self-Affinity in Physics*. New York: Springer.
- Mandler, G. (1976). *Mind and emotion*. New York: Wiley.
- Mandler, G. (1984a). Another theory of emotion claims too much and specifies too little. *Current Psychology of Cognition*, 4(1), 84-87.
- Mandler, G. (1984b). *Mind and Body: Psychology of Emotion and Stress*. New York: W.W. Norton & Company.
- Mandler, G. (1999). Emotion. In B. M. Bly & D. E. Rumelhart (Eds.), *Cognitive science. Handbook of perception and cognition* (2nd ed.). San Diego, CA: Academic Press.
- Mandryk, R., & Atkins, M. (2007, 2005). *A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies*. Paper

presented at the Conference on Human Factors in Computing Systems (CHI 2005), Portland, OR.

Marsic, I., Medl, A., & Flanagan, J. (2000). Natural communication with information systems. *Proceedings of the Ieee*, 88(8), 1354-1366.

Matsubara, Y., & Nagamachi, M. (1996). Motivation Systems and Motivation Models for Intelligent Tutoring. In P. o. t. T. I. C. i. I. T. Systems (Ed.), (Vol. 1086, pp. 139-147). Berlin / Heidelberg: Springer.

Mauss, I., Levenson, R., McCarter, L., Wilhelm, F., & Gross, J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, 5(2), 175-190.

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

McNamara, D., Louwerse, M., & Graesser, A. (2008). *Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Final report on Institute of Education Science grant (R305G020018)*. Memphis, TN: University of Memphis.

McNamara, D., Ozuru, Y., Graesser, A., & Louwerse, M. (2006). *Validating Coh-Metrix*. Paper presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver, BC.

- McNeese, M. (2003). New visions of human-computer interaction: making affect compute. *International Journal of Human-Computer Studies*, 59(1-2), 33-53.
- McQuiggan, S., Mott, B., & Lester, J. (2008). Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *User Modeling and User-Adapted Interaction*, 18(1-2), 81-123.
- Mehrabian, A. (1968a). Inference of Attitudes from Posture Orientation and Distance of a Communicator. *Journal of Consulting and Clinical Psychology*, 32(3), 296-308.
- Mehrabian, A. (1968b). Relationship of Attitude to Seated Posture Orientation and Distance. *Journal of Personality and Social Psychology*, 10(1), 26-&.
- Mehrabian, A. (1971). Nonverbal Betrayal Of Feelings. *Journal of Experimental Research in Personality*, 5, 64-73.
- Mehrabian, A. (1972). *Nonverbal communication*. Chicago, Illinois: Aldine-Atherton.
- Mehrabian, A., & Friar, J. (1969). Encoding of Attitude by a Seated Communicator Via Posture and Position Cues. *Journal of Consulting and Clinical Psychology*, 33(3), 330-&.
- Metcalfe, J., & Kornell, N. (2005). A Region of Proximal Learning model of study time allocation. *Journal of Memory and Language*, 52(4), 463-477.
- Meyer, D., & Turner, J. (2006). Re-conceptualizing emotion and motivation to learn in classroom contexts. *Educational Psychology Review*, 18(4), 377-390.
- Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88(2), 203-214.

- Mitchell, T. (1997). *Machine Learning*: Mc-Graw-Hill.
- Montepare, J., Koff, E., Zaitchik, D., & Albert, M. (1999). The use of body movements and gestures as cues to emotions in younger and older adults. *Journal of Nonverbal Behavior*, 23(2), 133-152.
- Morimoto, C., Koons, D., Amir, A., & Flickner, M. (1998). *Pupil Detection and Tracking using Multiple Light Sources*: IBM: Almaden Research Center.
- Mota, S., & Picard, R. (2003). Automated Posture Analysis for Detecting Learner's Interest Level. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on* (Vol. 5, pp. 49-49).
- Norman, D. (1994). How might people interact with agents. 37(7), 68-71.
- Obudho, C. (1979). *Human nonverbal behavior. An annotated bibliography*. Westport, Conn: Greenwood Press.
- Ohman, A. (2004, Jun 02-04). *The role of the amygdala in human fear: Automatic detection of threat*. Paper presented at the Meeting on Somatisation, Sensitisation and Psychosomatic Medicine held in honor of Holger Ursin, Bergen, NORWAY.
- Ohman, A., & Soares, J. (1993). On the Automatic Nature of Phobic Fear - Conditioned Electrodermal Responses to Masked Fear-Relevant Stimuli. *Journal of Abnormal Psychology*, 102(1), 121-132.
- Ohman, A., & Soares, J. (1994). Unconscious Anxiety - Phobic Responses to Masked Stimuli. *Journal of Abnormal Psychology*, 103(2), 231-240.
- Olney, A., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. (2003). *Utterance classification in AutoTutor*. Paper presented at the HLT-

NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Ortony, A., & Turner, T. (1990). What's Basic About Basic Emotions. *Psychological Review*, 97(3), 315-331.

Paiva, A., Prada, R., & Picard, R. (Eds.). (2007). *Affective Computing and Intelligent Interaction*. Heidelberg Springer.

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. New York: Oxford University Press.

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

Pantic, M., & Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9), 1370-1390.

Patrick, B., Skinner, E., & Connell, J. (1993). What Motivates Children's Behavior and Emotion - Joint Effects of Perceived Control and Autonomy in the Academic Domain. *Journal of Personality and Social Psychology*, 65(4), 781-791.

Pentland, A. (2000). Looking at people: Sensing for ubiquitous and wearable computing. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 107-119.

- Person, N., & Graesser, A. (2002, Jun 02-07). *Human or computer? AutoTutor, in a Bystander Turing Test*. Paper presented at the 6th International Conference on Intelligent Tutoring Systems, San Sebastian, Spain.
- Piaget, J. (1952). *The origins of intelligence*. New York: International University Press.
- Picard, R. (1997). *Affective Computing*. Cambridge, Mass: MIT Press.
- Picard, R., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 1175-1191.
- Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction*, 18(1-2), 125-173.
- Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, 19(3-4), 267-285.
- Rani, P., Sarkar, N., & Smith, C. (2003). *Affect-sensitive human-robot cooperation - theory and experiments*. Paper presented at the IEEE International Conference on Robotics and Automation.
- Robson, C. (1993). *Real world research: A resource for social scientist and practitioner researchers*. Oxford: Blackwell.
- Rosenthal, R., & Rosnow, R. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

- Rumelhart, D., McClelland, J., & Group, P. R. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.
- Rus, V., & Graesser, A. (2007). Lexico-Syntactic Subsumption for Textual Entailment. In N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing IV: Selected Papers from RANLP 2005* (pp. 187-196): John Benjamins Publishing Company.
- Rus, V., McCarthy, P., McNamara, D., & Graesser, A. (2008). A Study of Textual Entailment. *International Journal on Artificial Intelligence Tools*, 17(4), 659-685.
- Russell, J. (1994). Is There Universal Recognition of Emotion from Facial Expression - a Review of the Cross-Cultural Studies. *Psychological Bulletin*, 115(1), 102-141.
- Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110, 145-172.
- Schank, R. (1986). *Explanation Patterns: Understanding Mechanically and Creatively*. Hillsdale, NJ: Erlbaum.
- Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2), 227-256.
- Scherer, K., & Ellgring, H. (2007a). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7(1), 113-130.
- Scherer, K., & Ellgring, H. (2007b). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion*, 7(1), 158-171.

- Scherer, K., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer & H. Goldsmith (Eds.), *Handbook of the Affective Sciences* (pp. 433–456). New York and Oxford: Oxford University Press.
- Scherer, K., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. London: London University Press.
- Selfridge, O. (1959). Pandemonium: A Paradigm For Learning. In *Symposium on the Mechanization of Thought Processes* (pp. 511-531). London: Her Majesty's Stationery Office.
- Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Shafran, I., Riley, M., & Mohri, M. (2003). *Voice signatures*. Paper presented at the IEEE Workshop on Automatic Speech Recognition and Understanding.
- Shah, F., Evens, M., Michael, J., & Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, 33(1), 23-52.
- Shaikh, M., Prendinger, H., & Ishizuka, M. (2007). An analytical approach to assess sentiment of text (Ieee, Trans.). In *Proceedings of 10th International Conference on Computer and Information Technology* (pp. 63-68): IEEE.
- Shaikh, M., Prendinger, H., & Ishizuka, M. (2008). Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence*, 22(6), 558-601.

- Shneiderman, B., & Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction*. Reading, MA: Addison-Wesley.
- Silvia, P., & Abele, A. (2002). Can positive affect induce self-focused attention? Methodological and measurement issues. *Cognition & Emotion*, 16(6), 845-853.
- Smith, C., & Ellsworth, P. (1985). Patterns of Cognitive Appraisal in Emotion. *Journal of Personality and Social Psychology*, 48(4), 813-838.
- Stein, N., & Hernandez, M. (2007). Assessing understanding and appraisals during emotional experience: The development and use of the Narcoder. In J. A. Coan & J. J. Allen (Eds.), *Handbook of emotion elicitation and assessment* (pp. 298-317). New York: Oxford University Press.
- Stein, N., Hernandez, M., & Trabasso, T. (2008). Advances in modeling emotions and thought: The importance of developmental, online, and multilevel analysis. In M. Lewis, J. M. Haviland-Jones & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 574-586). New York: Guilford Press.
- Stein, N., & Levine, L. (1991). Making sense out of emotion. In A. O. W. Kessen, & F. Kraik (Eds.) (Ed.), *Memories, thoughts, and emotions: Essays in honor of George Mandler* (pp. 295-322). Hillsdale, NJ: Erlbaum.
- Stemmler, G., Heldmann, M., Pauls, C., & Scherer, T. (2001). Constraints for emotion specificity in fear and anger: The context counts. *Psychophysiology*, 38(2), 275-291.

- Storey, J., Kopp, K., Wiemer, K., Chipman, P., & Graesser, A. (in press). Critical thinking tutor: Using AutoTutor to teach scientific critical thinking skills. *Behavioral Research Methods*.
- Sylwester, R. (1994). How Emotions Affect Learning. *Educational Leadership*, 52(2), 60-65.
- Tan, H., Lu, I., & Pentland, A. (1997). *The chair as a novel haptic user interface*. Paper presented at the Proceedings of the Workshop on Perceptual User Interfaces.
- Tan, H., Slivovsky, L., & Pentland, A. (2001). A sensing chair using pressure distribution sensors. *Ieee-Asme Transactions on Mechatronics*, 6(3), 261-268.
- Tekscan. (1997). *Body Pressure Measurement System User's Manual*. South Boston, MA: Tekscan Inc.
- Turner, T., & Ortony, A. (1992). Basic Emotions - Can Conflicting Criteria Converge. *Psychological Review*, 99(3), 566-571.
- Van Orden, G., Holden, J., & Turvey, M. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology-General*, 132(3), 331-350.
- VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.
- VanLehn, K., Graesser, A., Jackson, G., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1), 3-62.
- VanLehn, K., Jordan, P., Rose, C., Bhembe, D., Bottner, M., & A., G. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In S.

- A. Cerri, G. Gouarderes & F. Paraguacu (Eds.), *Proceedings of the Sixth International Conference on Intelligent Tutoring* (pp. 158-167). Berlin: Springer-Verlag.
- VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., et al. (2005). The Andes physics tutoring system: five years of evaluations. *International Journal of Artificial Intelligence in Education* 15, 147-204.
- Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 511-518): IEEE.
- Walk, R., & Homan, C. (1984). Emotion and Dance in Dynamic Light Displays. *Bulletin of the Psychonomic Society*, 22(5), 437-440.
- Walk, R., & Walters, K. (1988). Perception of the Smile and Other Emotions of the Body and Face at Different Distances. *Bulletin of the Psychonomic Society*, 26(6), 510-510.
- Walker, M., Langkilde-Geary, I., Hastie, H., & Gorin, A. (2002). Automatically training a problematic dialogue predictor for a spoken dialogue system. *Journal of Artificial Intelligence Research*, 16, 293-319.
- Wallbott, H. (1998). Bodily expression of emotion. *European Journal of Social Psychology*, 28(6), 879-896.
- Walters, K., & Walk, R. (1986). Perception of Emotion from Body Posture. *Bulletin of the Psychonomic Society*, 24(5), 329-329.

- Walters, K., & Walk, R. (1988). Perception of Emotion from Moving Body Cues in Photographs. *Bulletin of the Psychonomic Society*, 26(2), 112-114.
- Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.
- Whang, M., Lim, J., & Boucsein, W. (2003). Preparing computers for affective communication: A psychophysiological concept and preliminary results. *Human Factors*, 45(4), 623-634.
- Wiemer-Hastings, P., Graesser, A., & Harter, D. (1998). The foundations and architecture of AutoTutor. In B. Goettl, H. H., R. C. & S. V. (Eds.), *4th International Conference on Intelligent Tutoring Systems (ITS 98)* (pp. 334-343). San Antonio, Texas.
- Wiemer-Hastings, P., Wiemer-Hastings, K., & Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with latent semantic analysis. In S. P. Lajoie & M. Vivet (Eds.), *Proceedings of International Conference on Artificial Intelligence in Education* (pp. 535-542). Amsterdam: IOS Press.
- Witten, I., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.
- Woolf, B., Burleson, W., & Arroyo, I. (2007). *Emotional Intelligence for Computer Tutors*. Paper presented at the Workshop on Modeling and Scaffolding Affective Experiences to Impact Learning at 13th International Conference on Artificial Intelligence in Education, Los Angeles, USA.

- Yoshitomi, Y., K., S.-I., Kawano, T., & Kilazoe, T. (2000). Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In P. o. t. I. I. W. o. R. a. H. I. Communication (Ed.), (pp. 178 - 183): IEEE.
- Zeng, Z., Pantic, M., Roisman, G., & Huang, T. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 31(1), 39-58.

## **Appendices**

## Appendix A. Affect Measurement by Humans

Table 20. Mean difference in kappa scores for judge  $\times$  judgment type interaction.

JP	JP	SP		S1		S2		P1		P2		12	
		M	M	M	M	M	M	M	M	V	V	V	V
SP	M	.00	-.05	-.07	-.06	-.09	-.25	-.07	-.25	-.18	-.31	-.31	-.65
S1	M	.05	.00	-.02	-.01	-.04	-.21	-.02	-.20	-.13	-.26	-.26	-.60
S2	M	.07	.02	.00	.01	-.02	-.19	.00	-.18	-.11	-.24	-.24	-.58
P1	M	.06	.01	-.01	.00	-.03	-.20	-.01	-.19	-.12	-.25	-.25	-.59
P2	M	.09	.04	.02	.03	.00	-.16	.02	-.16	-.09	-.21	-.22	-.56
12	M	.25	.21	.19	.20	.16	.00	.19	.01	.08	.05	.06	-.40
SP	V	.07	.02	.00	.01	-.02	-.19	.00	-.18	-.11	-.24	-.24	-.58
S1	V	.25	.20	.18	.19	.16	-.01	.18	.00	.07	-.06	-.06	-.40
S2	V	.18	.13	.11	.12	.09	-.08	.11	-.07	.00	-.13	-.13	-.47
P1	V	.31	.26	.24	.25	.21	.05	.24	.06	.13	.00	.00	-.34
P2	V	.31	.26	.24	.25	.22	.06	.24	.06	.13	.00	.00	-.34
12	V	.65	.60	.58	.59	.56	.40	.58	.40	.47	.34	.34	.00

JP = Judge Pairs  
 SP = Self-Peer  
 JT = Judgment type      S1 = Self-Judge1  
 M = Mandatory

S2 = Self-Judge2  
 V = Voluntary

P1 = Peer-Judge1

P2 = Peer-Judge2

12 = Judge1-Judge2

**Table 21. Mean difference in kappa scores for mandatory judgments.**

JP	SP	SP	SP	SP	SP	S1	S1	S1	S1	S2	S2	S2	P1	P1	P1	P2	P2	P2	I2	I2	I2	I2								
SP	B	-.03	.04	.03	.03	-.04	-.09	-.01	.01	-.07	-.12	-.04	.06	-.02	-.12	-.07	-.02	-.01	-.19	-.09	-.02	-.03	-.19	-.29	-.23	-.04	-.20			
SP	C	.03	.00	.07	.06	.06	.01	-.06	.02	.04	-.10	-.01	.09	.01	-.09	-.04	.01	.02	.02	-.16	-.06	.01	.01	-.16	-.26	-.20	-.01	-.17		
SP	L	-.04	-.07	.00	-.01	-.08	-.13	-.05	-.03	-.05	-.10	-.16	-.08	.02	-.06	-.15	-.11	-.05	-.04	-.23	-.13	-.06	-.05	-.07	-.22	-.33	-.27	-.07	-.24	
SP	F	-.03	-.06	.00	.00	-.08	-.13	-.04	-.03	-.05	-.10	-.16	-.07	.03	-.06	-.15	-.10	-.05	-.04	-.22	-.12	-.06	-.05	-.07	-.22	-.33	-.27	-.07	-.24	
SP	N	-.03	-.06	.01	.00	-.08	-.12	-.04	-.02	-.05	-.10	-.16	-.07	.03	-.05	-.15	-.10	-.05	-.04	-.22	-.12	-.06	-.05	-.07	-.22	-.33	-.26	-.07	-.23	
S1	B	.04	.01	.08	.08	.00	-.05	.03	.05	.03	-.02	-.08	.01	.10	.02	-.07	-.02	.03	.04	.04	.15	-.05	.02	.03	.01	-.14	-.25	-.19	.01	-.16
S1	C	.09	.06	.13	.12	.05	.00	.08	.10	.08	.03	.03	.05	.15	.07	-.02	.02	.08	.08	.09	.10	.00	.07	.08	.06	-.20	-.14	.06	-.11	-.11
S1	L	.01	-.02	.05	.04	.04	-.03	-.08	.00	.02	.00	-.06	-.12	-.03	.07	-.01	-.11	-.06	-.01	.00	.18	-.08	-.01	-.01	-.03	-.18	-.28	-.22	-.03	-.19
S1	F	-.01	-.04	.03	.02	-.05	-.10	-.02	.00	-.02	-.07	-.13	-.05	.05	-.03	-.12	-.08	-.02	-.01	-.20	-.10	-.03	-.02	-.04	-.19	-.30	-.24	-.04	-.21	
S1	N	.01	-.02	.05	.05	.05	-.03	-.08	.00	.02	-.05	-.11	-.02	.07	-.01	-.10	-.05	.00	.01	.18	-.08	-.01	.00	.02	-.17	-.28	-.22	-.02	-.19	
S2	B	.07	.04	.10	.10	.02	-.03	.06	.07	.05	.00	-.06	.03	.13	.04	-.05	-.05	.05	.06	.12	-.02	.04	.05	.03	-.12	-.23	-.17	.03	-.14	
S2	C	.12	.10	.16	.16	.08	.03	.12	.13	.11	.06	.00	.09	.19	.10	.01	.06	.11	.12	.12	-.07	.04	.10	.11	.09	-.06	-.17	-.11	.09	-.08
S2	L	.04	.01	.08	.07	.07	-.01	-.05	.03	.05	.02	-.03	-.09	.00	.10	.02	-.08	-.03	.03	.15	-.05	.02	.02	.00	-.15	-.25	-.19	.00	-.16	
S2	F	-.06	-.09	-.02	-.03	-.03	-.10	-.15	-.07	-.05	-.07	-.13	-.19	-.10	.00	-.08	-.18	-.13	-.08	-.07	-.25	-.15	-.08	-.10	-.25	-.35	-.29	-.10	-.26	
S2	N	.02	-.01	.06	.06	.05	-.02	-.07	.01	-.03	.01	-.04	-.10	-.02	.08	.00	-.09	-.05	.01	.01	.02	-.17	-.07	.00	.01	-.16	-.27	-.21	-.01	-.18
P1	B	.12	.09	.15	.15	.07	.02	.11	.12	.10	.05	-.01	.08	.18	.09	.00	.05	.10	.11	-.07	.03	.09	.10	.08	-.07	-.18	-.11	.08	-.08	
P1	C	.07	.04	.11	.10	.10	.02	-.02	.06	.08	.05	.00	-.06	.03	.13	.05	-.05	.00	.05	.06	-.12	.02	.05	.05	.03	-.12	-.22	-.16	.03	-.13
P1	L	.02	-.01	.05	.05	.05	-.03	-.08	.01	.02	.00	-.05	-.11	-.02	.08	-.01	-.10	-.05	.00	.01	.01	.17	-.07	-.01	.00	-.02	-.21	-.21	-.02	-.18
P1	F	.01	-.02	.05	.04	.04	-.04	-.08	.00	.02	-.01	-.06	-.12	-.03	.07	-.01	-.11	-.06	-.01	.00	-.18	-.08	-.01	-.03	-.18	-.28	-.22	-.03	-.19	
P1	N	.01	-.02	.04	.04	.04	-.04	-.09	.00	.01	-.01	-.06	-.12	-.03	.07	-.02	-.11	-.06	-.01	.00	-.18	-.08	-.02	-.01	-.03	-.18	-.29	-.22	-.03	-.19
P2	B	.19	.16	.23	.22	.15	.10	.18	.20	.18	.12	.07	.15	.25	.17	.07	.12	.17	.18	.18	.00	.10	.17	.17	.16	.00	-.10	-.04	.15	-.01
P2	C	.09	.06	.13	.12	.05	.00	.08	.10	.08	.02	-.04	.05	.15	.07	-.03	.02	.07	.08	.08	-.10	.00	.07	.07	.06	-.10	-.20	-.14	.05	-.11
P2	L	.02	-.01	.06	.06	.06	-.02	-.07	.01	.03	.01	-.04	-.10	-.02	.08	-.00	-.09	-.05	.01	.01	.02	.17	-.07	.00	.01	-.16	-.27	-.21	-.01	-.18
P2	F	.02	-.01	.05	.05	.05	-.03	-.08	.01	-.02	.00	-.05	-.11	-.02	.08	-.01	-.10	-.05	.00	.01	.01	.17	-.07	-.01	.00	-.02	-.28	-.22	-.02	-.19
P2	N	.03	.01	.07	.07	-.01	-.06	.03	.04	.02	-.03	-.09	.00	.10	.01	-.08	-.03	.02	.03	.03	.16	-.06	.01	.02	.00	-.15	-.26	-.20	.00	-.17
I2	B	.19	.16	.22	.22	.14	.09	.18	.19	.17	.12	.06	.15	.25	.16	.07	.12	.17	.18	.18	.00	.10	.16	.17	.15	.00	-.11	-.05	.15	-.01
I2	C	.29	.26	.33	.33	.25	.20	.28	.30	.28	.23	.17	.25	.35	.27	.18	.22	.28	.29	.10	.20	.27	.28	.26	.11	.00	.06	.26	.09	
I2	L	.23	.20	.27	.27	.26	.19	.14	.22	.24	.22	.17	.11	.19	.29	.21	.11	.16	.21	.22	.04	.14	.21	.22	.20	.05	-.06	.00	.20	.03
I2	F	.04	.01	.07	.07	-.01	-.06	.03	.04	.02	-.03	-.09	.00	.10	.01	-.08	-.03	.02	.03	.03	-.15	-.05	.01	.02	.00	-.15	-.26	-.20	.00	-.16
I2	N	.20	.17	.24	.24	.23	.16	.11	.19	.21	.14	.08	.16	.26	.18	.08	.13	.18	.19	.01	.11	.18	.19	.17	.01	-.09	-.03	.16	.00	

Table 22. Mean difference in kappa scores for voluntary judgments.

	JP	SP	SP	SP	S1	S1	S2	S2	P1	P1	P2	P2	P2	12	12
	JP	E	C	D	C	D	C	D	C	D	C	D	C	D	F
SP	C	.00	-.03	.33	-.21	.13	.13	-.16	.09	.31	-.11	.00	.01	-.19	.04
SP	D	.03	.00	.36	-.18	.17	.16	-.13	.12	.35	-.07	.03	.04	-.16	.08
SP	F	-.33	-.36	.00	-.54	-.19	-.20	-.49	-.24	-.01	-.43	-.33	-.32	-.52	-.28
S1	C	.21	.18	.54	.00	.35	.34	.05	.30	.52	.11	.21	.22	.02	.26
S1	D	-.13	-.17	.19	-.35	.00	-.01	-.30	-.05	.18	-.24	-.14	-.13	-.33	-.09
S1	F	-.13	-.16	.20	-.34	.01	.00	-.29	-.04	.18	-.23	-.13	-.12	-.32	-.08
S2	C	.16	.13	.49	-.05	.30	.29	.00	.25	.48	.06	.16	.17	-.03	.21
S2	D	-.09	-.12	.24	-.30	.05	.04	-.25	.00	.22	-.19	-.09	-.08	-.28	-.04
S2	F	-.31	-.35	.01	-.52	-.18	-.18	-.48	-.22	.00	-.42	-.31	-.30	-.51	-.27
P1	C	.11	.07	.43	-.11	.24	.23	-.06	.19	.42	.00	.10	.11	-.09	.15
P1	D	.00	-.03	.33	-.21	.14	.13	-.16	.09	.31	-.10	.00	.01	-.19	.05
P1	F	-.01	-.04	.32	-.22	.13	.12	-.17	.08	.30	-.11	-.01	.00	-.20	.04
P2	C	.19	.16	.52	-.02	.33	.32	.03	.28	.51	.09	.19	.20	.00	.24
P2	D	-.04	-.08	.28	-.26	.09	.08	-.21	.04	.27	-.15	-.05	-.04	-.24	.00
P2	F	-.11	-.15	.21	-.33	.02	.01	-.28	-.03	.20	-.22	-.12	-.11	-.31	-.07
12	C	.36	.33	.69	.15	.49	.49	.20	.45	.67	.25	.36	.37	.17	.40
12	D	.38	.35	.71	.17	.52	.51	.22	.47	.69	.28	.38	.39	.19	.43
12	F	.11	.08	.44	-.10	.25	.24	-.05	.20	.43	.01	.11	.12	-.08	.16

Notes.

JP = Judge Pairs

SP = Self-Peer

E = Emotions

S2 = Self-Judge2

D = Delight

P1 = Peer-Judge1

F = Frustration

P2 = Peer-Judge2

12 = Judge1-Judge2

## Appendix B. Affect Classification from Conversational Cues

*Detailed classification results for 4-way emotion discriminations*

Table 23. Classification accuracies for 4-way discrimination.

Classifier Type	Classifier Name	Classification Accuracy Boredom, Confusion, Flow, and Frustration						Mean
		Self	Peer	Judge 1	Judge 2	Judges Agree	2 Agree	
Bayesian	Naive Bayes	33.2	37.1	45.4	46.3	47.4	49.0	49.2
	Naive Bayes Updatable	33.5	37.7	44.6	46.4	46.8	48.8	49.1
Functions	Logistic Regression	35.1	38.4	50.0	50.4	54.0	51.5	52.5
	Multilayer Perceptron	32.3	36.3	44.9	46.6	47.2	49.2	48.8
	Support Vector Machines	34.5	38.2	48.1	49.0	48.4	51.2	50.5
Instance	Nearest Neighbor	28.7	31.0	40.0	40.2	42.7	43.0	39.4
	K*	29.1	35.1	31.1	39.5	31.4	30.3	31.3
	Locally Weighted Learning	32.5	37.3	45.0	44.2	47.0	44.8	43.2
Meta	AdaBoost	28.5	31.3	41.8	41.9	42.7	38.3	41.5
	Bagging Predictors	32.1	37.0	46.2	47.9	48.8	49.2	47.6
	Additive Logistic Regression	34.1	38.8	45.9	48.6	49.9	52.7	49.6
Rules	Decision Tables	27.2	36.7	42.2	47.0	48.8	44.8	46.6
	Nearest Neighbor Gen.	29.2	31.6	41.6	41.3	44.8	46.1	42.5
	PART	30.2	34.6	41.2	41.9	43.3	46.3	44.5
Trees	C4.5 Decision Tree	28.6	36.3	41.6	43.5	45.2	43.5	43.4
	Logistic Model Trees	34.3	37.0	50.5	49.9	52.9	50.3	51.7
	REP Tree	30.3	36.7	42.2	44.5	46.2	44.9	44.2
Mean		31.4	35.9	43.7	45.2	46.3	46.1	45.6

*Detailed classification results for affect-neutral discriminations*

**Table 24.** Classification accuracies in discriminating between boredom and neutral.

Classifier Type	Classifier Name	Classification Accuracy						Mean
		Boredom and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree	
Self	Peer							Mean
Bayesian	Naive Bayes	58.1	57.7	58.6	61.2	61.3	65.6	60.6
	Naive Bayes Updatable	57.9	58.2	59.2	61.8	62.4	66.3	60.7
Functions	Logistic Regression	58.6	59.2	59.4	63.9	63.2	67.6	61.4
	MultiLayer Perceptron	58.4	59.3	56.5	60.0	61.7	67.7	61.3
	Support Vector Machines	61.3	58.8	59.5	61.7	62.0	65.3	61.4
Instance	Nearest Neighbor	54.0	58.5	61.2	60.8	67.2	69.0	61.6
	K*	54.7	59.5	59.2	62.6	65.6	65.9	62.7
	Locally Weighted Learning	52.0	55.5	58.5	60.6	64.4	62.7	61.4
Meta	AdaBoost	53.6	59.8	56.0	56.2	61.7	65.0	50.7
	Bagging Predictors	56.0	58.8	57.1	53.7	51.9	65.5	60.4
	Additive Logistic Regression	56.8	61.2	60.1	62.4	63.1	68.1	62.3
Rules	Decision Tables	52.7	58.3	61.4	62.2	66.7	67.1	61.1
	Nearest Neighbor Gen.	50.9	59.3	59.4	63.2	63.2	65.7	61.5
	PART	58.3	58.7	59.5	59.8	62.6	65.8	61.0
Trees	C4.5 Decision Tree	54.0	57.1	56.0	58.2	64.4	65.9	60.4
	Logistic Model Trees	56.1	57.0	57.8	61.3	62.5	63.8	56.8
	REP Tree	55.1	57.9	59.4	61.3	62.7	64.0	60.7
Mean		55.8	58.5	58.8	60.6	62.7	65.9	60.1

**Table 25. Classification accuracies in discriminating between confusion and neutral.**

Classifier Type	Classifier Name	Classification Accuracy Confusion and Neutral						Mean
		Self	Peer	Judge 1	Judge 2	Judges Agree	2 Agree	
Bayesian	Naive Bayes	57.9	58.5	61.2	59.0	59.7	58.7	59.9
	Naive Bayes Updatable	57.5	59.0	61.0	58.9	60.5	59.1	59.9
	Logistic Regression	58.9	56.9	62.6	60.8	60.9	67.1	65.4
	Multilayer Perception	56.9	56.6	60.8	59.2	55.5	66.3	62.1
Functions	SupportVector Machines	58.0	54.9	61.7	60.9	58.6	64.4	63.8
	Nearest Neighbor	56.8	57.2	58.2	59.2	54.9	67.8	58.0
	K*	56.7	59.2	61.9	60.7	56.3	62.6	62.6
	Locally Weighted Learning	53.8	57.2	60.3	60.1	55.1	58.4	64.5
Instance	AdaBoost	54.7	58.9	53.6	56.4	52.7	57.6	56.5
	Bagging Predictors	56.9	55.8	56.9	51.6	48.4	64.9	61.9
	Additive Logistic Regression	56.8	59.4	60.6	61.2	56.0	65.9	64.2
	Decision Tables	54.0	53.7	60.9	59.1	57.7	59.8	61.0
Rules	Nearest Neighbor Gen.	53.0	57.0	62.3	60.0	55.0	65.9	59.6
	PART	57.7	57.1	60.6	59.5	60.3	66.1	63.6
	C4.5 Decision Tree	56.6	57.2	56.5	56.0	54.8	63.9	60.5
	Logistic Model Trees	57.0	57.9	59.8	58.4	52.9	61.4	62.1
Trees	REP Tree	57.6	57.7	59.4	57.3	55.2	66.4	60.6
	Mean	56.5	57.3	59.9	58.7	56.1	63.3	61.5

**Table 26.** Classification accuracies in discriminating between flow and neutral.

Classifier Type	Classifier Name	Classification Accuracy						Mean
		Flow and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree	
Bayesian	Naive Bayes	50.2	55.3	62.8	65.8	60.1	61.2	60.9
	Naive Bayes Updatable	49.6	55.1	62.4	65.9	60.1	61.3	61.5
	Logistic Regression	50.5	55.2	70.0	66.8	70.5	66.9	63.9
	Multilayer Perceptron	50.1	52.5	67.1	64.2	68.7	67.3	62.1
Functions	Support Vector Machines	51.7	54.2	63.9	66.7	66.6	66.5	61.5
	Nearest Neighbor	49.9	52.5	61.4	55.6	65.6	65.1	58.4
	K*	51.7	54.6	65.7	63.2	66.9	58.7	60.5
	Locally Weighted Learning	52.0	53.7	65.4	61.8	66.0	57.3	53.6
Instance	AdaBoost	52.8	56.0	56.4	58.6	59.1	55.3	54.5
	Bagging Predictors	50.2	54.3	55.7	57.1	54.1	65.1	60.3
	Additive Logistic Regression	52.9	55.4	64.6	62.9	66.3	63.7	58.3
	Decision Tables	48.9	51.0	62.3	60.6	61.3	65.9	58.0
Rules	Nearest Neighbor Gen.	51.9	55.6	69.2	66.1	65.7	66.5	57.6
	PART	49.5	55.5	62.7	62.2	68.2	65.3	62.5
	C4.5 Decision Tree	51.8	52.9	62.0	58.8	63.4	65.2	57.6
Trees	Logistic Model Trees	50.9	53.4	62.0	62.3	61.6	59.3	57.8
	REP Tree	50.2	54.3	62.5	63.3	64.2	66.0	58.2
	Mean	50.9	54.2	63.3	62.5	64.0	63.3	59.2

**Table 27.** Classification accuracies in discriminating between frustration and neutral.

Classifier Type	Classifier Name	Classification Accuracy						Mean
		Frustration and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree	
Bayesian	Naive Bayes	63.7	65.9	72.4	68.9	71.9	73.7	72.2
	Naive Bayes Updatable	63.8	65.6	72.8	68.9	71.3	72.8	71.9
Functions	Logistic Regression	63.0	67.6	73.2	71.0	73.5	75.0	72.8
	Multilayer Perceptron	62.7	65.4	73.7	69.9	73.4	75.3	73.4
Instance	Support Vector Machines	62.8	67.0	73.2	71.6	73.1	77.2	73.4
	Nearest Neighbor	63.7	67.7	76.1	73.5	76.6	75.3	76.0
K*	K*	62.0	66.0	76.7	72.5	73.5	72.5	72.7
	Locally Weighted Learning	55.2	58.3	76.2	71.5	75.5	63.2	71.1
Meta	AdaBoost	62.3	61.0	65.3	61.6	65.2	48.7	66.5
	Bagging Predictors	64.1	69.2	67.7	49.3	51.6	76.5	74.2
Rules	Additive Logistic Regression	62.5	64.8	74.0	71.8	71.6	73.3	70.3
	Decision Tables	64.1	68.5	76.4	73.2	76.6	75.8	72.9
Trees	Nearest Neighbor Gen.	55.2	67.8	74.8	70.9	74.4	77.7	74.3
	PART	63.1	66.6	74.7	72.0	73.9	76.0	70.7
Mean	C4.5 Decision Tree	63.3	59.7	68.1	64.6	76.5	76.5	71.4
	Logistic Model Trees	62.4	67.0	75.6	71.0	68.9	60.3	69.0
	REP Tree	62.9	67.7	76.7	70.9	74.6	77.3	67.5
	Mean	62.2	65.6	73.4	69.0	71.9	72.2	72.0

## Appendix C. Affect Classification from Gross Body Language

*Detailed classification results for affect-neutral discriminations*

Table 28. Affect neutral classification for self judgments.

SC	MEAN P	C	STDEV P	PRESSURE (P)			CONTOURS (C)			
				A	B	C	D	E	A	
1	.20	.28	.06	.25	.19	.16	.26	.13	.29	.19
2	.20	.28	.05	.24	.20	.15	.26	.14	.29	.19
3	.22	.37	.04	.16	.28	.20	.17	.24	.21	.46
4	.23	.41	.05	.09	.29	.20	.17	.25	.25	.45
5	.40	.40	.06	.13	.41	.35	.34	.47	.43	.46
6	.39	.43	.09	.14	.41	.37	.26	.49	.43	.47
7	.39	.39	.07	.12	.38	.35	.29	.49	.42	.57
8	.21	.28	.03	.05	.20	.24	.23	.25	.16	.28
9	.27	.32	.05	.05	.31	.26	.27	.33	.19	.32
10	.36	.43	.07	.12	.38	.33	.27	.46	.34	.45
11	.28	.36	.06	.07	.34	.26	.26	.35	.21	.38
12	.22	.35	.06	.08	.27	.24	.21	.27	.11	.36
13	.24	.33	.07	.08	.29	.20	.16	.32	.22	.35
14	.27	.33	.06	.08	.32	.27	.20	.34	.23	.33
15	.29	.33	.06	.09	.34	.27	.22	.37	.25	.33
16	.31	.40	.07	.10	.34	.30	.22	.41	.30	.46
17	.28	.33	.06	.09	.33	.24	.24	.36	.23	.35

- SC = Classifier  
 6 = 3 Nearest Neighbors  
 12 = Decision Tables  
 A = Boredom-Neutral,  
 B = Confusion-Neutral,  
 C = Delight-Neutral,  
 D = Flow-Neutral,  
 E = Frustration-Neutral
- 1 = Naïve Bayes  
 7 = K\*  
 13 = Nearest Neighbor Gen.  
 14 = PART
- 2 = Naïve Bayes Updatable  
 8 = Locally Weighted Learning
- 3 = Logistic Regression  
 9 = AdaBoost  
 15 = C4.5 Decision Trees
- 4 = Support Vector Machines  
 10 = Bagging Predictors  
 16 = Logistic Model Trees
- 5 = 1 Nearest Neighbor  
 11 = Additive Logistic Regression  
 17 = REP Tree

Table 29. Affect neutral classification for peer judgments.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	P	C	C	A	B	C	D	E	A	B	C	D	E
1	.20	.28	.10	.10	.20	.15	.07	.33	.23	.20	.19	.31	.43	.25		
2	.19	.28	.10	.10	.20	.15	.07	.33	.23	.19	.20	.33	.42	.27		
3	.21	.35	.07	.14	.22	.17	.15	.33	.17	.35	.35	.19	.58	.30		
4	.21	.40	.09	.13	.23	.16	.11	.35	.17	.35	.38	.30	.62	.35		
5	.37	.43	.10	.11	.34	.32	.28	.53	.35	.40	.32	.41	.61	.39		
6	.37	.42	.13	.14	.34	.35	.22	.57	.36	.41	.38	.27	.65	.38		
7	.36	.39	.11	.11	.33	.30	.27	.54	.33	.40	.33	.31	.58	.35		
8	.19	.28	.10	.08	.21	.13	.08	.35	.19	.24	.25	.26	.42	.22		
9	.26	.30	.11	.12	.24	.20	.19	.45	.21	.17	.31	.33	.49	.21		
10	.32	.43	.12	.11	.30	.29	.23	.52	.26	.39	.41	.36	.63	.38		
11	.25	.35	.11	.12	.23	.19	.16	.44	.21	.27	.34	.28	.55	.30		
12	.21	.34	.15	.11	.23	.11	.03	.43	.25	.29	.34	.25	.53	.30		
13	.21	.31	.09	.11	.22	.17	.10	.35	.21	.25	.28	.26	.51	.24		
14	.22	.32	.11	.11	.22	.17	.10	.41	.22	.25	.29	.27	.51	.29		
15	.25	.33	.11	.11	.23	.20	.14	.44	.23	.30	.30	.21	.52	.31		
16	.27	.38	.11	.12	.26	.23	.16	.46	.22	.32	.36	.29	.60	.36		
17	.24	.32	.11	.12	.24	.19	.14	.43	.21	.29	.31	.22	.53	.26		

SC = Classifier  
1 = Naïve Bayes  
6 = 3 Nearest Neighbors  
12 = Decision Tables  
A = Boredom-Neutral,

2 = Naïve Bayes Undataable  
7 = K\*  
13 = Nearest Neighbor Gen.  
B = Confusion-Neutral,

3 = Logistic Regression  
8 = Locally Weighted Learning  
14 = PART

4 = Support Vector Machines  
9 = AdaBoost  
15 = C4.5 Decision Trees  
D = Delight-Neutral,  
E = Frustration-Neutral

5 = 1 Nearest Neighbor  
10 = Bagging Predictors  
16 = Logistic Model Trees  
E = Frustration-Neutral

**Table 30. Affect neutral classification for judgments by trained judge 1.**

SC	MEAN			STDDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	P	C	C	A	B	C	D	E	A	B	C	D	E
1	.14	.15	.04	.07	.14	.11	.19	.18	.10	.13	.09	.26	.17	.08		
2	.14	.15	.04	.07	.13	.11	.19	.17	.09	.13	.09	.26	.17	.09		
3	.18	.19	.06	.03	.15	.10	.27	.20	.16	.22	.17	.18	.22	.13		
4	.18	.21	.05	.06	.17	.11	.25	.20	.16	.24	.17	.29	.23	.14		
5	.23	.22	.04	.06	.26	.17	.24	.21	.25	.28	.14	.30	.20	.21		
6	.23	.23	.04	.04	.24	.17	.23	.23	.27	.27	.18	.27	.21	.23		
7	.23	.22	.05	.05	.29	.15	.27	.22	.22	.24	.15	.29	.19	.21		
8	.13	.15	.07	.02	.09	.06	.24	.15	.12	.13	.12	.18	.16	.14		
9	.14	.19	.07	.06	.13	.04	.23	.18	.13	.17	.14	.29	.19	.15		
10	.22	.23	.03	.04	.21	.19	.27	.24	.19	.25	.18	.28	.25	.19		
11	.16	.18	.04	.04	.14	.13	.23	.17	.12	.18	.15	.26	.17	.16		
12	.10	.17	.08	.04	.06	.02	.21	.14	.06	.15	.17	.24	.19	.12		
13	.14	.15	.03	.04	.16	.09	.18	.13	.15	.15	.11	.21	.15	.13		
14	.13	.14	.04	.04	.09	.10	.20	.12	.13	.18	.11	.20	.13	.11		
15	.15	.15	.04	.04	.12	.13	.22	.15	.15	.18	.12	.20	.15	.12		
16	.18	.20	.04	.04	.16	.14	.25	.19	.16	.21	.17	.25	.20	.15		
17	.15	.16	.03	.03	.13	.12	.20	.15	.13	.16	.13	.21	.17	.12		

SC = Classifier  
 1 = Naïve Bayes  
 6 = 3 Nearest Neighbors  
 12 = Decision Tables  
 A = Boredom-Neutral,  
 B = Confusion-Neutral,

2 = Naïve Bayes Updatable  
 8 = Locally Weighted Learning  
 14 = PART  
 C = Delight-Neutral,

3 = Logistic Regression  
 9 = AdaBoost  
 15 = C4.5 Decision Trees  
 D = Flow-Neutral,

4 = Support Vector Machines  
 10 = Bagging Predictors  
 16 = Logistic Model Trees  
 E = Frustration-Neutral

5 = 1 Nearest Neighbor  
 11 = Additive Logistic Regression  
 17 = REP Tree

**Table 31.** Affect neutral classification for judgments by trained judge 2.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D	E
1	.14	.17	.06	.07	.10	.08	.21	.21	.13	.17	.07	.25	.14	.21		
2	.15	.17	.07	.06	.09	.08	.22	.22	.14	.17	.08	.24	.16	.21		
3	.22	.17	.11	.03	.20	.07	.38	.25	.23	.20	.12	.17	.17	.20		
4	.21	.24	.10	.08	.19	.07	.35	.23	.24	.22	.13	.31	.22	.32		
5	.24	.26	.06	.09	.23	.16	.31	.22	.29	.24	.14	.35	.22	.34		
6	.26	.26	.07	.09	.27	.14	.32	.25	.31	.26	.13	.37	.24	.28		
7	.23	.25	.05	.10	.25	.15	.30	.22	.23	.22	.12	.34	.22	.35		
8	.16	.16	.09	.06	.16	.08	.30	.15	.10	.17	.05	.20	.17	.20		
9	.20	.18	.09	.06	.18	.09	.34	.22	.16	.21	.08	.24	.17	.22		
10	.25	.25	.08	.06	.26	.13	.35	.26	.25	.25	.14	.30	.24	.29		
11	.21	.20	.09	.07	.20	.13	.35	.23	.17	.23	.10	.23	.18	.27		
12	.14	.18	.12	.08	.17	.08	.33	.15	.00	.19	.04	.24	.18	.24		
13	.16	.16	.07	.06	.16	.08	.28	.16	.12	.14	.07	.23	.16	.20		
14	.16	.15	.06	.06	.17	.09	.26	.16	.14	.13	.07	.23	.14	.20		
15	.18	.18	.08	.05	.18	.09	.30	.18	.15	.16	.10	.23	.17	.23		
16	.22	.21	.07	.05	.20	.12	.32	.24	.21	.21	.12	.24	.19	.27		
17	.17	.17	.06	.05	.17	.10	.26	.20	.13	.20	.09	.21	.16	.19		

SC = Classifier  
 1 = Naïve Bayes  
 6 = 3 Nearest Neighbors  
 12 = Decision Tables  
 A = Boredom-Neutral,  
 B = Confusion-Neutral,  
 C = Delight-Neutral,

2 = Naïve Bayes Updatable  
 8 = Locally Weighted Learning  
 13 = Nearest Neighbor Gen.  
 D = Flow-Neutral,

3 = Logistic Regression  
 9 = AdaBoost  
 14 = PART  
 C = Delight-Neutral,  
 D = Flow-Neutral,

4 = Support Vector Machines  
 10 = Bagging Predictors  
 15 = C4.5 Decision Trees  
 E = Frustration-Neutral

5 = 1 Nearest Neighbor  
 11 = Additive Logistic Regression  
 17 = REP Tree

**Table 32.** Affect neutral classification for judgments where trained judges agree.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D	E
1	.19	.22	.10	.08	.18	.10	.28	.31	.09	.19	.10	.33	.26	.20		
2	.20	.21	.10	.08	.18	.10	.29	.32	.12	.20	.10	.32	.24	.17		
3	.22	.22	.11	.05	.20	.09	.32	.35	.15	.22	.19	.24	.28	.16		
4	.21	.28	.11	.06	.17	.09	.31	.34	.12	.26	.21	.37	.28	.29		
5	.25	.29	.05	.08	.27	.19	.21	.31	.27	.36	.18	.34	.33	.25		
6	.26	.31	.07	.09	.32	.19	.24	.34	.19	.33	.21	.40	.37	.22		
7	.25	.27	.06	.07	.28	.18	.25	.34	.20	.30	.16	.34	.32	.25		
8	.20	.19	.12	.05	.15	.11	.38	.27	.08	.13	.18	.24	.26	.17		
9	.22	.23	.11	.08	.19	.11	.38	.30	.13	.20	.12	.31	.26	.27		
10	.28	.31	.10	.06	.25	.18	.41	.36	.21	.30	.22	.36	.36	.29		
11	.22	.24	.10	.06	.17	.12	.33	.33	.14	.22	.15	.32	.27	.25		
12	.18	.22	.14	.06	.16	.13	.38	.26	.01	.21	.13	.28	.28	.19		
13	.19	.21	.09	.05	.18	.12	.31	.25	.11	.20	.15	.26	.25	.19		
14	.19	.20	.11	.06	.14	.11	.34	.26	.08	.18	.11	.27	.25	.22		
15	.19	.21	.11	.05	.17	.12	.31	.28	.07	.20	.14	.28	.24	.20		
16	.22	.25	.10	.06	.18	.13	.31	.34	.13	.24	.19	.35	.28	.20		
17	.19	.20	.12	.05	.16	.11	.34	.28	.05	.18	.12	.22	.27	.19		

SC = Classifier  
1 = Naive Bayes

2 = Native Bayes Updatable  
3 = Logistic Regression

4 = Support Vector Machines  
5 = 1 Nearest Neighbor

6 = 3 Nearest Neighbors  
7 = K\*  
8 = Locally Weighted Learning  
9 = AdaBoost

10 = Bagging Predictors  
11 = Additive Logistic Regression

12 = Decision Tables  
13 = Nearest Neighbor Gen.  
14 = PART  
15 = C4.5 Decision Trees

16 = Logistic Model Trees  
17 = REP Tree

C = Delight-Neutral,  
D = Boredom-Neutral,

E = Frustration-Neutral

*Detailed classification results for 2-way classifications*

**Table 33. Two-way classification for self judgments.**

SC	MEAN			STDEV			PRESSURE (P)							CONTOURS (C)										
	P	C	P	C	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
1	.14	.21	.06	.05	.23	.17	.25	.14	.10	.16	.10	.06	.12	.08	.20	.26	.21	.21	.15	.19	.20	.30	.14	
2	.14	.21	.07	.05	.23	.17	.26	.13	.11	.16	.10	.03	.10	.09	.20	.28	.26	.22	.19	.15	.18	.30	.15	
3	.16	.24	.06	.07	.20	.25	.24	.18	.13	.13	.08	.08	.18	.15	.32	.17	.36	.29	.18	.23	.24	.14	.22	
4	.15	.29	.07	.06	.21	.21	.24	.20	.09	.12	.10	.02	.14	.16	.36	.26	.35	.30	.19	.28	.25	.24	.26	
5	.28	.30	.06	.06	.30	.40	.29	.27	.26	.19	.22	.27	.34	.26	.35	.36	.38	.30	.26	.24	.23	.29	.40	
6	.27	.29	.08	.07	.35	.34	.35	.34	.15	.21	.25	.17	.27	.28	.39	.32	.39	.30	.19	.26	.24	.22	.35	
7	.28	.27	.05	.05	.30	.37	.32	.27	.23	.24	.23	.28	.32	.24	.31	.31	.37	.26	.21	.22	.21	.24	.30	
8	.15	.19	.03	.04	.14	.23	.16	.15	.17	.11	.14	.15	.16	.12	.23	.27	.23	.20	.17	.16	.15	.17	.18	
9	.18	.22	.06	.04	.21	.29	.25	.17	.16	.13	.10	.10	.18	.16	.23	.27	.27	.18	.18	.24	.18	.21	.23	
10	.25	.29	.07	.06	.32	.36	.35	.29	.21	.21	.17	.17	.20	.26	.39	.33	.37	.33	.20	.28	.25	.22	.31	
11	.19	.25	.08	.05	.26	.35	.28	.17	.16	.17	.11	.12	.17	.16	.30	.28	.31	.25	.16	.27	.23	.22	.29	
12	.10	.22	.09	.05	.20	.20	.27	.07	.06	.09	.03	.00	.02	.06	.28	.25	.28	.21	.13	.24	.21	.20	.23	
13	.17	.23	.06	.05	.21	.30	.21	.14	.14	.13	.11	.14	.20	.13	.26	.28	.26	.20	.22	.17	.18	.23	.16	
14	.17	.21	.06	.04	.21	.28	.24	.14	.14	.13	.10	.10	.16	.15	.24	.28	.25	.21	.18	.19	.19	.17	.26	
15	.17	.22	.06	.04	.24	.27	.25	.17	.16	.14	.10	.09	.13	.17	.25	.29	.27	.25	.20	.19	.19	.17	.22	
16	.18	.25	.06	.05	.26	.26	.27	.20	.12	.16	.11	.12	.15	.19	.32	.27	.32	.28	.19	.22	.22	.28	.21	
17	.18	.22	.06	.04	.25	.28	.26	.19	.13	.15	.12	.12	.14	.15	.28	.23	.29	.22	.15	.22	.20	.22	.18	

<sup>1</sup> = Naïve Bayes, <sup>2</sup> = Native Bayes Updatable, <sup>3</sup> = Logistic Regression, <sup>4</sup> = Support Vector Machines, <sup>5</sup> = 1 Nearest Neighbor, <sup>6</sup> = 3 Nearest Neighbors, <sup>7</sup> = K\*, <sup>8</sup> = Locally Weighted Learning, <sup>9</sup> = AdaBoost, <sup>10</sup> = Bagging Predictors, <sup>11</sup> = Additive Logistic Regression, <sup>12</sup> = Decision Tables, <sup>13</sup> = Nearest Neighbor Generalization, <sup>14</sup> = PART, <sup>15</sup> = C4.5 Decision Trees, <sup>16</sup> = Logistic Model Trees, <sup>17</sup> = REP Tree.

A = Boredom-Confusion, B = Boredom-Delight, C = Boredom-Flow, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Flow, G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

**Table 34. Two-way classification for peer judgments.**

SC	MEAN			STDEV			PRESSURE (P)			G			H			I			J			CONTOURS (C)				
	P	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J	A	B	C		
1	.19	.24	.07	.09	.22	.24	.23	.06	.30	.22	.16	.21	.17	.36	.38	.10	.24	.26	.17	.16	.24	.27				
2	.19	.24	.08	.10	.22	.26	.23	.06	.05	.30	.22	.16	.23	.21	.16	.40	.38	.08	.23	.26	.16	.16	.27	.28		
3	.23	.23	.09	.11	.25	.33	.35	.07	.10	.26	.25	.24	.17	.31	.25	.24	.47	.17	.11	.33	.11	.17	.16	.27		
4	.22	.29	.10	.10	.25	.28	.33	.04	.03	.26	.23	.24	.21	.33	.28	.36	.52	.20	.18	.36	.19	.24	.26	.34		
5	.27	.31	.12	.12	.31	.28	.49	.19	.07	.23	.18	.31	.26	.39	.30	.44	.56	.24	.18	.31	.19	.30	.23	.38		
6	.27	.31	.13	.12	.30	.25	.52	.23	.03	.31	.20	.28	.19	.40	.33	.42	.57	.24	.18	.36	.20	.21	.23	.38		
7	.27	.27	.11	.13	.31	.29	.50	.18	.11	.29	.16	.22	.28	.39	.32	.37	.55	.17	.11	.29	.16	.22	.24	.30		
8	.21	.22	.09	.10	.23	.19	.33	.03	.11	.21	.24	.23	.24	.30	.14	.30	.38	.13	.16	.27	.14	.11	.29	.32		
9	.22	.24	.10	.10	.23	.19	.35	.02	.11	.24	.22	.27	.18	.35	.18	.37	.38	.13	.18	.28	.15	.13	.25	.32		
10	.26	.31	.12	.12	.29	.25	.49	.11	.09	.29	.18	.28	.24	.36	.32	.39	.56	.21	.21	.37	.16	.22	.30	.38		
11	.21	.26	.10	.09	.23	.23	.34	.06	.05	.24	.16	.21	.25	.33	.26	.27	.45	.17	.22	.33	.18	.15	.28	.33		
12	.19	.24	.11	.13	.24	.15	.34	-.02	.06	.25	.23	.16	.17	.29	.22	.36	.45	.06	.14	.29	.15	.09	.32	.33		
13	.18	.24	.08	.10	.18	.17	.33	.06	.05	.22	.15	.22	.17	.26	.20	.32	.43	.11	.16	.25	.12	.23	.32	.27		
14	.19	.21	.10	.10	.23	.21	.34	.00	.04	.23	.19	.25	.21	.27	.19	.27	.44	.12	.17	.24	.10	.12	.25	.25		
15	.20	.23	.11	.10	.23	.13	.40	.00	.08	.23	.18	.24	.18	.28	.22	.29	.45	.15	.12	.25	.14	.16	.22	.27		
16	.23	.28	.10	.10	.24	.27	.42	.06	.08	.25	.23	.23	.25	.31	.27	.38	.50	.16	.22	.34	.19	.18	.29	.32		
17	.20	.24	.11	.10	.23	.15	.39	.02	.02	.23	.18	.25	.22	.29	.22	.29	.46	.14	.17	.29	.15	.11	.25	.29		

SC = Supervised Classifier

1 = Naïve Bayes, 2 = Native Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion, B = Boredom-Frustration, C = Delight-Confusion, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Frustration, G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

Table 35. Two-way classification for judgments by trained judge 1.

SC	MEAN			STDEV			PRESSURE (P)							CONTOURS (C)												
	P	C	P	C	P	C	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
1	.19	.20	.05	.05	.21	.26	.27	.20	.14	.20	.11	.16	.17	.16	.21	.30	.21	.17	.21	.18	.11	.26	.17	.21		
2	.19	.20	.05	.05	.22	.25	.27	.19	.14	.20	.11	.15	.16	.16	.22	.29	.21	.16	.21	.17	.10	.26	.18	.21		
3	.26	.22	.06	.05	.23	.39	.34	.23	.25	.23	.20	.27	.19	.24	.25	.26	.30	.18	.15	.24	.14	.23	.17	.25		
4	.25	.27	.07	.07	.23	.36	.39	.23	.22	.22	.20	.26	.18	.25	.29	.34	.35	.24	.22	.25	.16	.35	.19	.31		
5	.34	.35	.07	.08	.35	.45	.43	.34	.27	.23	.28	.41	.30	.34	.33	.50	.39	.33	.27	.28	.22	.45	.38	.33		
6	.32	.33	.08	.08	.33	.43	.44	.32	.24	.26	.26	.37	.23	.36	.34	.47	.39	.31	.20	.31	.24	.40	.32	.35		
7	.33	.32	.08	.08	.34	.44	.44	.37	.28	.22	.25	.40	.30	.31	.30	.45	.35	.33	.23	.26	.21	.41	.30	.34		
8	.16	.19	.06	.06	.14	.13	.23	.12	.13	.14	.18	.25	.08	.23	.23	.32	.18	.18	.15	.20	.14	.20	.11	.20		
9	.21	.21	.06	.07	.17	.22	.30	.21	.15	.21	.18	.28	.12	.26	.25	.35	.26	.19	.17	.22	.15	.25	.11	.20		
10	.28	.30	.07	.07	.28	.35	.42	.28	.20	.24	.24	.32	.19	.33	.31	.44	.37	.29	.23	.27	.20	.36	.25	.35		
11	.23	.24	.05	.06	.20	.25	.31	.22	.17	.22	.21	.25	.14	.30	.28	.34	.25	.19	.20	.23	.18	.27	.14	.27		
12	.17	.20	.09	.08	.16	.18	.34	.11	.11	.14	.19	.25	.01	.23	.23	.36	.25	.18	.18	.21	.12	.25	.04	.21		
13	.19	.21	.06	.05	.20	.23	.30	.16	.16	.14	.14	.21	.11	.22	.21	.31	.27	.18	.17	.20	.14	.24	.18	.23		
14	.18	.21	.05	.07	.17	.23	.27	.18	.14	.17	.17	.23	.10	.20	.20	.35	.27	.16	.18	.15	.12	.27	.18	.21		
15	.21	.23	.05	.07	.19	.24	.32	.18	.18	.17	.19	.23	.14	.24	.21	.37	.28	.18	.16	.18	.16	.31	.19	.24		
16	.25	.25	.06	.07	.22	.33	.37	.21	.20	.22	.20	.26	.18	.30	.26	.35	.31	.21	.20	.22	.13	.31	.20	.27		
17	.20	.21	.06	.06	.17	.23	.32	.19	.17	.18	.20	.24	.11	.22	.21	.32	.24	.18	.14	.22	.13	.26	.14	.22		

SC = Supervised Classifier

1 = Naïve Bayes, 2 = Naïve Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion, B = Boredom-Delight, C = Boredom-Flow, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Flow, G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

Table 36. Two-way classification for judgments by trained judge 2.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)				
	P	C	P	C	A	B	C	D	E	F	G	H	I	J
1	.19	.19	.06	.06	.15	.21	.25	.18	.24	.24	.08	.28	.14	.22
2	.19	.19	.06	.06	.15	.22	.25	.17	.21	.23	.09	.28	.17	.22
3	.26	.19	.08	.08	.22	.27	.35	.18	.30	.26	.14	.40	.22	.27
4	.26	.27	.07	.09	.23	.27	.36	.17	.26	.27	.18	.41	.23	.26
5	.31	.32	.06	.08	.26	.35	.41	.33	.26	.27	.23	.40	.30	.32
6	.32	.31	.07	.10	.27	.32	.45	.31	.26	.30	.20	.40	.30	.38
7	.28	.29	.07	.07	.25	.32	.39	.29	.26	.24	.16	.38	.25	.31
8	.19	.19	.06	.08	.15	.19	.21	.13	.31	.17	.13	.29	.19	.15
9	.22	.21	.08	.08	.17	.25	.27	.14	.32	.23	.13	.35	.12	.22
10	.28	.30	.06	.08	.26	.28	.36	.22	.28	.28	.19	.39	.21	.30
11	.23	.25	.06	.08	.20	.23	.30	.17	.27	.24	.12	.33	.18	.27
12	.16	.19	.12	.09	.19	.13	.26	.01	.31	.21	.03	.34	.01	.15
13	.20	.21	.06	.08	.16	.21	.28	.16	.18	.19	.09	.32	.17	.21
14	.18	.22	.06	.07	.18	.17	.24	.11	.25	.20	.07	.27	.14	.17
15	.20	.22	.06	.07	.19	.22	.24	.13	.25	.23	.11	.30	.12	.20
16	.25	.25	.08	.09	.21	.24	.33	.18	.29	.25	.14	.40	.20	.26
17	.20	.20	.07	.07	.18	.21	.26	.15	.27	.20	.10	.33	.15	.21

SC = Supervised Classifier

1 = Naïve Bayes, 2 = Naïve Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion, B = Boredom-Delight, C = Boredom-Flow, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Flow, G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

Table 37. Two-way classification for judgments where trained judges agree.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)			
	P	C	P	A	B	C	D	E	F	G	H	I	J
1	.24	.23	.10	.07	.23	.27	.37	.18	.26	.29	.03	.35	.15
2	.24	.23	.10	.07	.23	.26	.37	.17	.27	.29	.02	.36	.16
3	.27	.21	.11	.11	.22	.28	.45	.14	.28	.30	.09	.43	.19
4	.27	.32	.13	.10	.22	.26	.45	.12	.30	.32	.08	.45	.18
5	.31	.40	.07	.09	.28	.29	.42	.37	.24	.31	.22	.41	.22
6	.32	.40	.09	.10	.32	.27	.49	.35	.23	.36	.20	.40	.26
7	.31	.35	.07	.09	.32	.33	.47	.31	.25	.31	.20	.38	.24
8	.19	.23	.08	.09	.14	.14	.28	.14	.26	.19	.04	.27	.24
9	.24	.27	.09	.08	.20	.23	.33	.19	.31	.27	.06	.33	.15
10	.31	.34	.09	.09	.27	.33	.44	.25	.36	.35	.13	.38	.22
11	.24	.29	.09	.09	.22	.27	.37	.19	.30	.30	.08	.31	.12
12	.18	.24	.12	.09	.19	.13	.36	.03	.24	.21	-.04	.28	.19
13	.21	.26	.09	.08	.17	.19	.33	.20	.23	.21	.05	.35	.13
14	.20	.27	.09	.08	.19	.18	.33	.18	.23	.28	.01	.26	.13
15	.21	.27	.09	.08	.24	.17	.37	.20	.22	.29	.02	.24	.13
16	.27	.30	.12	.09	.22	.25	.43	.21	.29	.30	.04	.44	.22
17	.22	.23	.08	.08	.21	.18	.33	.18	.26	.27	.04	.28	.18

SC = Supervised Classifier

1 = Naïve Bayes, 2 = Native Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Decision Trees, 17 = REP Tree.

A = Boredom-Confusion, B = Boredom-Delight, C = Boredom-Frustration, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Frustration  
 G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

## Appendix B. Affect Classification from Conversational Cues

*Detailed classification results for 4-way emotion discriminations*

Table 23. Classification accuracies for 4-way discrimination.

Classifier Type	Classifier Name	Classification Accuracy Boredom, Confusion, Flow, and Frustration						Mean
		Self	Peer	Judge 1	Judge 2	Judges Agree	2 Agree	
Bayesian	Naive Bayes	33.2	37.1	45.4	46.3	47.4	49.0	49.2
	Naive Bayes Updatable	33.5	37.7	44.6	46.4	46.8	48.8	49.1
Functions	Logistic Regression	35.1	38.4	50.0	50.4	54.0	51.5	52.5
	Multilayer Perceptron	32.3	36.3	44.9	46.6	47.2	49.2	48.8
	Support Vector Machines	34.5	38.2	48.1	49.0	48.4	51.2	50.5
Instance	Nearest Neighbor	28.7	31.0	40.0	40.2	42.7	43.0	39.4
	K*	29.1	35.1	31.1	39.5	31.4	30.3	31.3
Locally Weighted Learning	Locally Weighted Learning	32.5	37.3	45.0	44.2	47.0	44.8	43.2
	AdaBoost	28.5	31.3	41.8	41.9	42.7	38.3	41.5
Meta	Bagging Predictors	32.1	37.0	46.2	47.9	48.8	49.2	47.6
	Additive Logistic Regression	34.1	38.8	45.9	48.6	49.9	52.7	49.6
	Decision Tables	27.2	36.7	42.2	47.0	48.8	44.8	46.6
Rules	Nearest Neighbor Gen.	29.2	31.6	41.6	41.3	44.8	46.1	42.5
	PART	30.2	34.6	41.2	41.9	43.3	46.3	44.5
Trees	C4.5 Decision Tree	28.6	36.3	41.6	43.5	45.2	43.5	43.4
	Logistic Model Trees	34.3	37.0	50.5	49.9	52.9	50.3	51.7
	REP Tree	30.3	36.7	42.2	44.5	46.2	44.9	44.2
Mean		31.4	35.9	43.7	45.2	46.3	46.1	45.6

*Detailed classification results for affect-neutral discriminations*

**Table 24.** Classification accuracies in discriminating between boredom and neutral.

Classifier Type	Classifier Name	Classification Accuracy						Mean
		Boredom and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree	
Self	Peer							
Bayesian	Naive Bayes	58.1	57.7	58.6	61.2	61.3	65.6	60.6
	Naive Bayes Updatable	57.9	58.2	59.2	61.8	62.4	66.3	60.7
Functions	Logistic Regression	58.6	59.2	59.4	63.9	63.2	67.6	61.4
	MultiLayer Perceptron	58.4	59.3	56.5	60.0	61.7	67.7	61.3
	Support Vector Machines	61.3	58.8	59.5	61.7	62.0	65.3	61.4
Instance	Nearest Neighbor	54.0	58.5	61.2	60.8	67.2	69.0	61.6
	K*	54.7	59.5	59.2	62.6	65.6	65.9	62.7
	Locally Weighted Learning	52.0	55.5	58.5	60.6	64.4	62.7	61.4
Meta	AdaBoost	53.6	59.8	56.0	56.2	61.7	65.0	50.7
	Bagging Predictors	56.0	58.8	57.1	53.7	51.9	65.5	60.4
	Additive Logistic Regression	56.8	61.2	60.1	62.4	63.1	68.1	62.3
Rules	Decision Tables	52.7	58.3	61.4	62.2	66.7	67.1	61.1
	Nearest Neighbor Gen.	50.9	59.3	59.4	63.2	63.2	65.7	61.5
	PART	58.3	58.7	59.5	59.8	62.6	65.8	61.0
Trees	C4.5 Decision Tree	54.0	57.1	56.0	58.2	64.4	65.9	60.4
	Logistic Model Trees	56.1	57.0	57.8	61.3	62.5	63.8	56.8
	REP Tree	55.1	57.9	59.4	61.3	62.7	64.0	60.7
Mean		55.8	58.5	58.8	60.6	62.7	65.9	60.1

**Table 25. Classification accuracies in discriminating between confusion and neutral.**

Classifier Type	Classifier Name	Classification Accuracy						Mean
		Confusion and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree	
Self	Peer							
Bayesian	Naive Bayes	57.9	58.5	61.2	59.0	59.7	58.7	59.9
	Naive Bayes Updatable	57.5	59.0	61.0	58.9	60.5	59.1	59.9
	Logistic Regression	58.9	56.9	62.6	60.8	60.9	67.1	65.4
	Multilayer Perception	56.9	56.6	60.8	59.2	55.5	66.3	62.1
Functions	SupportVector Machines	58.0	54.9	61.7	60.9	58.6	64.4	63.8
	Nearest Neighbor	56.8	57.2	58.2	59.2	54.9	67.8	58.0
	K*	56.7	59.2	61.9	60.7	56.3	62.6	62.6
	Locally Weighted Learning	53.8	57.2	60.3	60.1	55.1	58.4	64.5
Instance	AdaBoost	54.7	58.9	53.6	56.4	52.7	57.6	56.5
	Bagging Predictors	56.9	55.8	56.9	51.6	48.4	64.9	61.9
	Additive Logistic Regression	56.8	59.4	60.6	61.2	56.0	65.9	64.2
	Decision Tables	54.0	53.7	60.9	59.1	57.7	59.8	61.0
Rules	Nearest Neighbor Gen.	53.0	57.0	62.3	60.0	55.0	65.9	59.6
	PART	57.7	57.1	60.6	59.5	60.3	66.1	63.6
	C4.5 Decision Tree	56.6	57.2	56.5	56.0	54.8	63.9	60.5
	Logistic Model Trees	57.0	57.9	59.8	58.4	52.9	61.4	62.1
Trees	REP Tree	57.6	57.7	59.4	57.3	55.2	66.4	60.6
	Mean	56.5	57.3	59.9	58.7	56.1	63.3	61.5

**Table 26.** Classification accuracies in discriminating between flow and neutral.

Classifier Type	Classifier Name	Classification Accuracy						Mean
		Flow and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree	
Bayesian	Naive Bayes	50.2	55.3	62.8	65.8	60.1	61.2	60.9
	Naive Bayes Updatable	49.6	55.1	62.4	65.9	60.1	61.3	61.5
	Logistic Regression	50.5	55.2	70.0	66.8	70.5	66.9	63.9
	Multilayer Perceptron	50.1	52.5	67.1	64.2	68.7	67.3	62.1
Functions	Support Vector Machines	51.7	54.2	63.9	66.7	66.6	66.5	61.5
	Nearest Neighbor	49.9	52.5	61.4	55.6	65.6	65.1	58.4
	K*	51.7	54.6	65.7	63.2	66.9	58.7	60.5
	Locally Weighted Learning	52.0	53.7	65.4	61.8	66.0	57.3	53.6
Instance	AdaBoost	52.8	56.0	56.4	58.6	59.1	55.3	54.5
	Bagging Predictors	50.2	54.3	55.7	57.1	54.1	65.1	60.3
	Additive Logistic Regression	52.9	55.4	64.6	62.9	66.3	63.7	58.3
	Decision Tables	48.9	51.0	62.3	60.6	61.3	65.9	58.0
Rules	Nearest Neighbor Gen.	51.9	55.6	69.2	66.1	65.7	66.5	57.6
	PART	49.5	55.5	62.7	62.2	68.2	65.3	62.5
	C4.5 Decision Tree	51.8	52.9	62.0	58.8	63.4	65.2	57.6
	Logistic Model Trees	50.9	53.4	62.0	62.3	61.6	59.3	57.8
Trees	REP Tree	50.2	54.3	62.5	63.3	64.2	66.0	58.2
	Mean	50.9	54.2	63.3	62.5	64.0	63.3	59.2

**Table 27.** Classification accuracies in discriminating between frustration and neutral.

Classifier Type	Classifier Name	Classification Accuracy					
		Frustration and Neutral		Judge 1	Judge 2	Judges Agree	2 Agree
Bayesian	Naive Bayes	63.7	65.9	72.4	68.9	71.9	73.7
	Naive Bayes Updatable	63.8	65.6	72.8	68.9	71.3	72.8
Functions	Logistic Regression	63.0	67.6	73.2	71.0	73.5	75.0
	Multilayer Perceptron	62.7	65.4	73.7	69.9	73.4	75.3
Instance	Support Vector Machines	62.8	67.0	73.2	71.6	73.1	77.2
	Nearest Neighbor	63.7	67.7	76.1	73.5	76.6	75.3
K*	K*	62.0	66.0	76.7	72.5	73.5	72.5
	Locally Weighted Learning	55.2	58.3	76.2	71.5	75.5	63.2
Meta	AdaBoost	62.3	61.0	65.3	61.6	65.2	48.7
	Bagging Predictors	64.1	69.2	67.7	49.3	51.6	76.5
Rules	Additive Logistic Regression	62.5	64.8	74.0	71.8	71.6	73.3
	Decision Tables	64.1	68.5	76.4	73.2	76.6	75.8
Trees	Nearest Neighbor Gen.	55.2	67.8	74.8	70.9	74.4	77.7
	PART	63.1	66.6	74.7	72.0	73.9	76.0
Mean	C4.5 Decision Tree	63.3	59.7	68.1	64.6	76.5	76.5
	Logistic Model Trees	62.4	67.0	75.6	71.0	68.9	60.3
	REP Tree	62.9	67.7	76.7	70.9	74.6	77.3
	Mean	62.2	65.6	73.4	69.0	71.9	72.2
							72.0

## Appendix C. Affect Classification from Gross Body Language

*Detailed classification results for affect-neutral discriminations*

Table 28. Affect neutral classification for self judgments.

SC	MEAN P	C	STDEV P	PRESSURE (P)			CONTOURS (C)			
				A	B	C	D	E	A	
1	.20	.28	.06	.25	.19	.16	.26	.13	.29	.19
2	.20	.28	.05	.24	.20	.15	.26	.14	.29	.19
3	.22	.37	.04	.16	.28	.20	.17	.24	.21	.46
4	.23	.41	.05	.09	.29	.20	.17	.25	.25	.45
5	.40	.40	.06	.13	.41	.35	.34	.47	.43	.46
6	.39	.43	.09	.14	.41	.37	.26	.49	.43	.47
7	.39	.39	.07	.12	.38	.35	.29	.49	.42	.57
8	.21	.28	.03	.05	.20	.24	.23	.25	.16	.28
9	.27	.32	.05	.05	.31	.26	.27	.33	.19	.32
10	.36	.43	.07	.12	.38	.33	.27	.46	.34	.45
11	.28	.36	.06	.07	.34	.26	.26	.35	.21	.38
12	.22	.35	.06	.08	.27	.24	.21	.27	.11	.36
13	.24	.33	.07	.08	.29	.20	.16	.32	.22	.35
14	.27	.33	.06	.08	.32	.27	.20	.34	.23	.33
15	.29	.33	.06	.09	.34	.27	.22	.37	.25	.33
16	.31	.40	.07	.10	.34	.30	.22	.41	.30	.46
17	.28	.33	.06	.09	.33	.24	.24	.36	.23	.35

- SC = Classifier  
 6 = 3 Nearest Neighbors  
 12 = Decision Tables  
 A = Boredom-Neutral,  
 B = Confusion-Neutral,  
 C = Delight-Neutral,  
 D = Flow-Neutral,  
 E = Frustration-Neutral
- 1 = Naïve Bayes  
 7 = K\*  
 13 = Nearest Neighbor Gen.  
 14 = PART
- 2 = Naïve Bayes Updatable  
 8 = Locally Weighted Learning
- 3 = Logistic Regression  
 9 = AdaBoost  
 15 = C4.5 Decision Trees
- 4 = Support Vector Machines  
 10 = Bagging Predictors  
 16 = Logistic Model Trees
- 5 = 1 Nearest Neighbor  
 11 = Additive Logistic Regression  
 17 = REP Tree

Table 29. Affect neutral classification for peer judgments.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)					
	P	C	P	P	C	C	A	B	C	D	E	C	D	E	
1	.20	.28	.10	.10	.20	.07	.15	.07	.33	.23	.20	.19	.31	.43	.25
2	.19	.28	.10	.10	.20	.07	.15	.07	.33	.23	.19	.20	.33	.42	.27
3	.21	.35	.07	.14	.22	.17	.15	.33	.17	.35	.35	.19	.58	.30	
4	.21	.40	.09	.13	.23	.16	.11	.35	.17	.35	.38	.30	.62	.35	
5	.37	.43	.10	.11	.34	.32	.28	.53	.55	.40	.40	.32	.41	.61	.39
6	.37	.42	.13	.14	.34	.35	.22	.57	.36	.41	.38	.27	.65	.38	
7	.36	.39	.11	.11	.33	.30	.27	.54	.33	.40	.33	.31	.58	.35	
8	.19	.28	.10	.08	.21	.13	.08	.35	.19	.24	.25	.26	.42	.22	
9	.26	.30	.11	.12	.24	.20	.19	.45	.21	.17	.31	.33	.49	.21	
10	.32	.43	.12	.11	.30	.29	.23	.52	.26	.39	.41	.36	.63	.38	
11	.25	.35	.11	.12	.23	.19	.16	.44	.21	.27	.34	.28	.55	.30	
12	.21	.34	.15	.11	.23	.11	.03	.43	.25	.29	.34	.25	.53	.30	
13	.21	.31	.09	.11	.22	.17	.10	.35	.21	.25	.28	.26	.51	.24	
14	.22	.32	.11	.11	.22	.17	.10	.41	.22	.25	.29	.27	.51	.29	
15	.25	.33	.11	.11	.23	.20	.14	.44	.23	.30	.30	.21	.52	.31	
16	.27	.38	.11	.12	.26	.23	.16	.46	.22	.32	.36	.29	.60	.36	
17	.24	.32	.11	.12	.24	.19	.14	.43	.21	.29	.31	.22	.53	.26	

SC = Classifier  
1 = Naïve Bayes  
6 = 3 Nearest Neighbors  
12 = Decision Tables  
A = Boredom-Neutral,  
B = Confusion-Neutral,

2 = Naïve Bayes Undataable  
7 = K\*  
13 = Nearest Neighbor Gen.  
C = Delight-Neutral,

3 = Logistic Regression  
8 = Locally Weighted Learning  
14 = PART

4 = Support Vector Machines  
9 = AdaBoost  
15 = C4.5 Decision Trees  
D = Flow-Neutral,

5 = 1 Nearest Neighbor  
10 = Bagging Predictors  
16 = Logistic Model Trees  
E = Frustration-Neutral

11 = Additive Logistic Regression  
17 = REP Tree

**Table 30. Affect neutral classification for judgments by trained judge 1.**

SC	MEAN			STDDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	P	C	C	A	B	C	D	E	A	B	C	D	E
1	.14	.15	.04	.07	.14	.11	.19	.18	.10	.13	.09	.26	.17	.08		
2	.14	.15	.04	.07	.13	.11	.19	.17	.09	.13	.09	.26	.17	.09		
3	.18	.19	.06	.03	.15	.10	.27	.20	.16	.22	.17	.18	.22	.13		
4	.18	.21	.05	.06	.17	.11	.25	.20	.16	.24	.17	.29	.23	.14		
5	.23	.22	.04	.06	.26	.17	.24	.21	.25	.28	.14	.30	.20	.21		
6	.23	.23	.04	.04	.24	.17	.23	.23	.27	.27	.18	.27	.21	.23		
7	.23	.22	.05	.05	.29	.15	.27	.22	.22	.24	.15	.29	.19	.21		
8	.13	.15	.07	.02	.09	.06	.24	.15	.12	.13	.12	.18	.16	.14		
9	.14	.19	.07	.06	.13	.04	.23	.18	.13	.17	.14	.29	.19	.15		
10	.22	.23	.03	.04	.21	.19	.27	.24	.19	.25	.18	.28	.25	.19		
11	.16	.18	.04	.04	.14	.13	.23	.17	.12	.18	.15	.26	.17	.16		
12	.10	.17	.08	.04	.06	.02	.21	.14	.06	.15	.17	.24	.19	.12		
13	.14	.15	.03	.04	.16	.09	.18	.13	.15	.15	.11	.21	.15	.13		
14	.13	.14	.04	.04	.09	.10	.20	.12	.13	.18	.11	.20	.13	.11		
15	.15	.15	.04	.04	.12	.13	.22	.15	.15	.18	.12	.20	.15	.12		
16	.18	.20	.04	.04	.16	.14	.25	.19	.16	.21	.17	.25	.20	.15		
17	.15	.16	.03	.03	.13	.12	.20	.15	.13	.16	.13	.21	.17	.12		

SC = Classifier

1 = Naïve Bayes

2 = Native Bayes Updatable

3 = Logistic Regression

4 = Support Vector Machines

5 = 1 Nearest Neighbor

6 = K\*

7 = AdaBoost

8 = Locally Weighted Learning

9 = Bagging Predictors

10 = C4.5 Decision Trees

11 = Logistic Model Trees

12 = Decision Tables

13 = Nearest Neighbor Gen.

14 = PART

15 = C4.5 Decision Trees

16 = Logistic Model Trees

17 = REP Tree

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Delight-Neutral,

B = Confusion-Neutral,

C = Delight-Neutral,

D = Flow-Neutral,

E = Frustration-Neutral,

A = Del

**Table 31.** Affect neutral classification for judgments by trained judge 2.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D	E
1	.14	.17	.06	.07	.10	.08	.21	.21	.13	.17	.07	.25	.14	.21		
2	.15	.17	.07	.06	.09	.08	.22	.22	.14	.17	.08	.24	.16	.21		
3	.22	.17	.11	.03	.20	.07	.38	.25	.23	.20	.12	.17	.17	.20		
4	.21	.24	.10	.08	.19	.07	.35	.23	.24	.22	.13	.31	.22	.32		
5	.24	.26	.06	.09	.23	.16	.31	.22	.29	.24	.14	.35	.22	.34		
6	.26	.26	.07	.09	.27	.14	.32	.25	.31	.26	.13	.37	.24	.28		
7	.23	.25	.05	.10	.25	.15	.30	.22	.23	.22	.12	.34	.22	.35		
8	.16	.16	.09	.06	.16	.08	.30	.15	.10	.17	.05	.20	.17	.20		
9	.20	.18	.09	.06	.18	.09	.34	.22	.16	.21	.08	.24	.17	.22		
10	.25	.25	.08	.06	.26	.13	.35	.26	.25	.25	.14	.30	.24	.29		
11	.21	.20	.09	.07	.20	.13	.35	.23	.17	.23	.10	.23	.18	.27		
12	.14	.18	.12	.08	.17	.08	.33	.15	.00	.19	.04	.24	.18	.24		
13	.16	.16	.07	.06	.16	.08	.28	.16	.12	.14	.07	.23	.16	.20		
14	.16	.15	.06	.06	.17	.09	.26	.16	.14	.13	.07	.23	.14	.20		
15	.18	.18	.08	.05	.18	.09	.30	.18	.15	.16	.10	.23	.17	.23		
16	.22	.21	.07	.05	.20	.12	.32	.24	.21	.21	.12	.24	.19	.27		
17	.17	.17	.06	.05	.17	.10	.26	.20	.13	.20	.09	.21	.16	.19		

SC = Classifier  
 1 = Naïve Bayes  
 6 = 3 Nearest Neighbors  
 12 = Decision Tables  
 A = Boredom-Neutral,  
 B = Confusion-Neutral,  
 C = Delight-Neutral,

2 = Native Bayes Updatable  
 8 = Locally Weighted Learning  
 13 = Nearest Neighbor Gen.  
 D = Flow-Neutral,

3 = Logistic Regression  
 9 = AdaBoost  
 14 = PART  
 C = Delight-Neutral,  
 D = Flow-Neutral,

4 = Support Vector Machines  
 10 = Bagging Predictors  
 15 = C4.5 Decision Trees  
 E = Frustration-Neutral

5 = 1 Nearest Neighbor  
 11 = Additive Logistic Regression  
 17 = REP Tree

**Table 32.** Affect neutral classification for judgments where trained judges agree.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D	E
1	.19	.22	.10	.08	.18	.10	.28	.31	.09	.19	.10	.33	.26	.20		
2	.20	.21	.10	.08	.18	.10	.29	.32	.12	.20	.10	.32	.24	.17		
3	.22	.22	.11	.05	.20	.09	.32	.35	.15	.22	.19	.24	.28	.16		
4	.21	.28	.11	.06	.17	.09	.31	.34	.12	.26	.21	.37	.28	.29		
5	.25	.29	.05	.08	.27	.19	.21	.31	.27	.36	.18	.34	.33	.25		
6	.26	.31	.07	.09	.32	.19	.24	.34	.19	.33	.21	.40	.37	.22		
7	.25	.27	.06	.07	.28	.18	.25	.34	.20	.30	.16	.34	.32	.25		
8	.20	.19	.12	.05	.15	.11	.38	.27	.08	.13	.18	.24	.26	.17		
9	.22	.23	.11	.08	.19	.11	.38	.30	.13	.20	.12	.31	.26	.27		
10	.28	.31	.10	.06	.25	.18	.41	.36	.21	.30	.22	.36	.36	.29		
11	.22	.24	.10	.06	.17	.12	.33	.33	.14	.22	.15	.32	.27	.25		
12	.18	.22	.14	.06	.16	.13	.38	.26	.01	.21	.13	.28	.28	.19		
13	.19	.21	.09	.05	.18	.12	.31	.25	.11	.20	.15	.26	.25	.19		
14	.19	.20	.11	.06	.14	.11	.34	.26	.08	.18	.11	.27	.25	.22		
15	.19	.21	.11	.05	.17	.12	.31	.28	.07	.20	.14	.28	.24	.20		
16	.22	.25	.10	.06	.18	.13	.31	.34	.13	.24	.19	.35	.28	.20		
17	.19	.20	.12	.05	.16	.11	.34	.28	.05	.18	.12	.22	.27	.19		

SC = Classifier  
1 = Naive Bayes  
6 = 3 Nearest Neighbors  
12 = Decision Tables  
A = Boredom-Neutral,

2 = Native Bayes Updatable  
7 = K\*  
13 = Nearest Neighbor Gen.  
B = Confusion-Neutral,

3 = Logistic Regression  
8 = Locally Weighted Learning  
14 = PART

9 = AdaBoost  
15 = C4.5 Decision Trees  
D = Delight-Neutral,

4 = Support Vector Machines  
10 = Bagging Predictors  
16 = Logistic Model Trees

E = Frustration-Neutral

5 = 1 Nearest Neighbor  
11 = Additive Logistic Regression  
17 = REP Tree

**Table 34. Two-way classification for peer judgments.**

SC	MEAN			STDEV			PRESSURE (P)			G			H			I			J			CONTOURS (C)				
	P	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J	A	B	C		
1	.19	.24	.07	.09	.22	.24	.23	.06	.30	.22	.16	.21	.17	.36	.38	.10	.24	.26	.17	.16	.24	.27				
2	.19	.24	.08	.10	.22	.26	.23	.06	.05	.30	.22	.16	.23	.21	.16	.40	.38	.08	.23	.26	.16	.16	.27	.28		
3	.23	.23	.09	.11	.25	.33	.35	.07	.10	.26	.25	.24	.17	.31	.25	.24	.47	.17	.11	.33	.11	.17	.16	.27		
4	.22	.29	.10	.10	.25	.28	.33	.04	.03	.26	.23	.24	.21	.33	.28	.36	.52	.20	.18	.36	.19	.24	.26	.34		
5	.27	.31	.12	.12	.31	.28	.49	.19	.07	.23	.18	.31	.26	.39	.30	.44	.56	.24	.18	.31	.19	.30	.23	.38		
6	.27	.31	.13	.12	.30	.25	.52	.23	.03	.31	.20	.28	.19	.40	.33	.42	.57	.24	.18	.36	.20	.21	.23	.38		
7	.27	.27	.11	.13	.31	.29	.50	.18	.11	.29	.16	.22	.28	.39	.32	.37	.55	.17	.11	.29	.16	.22	.24	.30		
8	.21	.22	.09	.10	.23	.19	.33	.03	.11	.21	.24	.23	.24	.30	.14	.30	.38	.13	.16	.27	.14	.11	.29	.32		
9	.22	.24	.10	.10	.23	.19	.35	.02	.11	.24	.22	.27	.18	.35	.18	.37	.38	.13	.18	.28	.15	.13	.25	.32		
10	.26	.31	.12	.12	.29	.25	.49	.11	.09	.29	.18	.28	.24	.36	.32	.39	.56	.21	.21	.37	.16	.22	.30	.38		
11	.21	.26	.10	.09	.23	.23	.34	.06	.05	.24	.16	.21	.25	.33	.26	.27	.45	.17	.22	.33	.18	.15	.28	.33		
12	.19	.24	.11	.13	.24	.15	.34	-.02	.06	.25	.23	.16	.17	.29	.22	.36	.45	.06	.14	.29	.15	.09	.32	.33		
13	.18	.24	.08	.10	.18	.17	.33	.06	.05	.22	.15	.22	.17	.26	.20	.32	.43	.11	.16	.25	.12	.23	.32	.27		
14	.19	.21	.10	.10	.23	.21	.34	.00	.04	.23	.19	.25	.21	.27	.19	.27	.44	.12	.17	.24	.10	.12	.25	.25		
15	.20	.23	.11	.10	.23	.13	.40	.00	.08	.23	.18	.24	.18	.28	.22	.29	.45	.15	.12	.25	.14	.16	.22	.27		
16	.23	.28	.10	.10	.24	.27	.42	.06	.08	.25	.23	.23	.25	.31	.27	.38	.50	.16	.22	.34	.19	.18	.29	.32		
17	.20	.24	.11	.10	.23	.15	.39	.02	.02	.23	.18	.25	.22	.29	.22	.29	.46	.14	.17	.29	.15	.11	.25	.29		

SC = Supervised Classifier

1 = Naïve Bayes, 2 = Native Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion, B = Boredom-Frustration, C = Delight-Confusion, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Frustration, G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

Table 37. Two-way classification for judgments where trained judges agree.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)			
	P	C	P	A	B	C	D	E	F	G	H	I	J
1	.24	.23	.10	.07	.23	.27	.37	.18	.26	.29	.03	.35	.15
2	.24	.23	.10	.07	.23	.26	.37	.17	.27	.29	.02	.36	.16
3	.27	.21	.11	.11	.22	.28	.45	.14	.28	.30	.09	.43	.19
4	.27	.32	.13	.10	.22	.26	.45	.12	.30	.32	.08	.45	.18
5	.31	.40	.07	.09	.28	.29	.42	.37	.24	.31	.22	.41	.22
6	.32	.40	.09	.10	.32	.27	.49	.35	.23	.36	.20	.40	.26
7	.31	.35	.07	.09	.32	.33	.47	.31	.25	.31	.20	.38	.24
8	.19	.23	.08	.09	.14	.14	.28	.14	.26	.19	.04	.27	.24
9	.24	.27	.09	.08	.20	.23	.33	.19	.31	.27	.06	.33	.15
10	.31	.34	.09	.09	.27	.33	.44	.25	.36	.35	.13	.38	.22
11	.24	.29	.09	.09	.22	.27	.37	.19	.30	.30	.08	.31	.12
12	.18	.24	.12	.09	.19	.13	.36	.03	.24	.21	-.04	.28	.19
13	.21	.26	.09	.08	.17	.19	.33	.20	.23	.21	.05	.35	.13
14	.20	.27	.09	.08	.19	.18	.33	.18	.23	.28	.01	.26	.13
15	.21	.27	.09	.08	.24	.17	.37	.20	.22	.29	.02	.24	.13
16	.27	.30	.12	.09	.22	.25	.43	.21	.29	.30	.04	.44	.22
17	.22	.23	.08	.08	.21	.18	.33	.18	.26	.27	.04	.28	.18

SC = Supervised Classifier

1 = Naïve Bayes, 2 = Native Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Decision Trees, 17 = REP Tree.

A = Boredom-Confusion, B = Boredom-Delight, C = Boredom-Frustration, D = Boredom-Frustration, E = Confusion-Delight, F = Confusion-Frustration  
 G = Confusion-Frustration, H = Delight-Flow, I = Delight-Frustration, J = Flow-Frustration

*Detailed classification results for 3-way classifications*

Table 38. Three-way classification for self judgments.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)					
	P	A	B	C	D	E	F	G	H	I	J	A	B	C	
												C	P	A	
1	.11	.18	.03	.02	.12	.13	.18	.12	.10	.06	.09	.13	.07	.18	.20
2	.11	.17	.04	.02	.13	.13	.18	.10	.12	.05	.09	.13	.06	.17	.20
3	.13	.17	.02	.06	.15	.14	.15	.14	.11	.09	.11	.13	.14	.13	.24
4	.11	.22	.02	.05	.13	.14	.15	.10	.10	.09	.08	.11	.13	.19	.25
5	.22	.24	.04	.04	.27	.21	.21	.26	.25	.18	.17	.21	.24	.27	.23
6	.21	.22	.04	.04	.24	.26	.24	.22	.21	.16	.14	.19	.20	.23	.24
7	.21	.20	.03	.03	.24	.22	.23	.25	.23	.19	.14	.18	.21	.20	.23
8	.09	.16	.02	.01	.08	.08	.12	.11	.11	.08	.07	.08	.09	.11	.17
9	.05	.08	.01	.03	.07	.06	.07	.05	.07	.04	.03	.04	.05	.07	.10
10	.19	.24	.04	.04	.24	.21	.24	.22	.22	.15	.12	.16	.22	.16	.23
11	.17	.22	.04	.03	.20	.18	.21	.20	.22	.12	.12	.12	.18	.16	.21
12	.07	.16	.05	.03	.07	.10	.15	.05	.10	-.01	.01	.02	.09	.08	.14
13	.13	.18	.03	.02	.15	.15	.16	.16	.16	.10	.07	.10	.14	.13	.17
14	.13	.17	.03	.02	.17	.12	.15	.15	.09	.08	.10	.14	.11	.18	.17
15	.14	.17	.03	.03	.17	.13	.17	.15	.19	.10	.12	.15	.13	.19	.18
16	.15	.20	.04	.04	.16	.18	.20	.15	.16	.11	.09	.12	.18	.13	.24
17	.13	.17	.03	.03	.15	.14	.18	.13	.17	.10	.08	.10	.15	.11	.16

1 = Naive Bayes, 2 = Naive Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion-Delight, B = Boredom-Confusion-Frustration, C = Boredom-Confusion-Flow, D = Boredom-Delight-Frustration, E = Boredom-Delight-Flow-Frustration, F = Delight-Flow-Frustration, G = Confusion-Delight-Flow, H = Confusion-Flow-Frustration, I = Confusion-Flow-Frustration, J = Delight-Delight-Frustration

*Detailed classification results for 3-way classifications*

Table 38. Three-way classification for self judgments.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)					
	P	A	B	C	D	E	F	G	H	I	J	A	B	C	
												C	P	A	
1	.11	.18	.03	.02	.12	.13	.18	.12	.10	.06	.09	.13	.07	.18	.20
2	.11	.17	.04	.02	.13	.13	.18	.10	.12	.05	.09	.13	.06	.17	.20
3	.13	.17	.02	.06	.15	.14	.15	.14	.11	.09	.11	.13	.14	.13	.24
4	.11	.22	.02	.05	.13	.14	.15	.10	.10	.09	.08	.11	.13	.19	.25
5	.22	.24	.04	.04	.27	.21	.21	.26	.25	.18	.17	.21	.24	.27	.23
6	.21	.22	.04	.04	.24	.26	.24	.22	.21	.16	.14	.19	.20	.23	.24
7	.21	.20	.03	.03	.24	.22	.23	.25	.23	.19	.14	.18	.21	.20	.23
8	.09	.16	.02	.01	.08	.08	.12	.11	.11	.08	.07	.08	.09	.11	.17
9	.05	.08	.01	.03	.07	.06	.07	.05	.07	.04	.03	.04	.05	.07	.10
10	.19	.24	.04	.04	.24	.21	.24	.22	.22	.15	.12	.16	.22	.16	.23
11	.17	.22	.04	.03	.20	.18	.21	.20	.22	.12	.12	.12	.18	.16	.21
12	.07	.16	.05	.03	.07	.10	.15	.05	.10	-.01	.01	.02	.09	.08	.14
13	.13	.18	.03	.02	.15	.15	.16	.16	.16	.10	.07	.10	.14	.13	.17
14	.13	.17	.03	.02	.17	.12	.15	.15	.09	.08	.10	.14	.11	.18	.17
15	.14	.17	.03	.03	.17	.13	.17	.15	.19	.10	.12	.15	.13	.19	.18
16	.15	.20	.04	.04	.16	.18	.20	.15	.16	.11	.09	.12	.18	.13	.24
17	.13	.17	.03	.03	.15	.14	.18	.13	.17	.10	.08	.10	.15	.11	.16

1 = Naive Bayes, 2 = Naive Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion-Delight, B = Boredom-Confusion-Frustration, C = Boredom-Confusion-Flow, D = Boredom-Delight-Frustration, E = Boredom-Delight-Flow-Frustration, F = Delight-Flow-Frustration, G = Confusion-Delight-Flow, H = Confusion-Flow-Frustration, I = Confusion-Flow-Frustration, J = Delight-Delight-Frustration

**Table 39.** Three-way classification for peer judgments.

SC	MEAN			STDEV			PRESSURE (P)						CONTOURS (C)										
	P	C	A	A B C			D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
				C	P	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
1	.14	.17	.04	.04	.09	.13	.20	.15	.17	.07	.10	.19	.14	.16	.13	.09	.23	.20	.18	.16	.17	.18	.14
2	.14	.17	.04	.03	.08	.14	.20	.20	.14	.09	.10	.18	.13	.13	.14	.10	.23	.18	.17	.16	.18	.19	.17
3	.16	.14	.06	.08	.07	.13	.26	.14	.19	.10	.13	.21	.20	.17	.05	.10	.30	.12	.13	.06	.13	.19	.25
4	.15	.23	.06	.07	.05	.14	.25	.16	.16	.09	.11	.20	.20	.15	.14	.15	.33	.22	.24	.15	.21	.27	.30
5	.19	.22	.06	.06	.13	.15	.28	.18	.23	.07	.15	.22	.27	.19	.17	.15	.32	.19	.27	.14	.20	.24	.31
6	.19	.22	.08	.07	.12	.15	.31	.14	.17	.07	.14	.26	.28	.22	.15	.16	.35	.22	.24	.16	.22	.26	.32
7	.19	.19	.07	.06	.12	.15	.30	.19	.24	.08	.16	.21	.28	.22	.16	.15	.32	.13	.19	.12	.19	.26	.18
8	.12	.16	.03	.04	.08	.12	.16	.10	.11	.08	.10	.15	.17	.12	.11	.09	.19	.17	.19	.17	.14	.17	.20
9	.10	.13	.04	.05	.07	.10	.16	.09	.09	.07	.08	.06	.11	.15	.09	.16	.03	.11	.20	.14	.20	.07	.12
10	.19	.23	.06	.06	.13	.14	.31	.18	.21	.10	.15	.24	.23	.22	.16	.15	.36	.20	.23	.18	.21	.26	.29
11	.17	.20	.06	.06	.11	.12	.25	.15	.20	.08	.14	.21	.22	.19	.13	.13	.32	.21	.21	.13	.19	.23	.27
12	.09	.16	.08	.07	.01	.09	.24	.03	.04	.01	.03	.17	.18	.08	.15	.04	.27	.18	.15	.22	.05	.19	.22
13	.13	.17	.06	.05	.07	.09	.22	.09	.14	.06	.09	.15	.17	.19	.12	.09	.26	.20	.19	.13	.13	.19	.20
14	.13	.14	.04	.05	.08	.10	.21	.12	.17	.08	.11	.15	.16	.13	.09	.09	.25	.14	.17	.08	.12	.17	.18
15	.13	.15	.06	.06	.09	.09	.24	.10	.17	.05	.10	.17	.17	.15	.10	.09	.26	.15	.17	.09	.12	.16	.22
16	.16	.21	.05	.06	.07	.13	.26	.16	.18	.12	.15	.21	.20	.17	.13	.12	.31	.22	.23	.17	.19	.26	.28
17	.14	.16	.05	.06	.11	.11	.24	.10	.15	.06	.13	.19	.19	.18	.12	.09	.28	.16	.16	.15	.11	.19	.21

1 = Naïve Bayes, 2 = Native Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = Nearest Neighbor, 6 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion-Delight, B = Boredom-Confusion-Frustration, C = Boredom-Confusion-Flow, D = Boredom-Delight-Frustration, E = Boredom-Delight-Flow, F = Delight-Flow-Frustration, G = Delight-Flow, H = Confusion-Flow-Frustration, I = Confusion-Flow-Frustration, J = Delight-Flow-Frustration

Table 40. Three-way classification for judgments by trained judge 1.

SC	MEAN			STDEV			PRESSURE (P)						CONTOURS (C)											
	P	C	P	C	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
1	.15	.16	.03	.02	.15	.12	.19	.15	.19	.12	.13	.17	.14	.17	.13	.17	.16	.19	.13	.18	.13	.15	.19	
2	.15	.16	.02	.03	.16	.13	.18	.15	.18	.12	.13	.12	.17	.13	.17	.12	.17	.16	.20	.12	.17	.14	.15	.18
3	.21	.18	.03	.04	.22	.16	.21	.21	.27	.19	.21	.20	.23	.21	.16	.22	.16	.23	.11	.17	.17	.21	.20	
4	.21	.22	.03	.04	.22	.16	.22	.19	.28	.16	.18	.20	.23	.21	.21	.19	.25	.21	.30	.15	.23	.20	.25	.24
5	.27	.29	.04	.05	.28	.25	.26	.30	.34	.23	.22	.22	.31	.28	.27	.24	.28	.33	.38	.24	.27	.23	.31	.33
6	.25	.27	.04	.06	.25	.23	.27	.26	.33	.19	.19	.22	.30	.26	.25	.20	.29	.29	.37	.21	.22	.24	.28	.34
7	.26	.27	.05	.04	.27	.25	.27	.28	.35	.20	.22	.20	.30	.28	.27	.22	.26	.30	.35	.23	.24	.24	.31	.30
8	.12	.15	.02	.03	.11	.09	.12	.09	.16	.10	.13	.13	.15	.14	.16	.12	.17	.16	.19	.10	.14	.16	.17	.15
9	.08	.08	.03	.02	.03	.06	.06	.04	.09	.09	.09	.11	.10	.11	.11	.09	.07	.11	.09	.05	.07	.08	.06	.05
10	.23	.26	.04	.05	.21	.21	.25	.22	.28	.17	.21	.22	.29	.24	.25	.22	.28	.30	.34	.18	.23	.24	.29	.28
11	.20	.21	.03	.03	.18	.19	.21	.17	.24	.15	.17	.20	.24	.21	.23	.17	.24	.22	.26	.16	.20	.19	.24	.23
12	.12	.16	.05	.04	.07	.11	.15	.05	.19	.07	.12	.14	.16	.09	.17	.13	.20	.17	.21	.08	.16	.15	.18	.14
13	.17	.19	.03	.04	.18	.15	.19	.18	.21	.14	.13	.15	.20	.17	.17	.15	.21	.21	.24	.12	.19	.16	.21	.23
14	.13	.18	.02	.03	.12	.13	.14	.12	.17	.11	.10	.14	.18	.13	.17	.15	.18	.21	.24	.14	.16	.16	.20	.20
15	.16	.18	.03	.04	.15	.15	.18	.14	.21	.11	.13	.15	.20	.14	.18	.16	.19	.21	.25	.14	.15	.15	.20	.23
16	.20	.21	.03	.04	.18	.17	.21	.19	.27	.17	.19	.20	.24	.20	.19	.18	.24	.22	.26	.14	.19	.20	.23	.24
17	.15	.17	.03	.03	.13	.13	.18	.13	.19	.11	.13	.14	.18	.15	.17	.14	.19	.18	.20	.10	.16	.17	.19	.19

<sup>1</sup> = Naïve Bayes, <sup>2</sup> = Native Bayes Updatable, <sup>3</sup> = Support Vector Machines, <sup>4</sup> = Logistic Regression, <sup>5</sup> = Nearest Neighbor, <sup>6</sup> = Nearest Neighbors, <sup>7</sup> = K\*, <sup>8</sup> = Locally Weighted Learning, <sup>9</sup> = AdaBoost, <sup>10</sup> = Bagging Predictors, <sup>11</sup> = Additive Logistic Regression, <sup>12</sup> = Decision Tables, <sup>13</sup> = Nearest Neighbor Generalization, <sup>14</sup> = PART, <sup>15</sup> = C4.5 Decision Trees, <sup>16</sup> = Logistic Model Trees, <sup>17</sup> = REP Tree.

A = Boredom-Confusion-Delight, B = Boredom-Confusion-Frustration, C = Boredom-Confusion-Flow, D = Boredom-Delight-Frustration, E = Boredom-Delight-Flow, F = Confusion-Delight-Frustration, G = Confusion-Confusion-Flow, H = Confusion-Flow-Frustration, I = Delight-Flow-Frustration, J = Delight-Flow-Frustration

Table 41. Three-way classification for judgments by trained judge 2.

SC	MEAN			STDEV			PRESSURE (P)						CONTOURS (C)												
	P	C	P	C	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J	
1	.13	.13	.04	.03	.15	.07	.16	.13	.18	.08	.11	.14	.13	.15	.11	.15	.11	.18	.07	.15	.11	.15	.15		
2	.14	.14	.03	.03	.15	.09	.17	.14	.17	.10	.19	.11	.13	.16	.11	.14	.11	.19	.09	.16	.11	.16	.15		
3	.21	.14	.05	.03	.20	.12	.23	.20	.28	.17	.26	.18	.22	.12	.13	.20	.13	.18	.10	.12	.12	.16	.15		
4	.20	.22	.04	.05	.19	.12	.23	.19	.26	.18	.26	.18	.20	.19	.19	.23	.20	.31	.13	.21	.20	.27	.24		
5	.24	.25	.04	.06	.20	.19	.24	.27	.31	.19	.24	.21	.26	.25	.20	.19	.26	.37	.19	.24	.20	.28	.32		
6	.23	.24	.04	.07	.19	.18	.26	.27	.29	.19	.24	.17	.26	.26	.21	.17	.27	.21	.37	.15	.25	.21	.26	.31	
7	.22	.24	.04	.06	.21	.16	.22	.24	.29	.16	.23	.16	.25	.26	.19	.20	.24	.33	.17	.22	.20	.30	.32		
8	.13	.14	.03	.04	.13	.08	.13	.12	.13	.16	.16	.09	.11	.13	.11	.11	.19	.11	.20	.09	.15	.15	.16	.14	
9	.11	.07	.04	.02	.14	.04	.12	.10	.11	.16	.13	.05	.08	.15	.05	.05	.08	.06	.11	.05	.07	.07	.08	.10	
10	.20	.23	.04	.05	.18	.15	.24	.20	.24	.16	.24	.17	.22	.25	.18	.16	.27	.22	.30	.16	.24	.18	.27	.27	
11	.19	.20	.04	.06	.20	.12	.22	.19	.23	.16	.23	.15	.21	.20	.16	.14	.24	.18	.28	.12	.22	.17	.25	.25	
12	.10	.12	.06	.06	.13	.01	.14	.08	.16	.12	.17	.04	.05	.13	.08	.06	.20	.08	.20	.03	.14	.11	.17	.12	
13	.15	.17	.03	.05	.15	.11	.18	.14	.20	.13	.17	.10	.17	.18	.13	.12	.20	.15	.24	.11	.21	.13	.19	.21	
14	.13	.16	.03	.04	.12	.06	.14	.13	.16	.10	.14	.10	.16	.15	.16	.12	.17	.16	.23	.11	.17	.13	.18	.20	
15	.14	.16	.03	.04	.13	.12	.17	.14	.18	.14	.17	.09	.15	.14	.11	.18	.16	.24	.11	.19	.12	.20	.20		
16	.20	.19	.04	.04	.17	.13	.23	.19	.25	.18	.24	.18	.20	.15	.17	.22	.19	.26	.13	.19	.16	.25	.22		
17	.14	.15	.03	.04	.14	.10	.16	.15	.11	.16	.11	.17	.11	.13	.18	.10	.11	.20	.12	.19	.10	.17	.13	.19	.17

1 = Naive Bayes, 2 = Naive Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\* 8 = Locally Weighted Learning, 9 = AddBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion-Delight, B = Boredom-Confusion-Frustration, C = Boredom-Confusion-Flow, D = Boredom-Delight-Frustration, E = Boredom-Delight-Flow, F = Confusion-Delight-Frustration, G = Confusion-Confusion-Flow, H = Confusion-Flow-Frustration, I = Boredom-Flow-Frustration, J = Delight-Flow-Frustration

**Table 42.** Three-way classification for judgments where trained judges agree.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)														
	P	C	P	C	A	B	C	D	E	F	G	H	I	J	A	B	C	D	E	F	G	H	I	J
1	.19	.18	.06	.04	.19	.08	.26	.18	.28	.11	.25	.15	.21	.20	.20	.14	.20	.14	.25	.14	.23	.15	.16	.21
2	.19	.18	.06	.04	.20	.08	.25	.16	.27	.10	.26	.14	.21	.21	.20	.15	.20	.14	.26	.13	.22	.16	.15	.21
3	.22	.16	.07	.06	.20	.12	.28	.15	.31	.15	.31	.19	.20	.25	.09	.13	.27	.11	.20	.09	.18	.14	.22	.19
4	.21	.26	.07	.05	.22	.10	.28	.14	.31	.14	.29	.18	.22	.25	.21	.23	.30	.20	.35	.19	.27	.25	.31	.30
5	.25	.32	.05	.06	.24	.22	.28	.27	.35	.15	.27	.24	.28	.25	.28	.28	.37	.31	.44	.22	.32	.29	.33	.38
6	.25	.31	.05	.07	.21	.19	.33	.24	.33	.19	.26	.21	.27	.25	.29	.25	.39	.30	.46	.21	.31	.26	.32	.36
7	.23	.28	.06	.05	.21	.16	.30	.22	.35	.16	.24	.19	.24	.21	.23	.23	.31	.26	.37	.19	.30	.27	.33	.32
8	.13	.17	.04	.05	.15	.06	.15	.10	.17	.13	.18	.09	.11	.14	.13	.14	.21	.15	.22	.08	.21	.21	.21	.18
9	.09	.09	.04	.04	.11	.02	.11	.06	.14	.12	.13	.07	.06	.12	.04	.03	.13	.05	.14	.05	.12	.09	.11	.12
10	.23	.28	.06	.06	.22	.13	.31	.21	.33	.15	.28	.19	.24	.25	.23	.22	.33	.26	.36	.17	.32	.27	.32	.30
11	.20	.26	.07	.05	.19	.09	.27	.17	.30	.13	.29	.18	.20	.22	.22	.22	.31	.23	.34	.16	.28	.26	.30	.27
12	.10	.16	.08	.06	.13	-.02	.18	.03	.21	.11	.17	.03	.04	.17	.09	.10	.24	.11	.23	.05	.20	.17	.21	.19
13	.16	.22	.05	.06	.15	.12	.22	.15	.24	.09	.21	.12	.16	.18	.20	.15	.27	.19	.29	.11	.26	.20	.26	.27
14	.15	.20	.05	.05	.17	.07	.23	.13	.21	.08	.20	.13	.17	.16	.16	.25	.20	.27	.12	.23	.19	.21	.22	
15	.16	.21	.05	.04	.14	.09	.23	.15	.23	.10	.20	.14	.19	.13	.19	.17	.26	.22	.26	.11	.24	.21	.24	.22
16	.21	.24	.07	.06	.18	.10	.27	.16	.31	.13	.31	.18	.20	.24	.19	.24	.27	.19	.34	.13	.26	.23	.27	.27
17	.16	.18	.05	.04	.16	.09	.21	.13	.23	.09	.21	.13	.17	.17	.14	.12	.23	.17	.22	.11	.20	.18	.21	.21

1 = Naive Bayes, 2 = Naive Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AddBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5 Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion-Delight, B = Boredom-Confusion-Frustration, C = Boredom-Confusion-Flow, D = Boredom-Delight-Frustration, E = Boredom-Delight-Flow, F = Delight-Flow-Frustration, G = Confusion-Delight-Flow, H = Confusion-Flow-Frustration, I = Boredom-Flow-Frustration, J = Delight-Flow-Frustration

*Detailed classification results for 4-way classifications*

Table 43. Four-way classification for judgments by self judgments.

SC	MEAN			STDDEV			PRESSURE (P)			CONTOURS (C)		
	P	C	P	P	C	C	A	B	C	D	E	C
1	.10	.15	.01	.02	.10	.11	.12	.10	.10	.17	.14	.16
2	.11	.15	.01	.02	.11	.11	.09	.10	.10	.17	.14	.17
3	.11	.13	.01	.04	.10	.13	.11	.12	.11	.12	.10	.11
4	.10	.20	.01	.02	.08	.12	.10	.11	.10	.18	.19	.23
5	.19	.21	.02	.01	.21	.20	.16	.21	.17	.22	.20	.22
6	.17	.19	.01	.01	.17	.16	.19	.18	.16	.19	.18	.21
7	.17	.19	.02	.02	.18	.18	.17	.18	.15	.20	.19	.18
8	.08	.14	.01	.02	.07	.07	.07	.08	.09	.13	.13	.15
9	.04	.05	.01	.01	.04	.04	.05	.04	.02	.07	.05	.04
10	.16	.20	.02	.02	.15	.17	.17	.18	.14	.20	.21	.22
11	.14	.18	.02	.02	.12	.15	.15	.15	.12	.18	.18	.21
12	.04	.12	.01	.03	.05	.04	.05	.04	.02	.12	.12	.16
13	.12	.16	.01	.01	.11	.13	.12	.13	.09	.17	.17	.16
14	.11	.14	.01	.01	.10	.11	.10	.13	.10	.16	.14	.13
15	.12	.14	.02	.02	.10	.12	.12	.15	.09	.16	.15	.14
16	.12	.19	.02	.02	.11	.12	.15	.14	.10	.19	.19	.16
17	.11	.14	.01	.02	.09	.11	.12	.11	.10	.14	.15	.16

SC = Supervised Classifier, 1 = Naïve Bayes, 2 = Naïve Bayes Updatable, 3 = Logistic Regression, 4 = Support Vector Machines, 5 = 1 Nearest Neighbor, 6 = 3 Nearest Neighbors, 7 = K\*, 8 = Locally Weighted Learning, 9 = AdaBoost, 10 = Bagging Predictors, 11 = Additive Logistic Regression, 12 = Decision Tables, 13 = Nearest Neighbor Generalization, 14 = PART, 15 = C4.5, Decision Trees, 16 = Logistic Model Trees, 17 = REP Tree.

A = Boredom-Confusion-Delight-Flow, B = Boredom-Confusion-Delight-Frustration, C = Boredom-Confusion-Flow-Frustration, D = Boredom -Delight-Flow-Frustration, E = Confusion-Delight-Flow-Frustration

Table 44. Four-way classification for judgments by peer judges.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D	E
1	.09	.13	.02	.02	.07	.08	.13	.08	.09	.11	.14	.13	.16	.13	.16	.13
2	.10	.13	.02	.02	.08	.09	.13	.08	.11	.10	.13	.13	.15	.12	.15	.12
3	.10	.09	.03	.05	.08	.09	.15	.11	.07	.07	.05	.17	.09	.08	.09	.08
4	.10	.18	.04	.03	.09	.07	.17	.09	.10	.14	.18	.22	.20	.19	.22	.15
5	.14	.17	.03	.04	.09	.14	.18	.16	.13	.12	.17	.20	.22	.15	.20	.15
6	.12	.17	.05	.04	.08	.12	.20	.12	.10	.11	.17	.21	.18	.16	.18	.16
7	.14	.14	.03	.04	.11	.15	.17	.13	.12	.08	.15	.17	.17	.16	.17	.16
8	.08	.12	.02	.02	.07	.06	.11	.08	.10	.12	.10	.14	.12	.10	.14	.10
9	.07	.10	.02	.02	.07	.03	.10	.07	.07	.07	.14	.09	.08	.10	.09	.09
10	.13	.17	.03	.04	.10	.14	.17	.14	.11	.12	.16	.20	.20	.16	.20	.16
11	.10	.15	.03	.02	.08	.08	.16	.11	.09	.12	.14	.18	.17	.14	.18	.17
12	.03	.10	.05	.04	.04	.00	.11	.00	.01	.10	.05	.15	.11	.08	.15	.10
13	.09	.14	.02	.01	.08	.07	.13	.10	.08	.12	.14	.15	.15	.15	.15	.15
14	.08	.10	.03	.02	.05	.10	.12	.06	.09	.07	.10	.13	.12	.11	.12	.11
15	.09	.11	.02	.03	.07	.10	.12	.07	.08	.06	.09	.13	.12	.13	.12	.13
16	.10	.15	.03	.04	.07	.10	.15	.10	.07	.12	.12	.20	.18	.15	.18	.15
17	.09	.11	.03	.02	.08	.08	.13	.10	.05	.10	.10	.15	.10	.15	.10	.11

SC = Supervised Classifier

- 1 = Naïve Bayes,
- 2 = Naïve Bayes Updatable,
- 3 = Logistic Regression,
- 4 = Support Vector Machines,
- 5 = 1 Nearest Neighbor,
- 6 = 3 Nearest Neighbors,
- 7 = K\*,
- 8 = Locally Weighted Learning,
- 9 = AdaBoost,
- 10 = Bagging Predictors,
- 11 = Additive Logistic Regression,
- 12 = Decision Tables,
- 13 = Nearest Neighbor Generalization,
- 14 = PART,
- 15 = C4.5 Decision Trees,
- 16 = Logistic Model Trees,
- 17 = REP Tree.

A = Boredom-Confusion-Delight-Flow,      B = Boredom-Confusion-Delight-Frustration,  
 C = Boredom-Confusion-Flow-Frustration,    D = Boredom -Delight-Flow-Frustration,  
 E = Confusion-Delight-Flow-Frustration

Table 45. Four-way classification for judgments by trained judge 1.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)				
	P	C	P	C	A	B	C	D	E	A	B	C		
1	.13	.13	.01	.02	.12	.14	.13	.12	.12	.13	.14	.11	.16	.13
2	.13	.13	.01	.01	.12	.14	.13	.14	.12	.13	.12	.12	.16	.13
3	.17	.15	.02	.02	.16	.17	.15	.19	.18	.12	.16	.16	.18	.15
4	.17	.19	.01	.03	.17	.18	.15	.19	.16	.16	.20	.19	.24	.18
5	.23	.24	.02	.03	.22	.23	.22	.27	.20	.23	.25	.22	.29	.23
6	.21	.23	.02	.04	.19	.21	.21	.23	.18	.20	.24	.20	.28	.21
7	.22	.23	.03	.03	.21	.21	.22	.27	.19	.23	.23	.21	.27	.21
8	.10	.13	.01	.01	.07	.10	.10	.11	.09	.13	.13	.13	.14	.11
9	.05	.06	.02	.02	.03	.06	.06	.07	.06	.08	.07	.05	.08	.03
10	.19	.22	.02	.03	.17	.18	.20	.22	.18	.19	.23	.22	.27	.20
11	.16	.17	.02	.02	.14	.17	.18	.18	.15	.16	.19	.18	.20	.15
12	.08	.12	.04	.03	.02	.08	.11	.10	.07	.12	.14	.13	.16	.08
13	.15	.18	.02	.02	.13	.15	.15	.17	.13	.16	.19	.17	.21	.16
14	.12	.15	.02	.03	.12	.13	.13	.14	.10	.12	.15	.13	.19	.14
15	.12	.15	.02	.03	.11	.13	.13	.14	.09	.13	.17	.13	.19	.13
16	.16	.18	.01	.02	.15	.16	.16	.17	.17	.16	.19	.19	.21	.16
17	.12	.14	.02	.02	.10	.13	.13	.14	.11	.11	.15	.13	.17	.12

SC = Supervised Classifier

- 1 = Naïve Bayes,
- 2 = Naïve Bayes Updatable,
- 3 = Logistic Regression,
- 4 = Support Vector Machines,
- 5 = 1 Nearest Neighbor,
- 6 = 3 Nearest Neighbors,
- 7 = K\*,
- 8 = Locally Weighted Learning,
- 9 = AdaBoost,
- 10 = Bagging Predictors,
- 11 = Additive Logistic Regression,
- 12 = Decision Tables,
- 13 = Nearest Neighbor Generalization,
- 14 = PART,
- 15 = C4.5 Decision Trees,
- 16 = Logistic Model Trees,
- 17 = REP Tree.

A = Boredom-Confusion-Delight-Flow,  
 B = Boredom-Confusion-Delight-Frustration,  
 C = Boredom-Confusion-Flow-Frustration,  
 D = Boredom -Delight-Flow-Frustration  
 E = Confusion-Delight-Flow-Frustration

Table 46. Four-way classification for judgments by trained judge 2.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)					
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D
1	.11	.11	.03	.02	.06	.14	.10	.12	.12	.08	.13	.13	.13	.13	.11
2	.11	.11	.03	.02	.07	.15	.10	.12	.12	.08	.12	.12	.13	.12	.12
3	.18	.13	.03	.02	.14	.21	.16	.20	.19	.10	.13	.14	.15	.15	.12
4	.18	.18	.03	.03	.14	.21	.15	.20	.19	.13	.18	.19	.22	.17	.17
5	.20	.21	.02	.03	.18	.21	.18	.23	.20	.18	.23	.19	.25	.20	.20
6	.18	.20	.02	.04	.16	.19	.16	.21	.19	.15	.23	.19	.24	.19	.19
7	.18	.21	.03	.04	.15	.18	.16	.22	.17	.17	.21	.20	.27	.21	.21
8	.09	.11	.02	.02	.08	.11	.06	.09	.10	.07	.13	.12	.13	.10	.10
9	.08	.06	.02	.02	.08	.10	.04	.08	.09	.03	.08	.06	.07	.06	.06
10	.17	.18	.03	.03	.14	.20	.14	.20	.18	.13	.19	.19	.21	.19	.21
11	.16	.17	.02	.03	.14	.18	.13	.16	.16	.12	.16	.17	.22	.15	.15
12	.08	.08	.04	.03	.05	.12	.02	.09	.11	.03	.09	.07	.11	.09	.09
13	.12	.15	.01	.03	.10	.14	.12	.13	.12	.10	.18	.14	.18	.14	.14
14	.12	.12	.02	.03	.09	.12	.10	.15	.13	.08	.12	.13	.17	.12	.12
15	.11	.13	.02	.03	.10	.13	.09	.12	.11	.09	.15	.12	.16	.11	.11
16	.17	.16	.03	.03	.12	.20	.15	.18	.18	.12	.16	.17	.20	.16	.16
17	.10	.12	.03	.02	.07	.12	.07	.13	.13	.08	.14	.14	.12	.12	.12

SC = Supervised Classifier

- 1 = Naïve Bayes,
- 2 = Naïve Bayes Updatable,
- 3 = Logistic Regression,
- 4 = Support Vector Machines,
- 5 = 1 Nearest Neighbor,
- 6 = 3 Nearest Neighbors,
- 7 = K\*,
- 8 = Locally Weighted Learning,
- 9 = AdaBoost,
- 10 = Bagging Predictors,
- 11 = Additive Logistic Regression,
- 12 = Decision Tables,
- 13 = Nearest Neighbor Generalization,
- 14 = PART,
- 15 = C4.5 Decision Trees,
- 16 = Logistic Model Trees,
- 17 = REP Tree.

- A = Boredom-Confusion-Delight-Flow,
- B = Boredom-Confusion-Delight-Frustration,
- C = Boredom-Confusion-Flow-Frustration, D = Boredom -Delight-Flow-Frustration,
- E = Confusion-Delight-Flow-Frustration

**Table 47.** Four-way classification for judgments where trained judges agree.

SC	MEAN			STDEV			PRESSURE (P)			CONTOURS (C)						
	P	C	P	C	P	C	A	B	C	D	E	A	B	C	D	E
1	.16	.16	.03	.03	.12	.19	.15	.20	.15	.13	.20	.14	.18	.15	.18	.15
2	.16	.16	.03	.03	.11	.18	.15	.20	.15	.12	.20	.14	.18	.16	.18	.16
3	.18	.13	.04	.02	.12	.21	.17	.20	.19	.11	.13	.13	.16	.12	.16	.12
4	.18	.23	.04	.03	.12	.22	.19	.20	.18	.18	.24	.23	.27	.21	.27	.21
5	.21	.27	.03	.04	.19	.22	.22	.25	.18	.21	.28	.27	.32	.27	.32	.27
6	.20	.28	.02	.04	.18	.23	.21	.22	.19	.22	.31	.26	.33	.27	.33	.27
7	.19	.25	.03	.03	.16	.23	.19	.21	.16	.21	.26	.24	.30	.25	.30	.25
8	.09	.13	.01	.03	.08	.11	.08	.09	.09	.08	.14	.17	.15	.13	.15	.13
9	.06	.06	.01	.03	.06	.07	.05	.07	.07	.02	.07	.07	.09	.04	.09	.04
10	.19	.24	.03	.04	.15	.23	.19	.21	.19	.18	.26	.24	.30	.23	.30	.23
11	.17	.22	.03	.05	.13	.21	.17	.19	.15	.16	.24	.22	.28	.20	.28	.20
12	.07	.11	.03	.04	.04	.11	.02	.08	.08	.04	.13	.12	.16	.11	.16	.11
13	.14	.19	.03	.03	.11	.17	.13	.16	.12	.15	.21	.17	.23	.19	.23	.19
14	.13	.17	.03	.03	.10	.16	.11	.16	.12	.14	.19	.18	.21	.16	.21	.16
15	.12	.18	.03	.04	.10	.17	.12	.14	.10	.14	.20	.16	.24	.15	.24	.15
16	.17	.21	.04	.03	.12	.21	.15	.19	.18	.16	.23	.22	.25	.19	.25	.19
17	.13	.15	.03	.04	.09	.16	.11	.14	.12	.10	.18	.15	.19	.14	.15	.14

SC = Supervised Classifier

- 1 = Naive Bayes,
- 2 = Naive Bayes Updatable,
- 3 = Logistic Regression,
- 4 = Support Vector Machines,
- 5 = Nearest Neighbor,
- 6 = 3 Nearest Neighbors,
- 7 = K\*,
- 8 = Locally Weighted Learning,
- 9 = AdaBoost,
- 10 = Bagging Predictors,
- 11 = Additive Logistic Regression,
- 12 = Decision Tables,
- 13 = Nearest Neighbor Generalization,
- 14 = PART,
- 15 = C4.5 Decision Trees,
- 16 = Logistic Model Trees,
- 17 = REP Tree.

A = Boredom-Confusion-Delight-Flow,      B = Boredom-Confusion-Delight-Frustration,  
 C = Boredom-Confusion-Flow-Frustration,      D = Boredom -Delight-Flow-Frustration,  
 E = Confusion-Delight-Flow-Frustration

*Detailed classification results for 5-way classifications*

Table 48. Five-way classification accuracies.

SC	SELF			PEER			JDG1			JDG2			JDGA		
	P	G	P	G	P	G	P	G	P	G	P	G	P	G	
1	.05	.12	.09	.10	.11	.13	.11	.12	.16	.14	.16	.14	.16	.14	.14
2	.06	.13	.07	.11	.11	.13	.10	.11	.16	.14	.16	.14	.16	.14	.14
3	.09	.09	.10	.04	.15	.13	.15	.11	.16	.12	.16	.12	.16	.12	.12
4	.07	.16	.09	.15	.15	.18	.14	.16	.16	.21	.16	.21	.16	.21	.21
5	.15	.19	.14	.13	.20	.20	.17	.18	.19	.24	.19	.24	.19	.24	.24
6	.15	.17	.10	.13	.17	.19	.17	.17	.18	.24	.18	.24	.18	.24	.24
7	.14	.17	.13	.09	.20	.21	.15	.18	.16	.23	.16	.23	.16	.23	.23
8	.05	.11	.09	.11	.08	.11	.08	.10	.09	.11	.09	.11	.09	.11	.11
9	.02	.04	.07	.08	.05	.05	.05	.06	.05	.05	.06	.05	.06	.05	.05
10	.13	.16	.13	.14	.16	.20	.15	.18	.18	.23	.18	.23	.18	.23	.23
11	.10	.15	.12	.10	.14	.16	.13	.16	.16	.21	.16	.21	.16	.21	.21
12	.00	.08	.03	.06	.06	.09	.04	.06	.07	.09	.07	.09	.07	.09	.09
13	.10	.15	.08	.12	.12	.16	.11	.15	.12	.18	.12	.18	.12	.18	.18
14	.11	.12	.09	.08	.10	.13	.10	.11	.12	.15	.12	.15	.12	.15	.15
15	.11	.11	.09	.08	.10	.13	.09	.11	.11	.15	.11	.15	.11	.15	.15
16	.10	.15	.11	.11	.14	.15	.15	.14	.15	.20	.16	.20	.16	.20	.20
17	.09	.11	.12	.09	.10	.13	.10	.10	.10	.11	.10	.10	.10	.10	.11

SC = Supervised Classifier

- 1 = Naïve Bayes,
- 2 = Naïve Bayes Updatable,
- 3 = Logistic Regression,
- 4 = Support Vector Machines,
- 5 = 1 Nearest Neighbor,
- 6 = 3 Nearest Neighbors,
- 7 = K\*,
- 8 = Locally Weighted Learning,
- 9 = AdaBoost,
- 10 = Bagging Predictors,
- 11 = Additive Logistic Regression,
- 12 = Decision Tables,
- 13 = Nearest Neighbor Generalization,
- 14 = PART,
- 15 = C4.5 Decision Trees,
- 16 = Logistic Model Trees,
- 17 = REP Tree.

Classifications are between boredom, confusion, delight, flow, and frustration, P = Pressure, C = Contours, JDG1 = Trained Judge 1, JDG2 = Trained Judge 2, JDGA = Trained Judges Agree