Accuracy vs. Availability Heuristic in Multimodal Affect Detection in the Wild

Nigel Bosch[1], Huili Chen[1], Ryan Baker[2], Valerie Shute[3], & Sidney D'Mello[1]
[1]University of Notre Dame, Notre Dame, IN 46556, USA
[2]Teachers College, Columbia University, New York, NY 10027, USA
[3]Florida State University, Tallahassee, FL 32306-4453, USA
pbosch1@nd.edu, hchen6@nd.edu, baker2@tc.columbia.edu, vshute@fsu.edu, sdmello@nd.edu

ABSTRACT
This paper discusses multimodal affect detection from a fusion of facial expressions and interaction features derived from students' interactions with an educational game in the noisy real-world context of a computer-enabled classroom. Log data of students' interactions with the game and face videos from 133 students were recorded in a computer-enabled classroom over a two day period. Human observers live annotated learning-centered affective states such as engagement, confusion, and frustration. The face-only detectors were more accurate than interaction-only detectors. Multimodal affect detectors did not show any substantial improvement in accuracy over the face-only detectors. However, the face-only detectors were only applicable to 65% of the cases due to face registration errors caused by excessive movement, occlusion, poor lighting, and other factors. Multimodal fusion techniques were able to improve the applicability of detectors to 98% of cases without sacrificing classification accuracy. Balancing the accuracy vs. applicability tradeoff appears to be an important feature of multimodal affect detection.

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems – *human factors*.

General Terms
Measurement, Performance, Experimentation, Human Factors.

Author Keywords
Missing data; Affect; Affect detection; Facial expressions; Interaction.

1. INTRODUCTION
Affect sensitivity has been shown to have considerable promise for improving learning in computerized educational environments [22]. There are many ways in which an interface can leverage knowledge of students' affective states to improve learning. For example, bored and disengaged students may be directed to new learning tasks to help them re-engage, while students who are frequently frustrated due to excessively difficult tasks might be presented with material more appropriate for their knowledge and skills.

Affect sensitivity requires affect detection. Consequently, there have been many proposed methods for affect detection [25,33], of which multimodal affect detection is of interest here. D'Mello and Kory [25] recently provided a review and meta-analysis of 90 multimodal affect detection system. Their review revealed that two of the most common modalities for affect detection were facial features and audio features, which were used in 76.7% (face) and 82.2% (audio) of surveyed studies. Thus, audiovisual affect detection is clearly the most prominent multimodal fusion approach. Taking a somewhat different approach, the current paper presents affect detection results from a fusion of facial and interaction features (e.g., clicks, response times).

1.1 Related Work
Face-based affect detection has been well studied in a variety of learning contexts in recent years. A variety of computer vision techniques have been used for affect detection, such as texture, shape, and motion features extracted from faces. The full scope of work is beyond this paper, but recent review articles cover face-based affect detection in some detail [8,25,33].

Interaction log-based affect detection are less common, but have been increasingly studied over the last decade [3–5,24,28,29]. Unlike physical sensor-based detectors, which rely upon the nonverbal responses, these detectors infer affective states from students' interactions with computerized learning systems. We expect these features to be useful for affect detection because a student's affect could alter how they interact with the learning environment. For example, a bored student might not interact very frequently, or might aimlessly repeat an action. An engaged student might be more likely to try various strategies in the game and succeed more. The fact that interaction-based affect detectors rely on logs of student actions makes it possible for them to be deployed at no extra cost to a school that is using the learning system. Their unobtrusive and cost-efficient nature also makes it feasible to apply interaction-based detectors at scale, leading to a growing field of research regarding discovery with models [6].

Multimodal combinations of face- and interaction- features are few and far between. In one of the first studies, Kapoor et al. [17] used multimodal techniques with face- and posture- based features collected from naturalistic data to develop a detector of student interest. Facial features, such as detected head nods, shakes, and smiles, were combined with posture features gathered from a pressure-sensitive chair and interaction log features from the learning environment. The best unimodal detector (posture-based) classified high interest vs. low interest vs. taking a break, and had 82% accuracy (chance being 52%). The best multimodal detector combining all channels had 87% accuracy, demonstrating a

ICMI '15, November 09-13, 2015, Seattle, WA, USA
 ISBN 978-1-4503-3912-4/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2818346.2820739

modest improvement over the best unimodal detector. Further, Kapoor et al. [16] used facial features, a pressure-sensing chair, a pressure-sensitive mouse, a skin conductance sensor, and interaction log data to predict frustration. They were able to predict when a user would self-report frustration with 79% accuracy (chance being 58%).

D'Mello et al. [23] developed a multimodal affect detection system that discriminated between naturally occurring affective states in a learning environment. They used conversational cues, gross body language, and facial features as inputs for classification. Using fixed affect labels (every 20 seconds), mean Cohen's kappa improved from .220 (best individual modality model) to .288 in the best multimodal models. Using spontaneous affect labels (labeled at any point at the coder's discretion), mean kappa improved from .374 (unimodal) to .391 (multimodal). These results also demonstrated a small advantage for multimodal affect detection in terms of accuracy. However, facial features were manually annotated in this paper, an approach which is not feasible for large datasets or real-time interventions in a learning environment.

More recently, Grafsgaard et al. [12] developed multimodal affect detectors that utilized both facial features and student interaction features (as well as tutor-student dialog) in computer-mediated human-human tutoring sessions. They predicted engagement, frustration, and learning (all self-reported by the students) using linear regression with leave-one-student-out cross validation. A combination of facial features and interaction features predicted engagement (R^2 = .112) and frustration (R^2 = .134) more accurately than unimodal features (best R^2 = .048 and -.010 respectively for engagement and frustration). Adding tutor-student dialogue features improved results even further, R^2 = .282 and R^2 = .520 for engagement and frustration, respectively. This shows that a fusion of features can outperform individual modalities. However, affect detection was done at a course-grained level across an entire learning session, which limits its applicability to drive real-time interventions.

While the previous studies were conducted in somewhat controlled settings, Arroyo et al. [2] tracked emotions of high school and college mathematics students in computer-enabled classrooms. They used self-reports (for ground truth) and several modalities (interaction data from log-files, facial features, posture, skin conductance, and mouse movements). Their best models explained 67% of the variance (R^2) for self-reported confidence, 52% for frustration, 69% for excitement, and 29% for interest. Average R^2 was .1 higher for multimodal models with all modalities combined compared to the single modality (interaction-only) models reported. Although this research suggests that it might be possible to perform automated affect detection in classroom, this conclusion should be interpreted with a modicum of caution. This is because the model was not validated with a separate testing set (i.e. no cross validation was performed), and the size of the data set was very small (20-36 instances depending on model) due to missing data. These issues raise concerns of overfitting, but also serve to illustrate how much missing data occurs in classroom data collection.

1.2 Current Study: Novelty and Contributions

The current study expands on previous research by building a multimodal face- and interaction-based affect detector for a game-based learning environment called Physics Playground [31]. Previous research on these data has demonstrated the possibility of building both face-only and interaction-only detectors individually ([7,15]). Integrating these modalities also provides some unique challenges. Facial expressions are often brief (such as a delighted smile after succeeding in a task), whereas interaction events tend to unfold over a longer period of time (e.g., number of mouse clicks may be zero unless aggregated over a longer period of time).

Multimodal approaches have been shown to produce "modest" improvements in affect detection accuracy, according to a recent meta-analysis of 90 studies [25]. The meta-analysis showed that median improvement of multimodal approaches over the best single modality was 6.60%. The meta-analysis also revealed that multimodal classification of natural affect has shown an average of only one third of the benefit of multimodal classification than observed with acted affect. The present focus is also on classifying naturalistic affect displays rather than acted or posed affect. Thus, we do not expect large gains in classification accuracy from multimodal fusion.

However, there is more to consider than mere accuracy when evaluating multimodal affect detectors in the wild. Our multimodal affect detector is designed to operate in computer-enabled classrooms, a context that is rife with noisy and missing data, thereby providing additional challenges for multimodal affect detection. One key limitation of unimodal affect detection is the inability to operate when data are missing, a limitation that affects nearly every modality commonly used for affect detection. Heart rate, skin conductance, electroencephalogram, and other modalities measured via bodily contact sensors can be unavailable do to noise in the data or lack of contact caused by movement. Facial features can be unavailable when the person's face is occluded by hand-to-face gestures, when the person falls outside the camera's field of view, or when rapid motion causes detection errors. Environmental factors such as lighting can also cause face registration errors. Similarly, interaction features can be unavailable when a person is not frequently interacting with an interface, due to disengagement or non-interactive tasks such as reading or video watching.

This missing data problem was particularly prevalent in the present context where data was collected in the wild, specifically in a computer-enabled classroom environment with up to 30 students at a time playing an educational physics game. Students talked to one another, laughed, moved around, and gestured, as is expected in a game-based learning environment and in naturalistic contexts. Face data was frequently unavailable due to face registration failures caused by excessive movements or occlusions, while interaction data was not be present when the student was passively processing information rather than actively interacting. To address this issue, this paper considers both accuracy and availability while developing a unique multimodal combination of face + interaction data in the noisy real-world context of computer-enabled classrooms.

2. METHOD

2.1 Data Collection

The sample consisted of 137 8th and 9th grade students (57 male, 80 female) who were enrolled in a public school in a medium-sized city in the Southeastern U.S. They were tested in groups of about 20 students per class period for a total of four periods on four different days (55 minutes per period). Students in the 8th and 9th grades were selected because of the alignment of the learning content and the state standards (relating to Newtonian Physics) at those grade levels.

2.1.1 Learning Environment (Physics Playground)

Physics Playground (PP) [31] is a two-dimensional game that requires the player to apply principles of Newtonian Physics in an attempt to guide a green ball to a red balloon in many challenging configurations (key goal). The player can nudge the ball to the left and right (if the surface is flat) but the primary way to move the ball is by drawing/creating simple machines (which are called "agents of force and motion" in the game) on the screen that "come to life" once the object is drawn (example in Figure 1). Thus, the problems in PP require the player to draw/create four different agents (which are simple machine-like objects): inclined plane/ramps, pendulums, levers, and springboards. All solutions are drawn with colored lines using the mouse. Everything in the game obeys the basic laws of physics relating to gravity and Newton's three laws of motion.

The game includes seven playgrounds (each one containing 10–11 problems, for a total of 74 problems) that progressively get more difficult. The difficulty of a problem is based on a number of factors including the relative location of ball to balloon, obstacles, the number of agents required to solve the problem, and the novelty of the problem. The game is nonlinear in that students have complete choice in selecting playgrounds and levels. Progress in the game is represented by silver and gold trophies, which are displayed in the top left part of the screen. While a silver trophy is obtained for any solution to a problem, students earn a gold trophy if a solution is under a certain number of objects (the threshold varies by problem, but is typically < 3). PP maintains detailed log files that record the problem, student actions and when they occurred, system responses, trophies awarded, and so on.

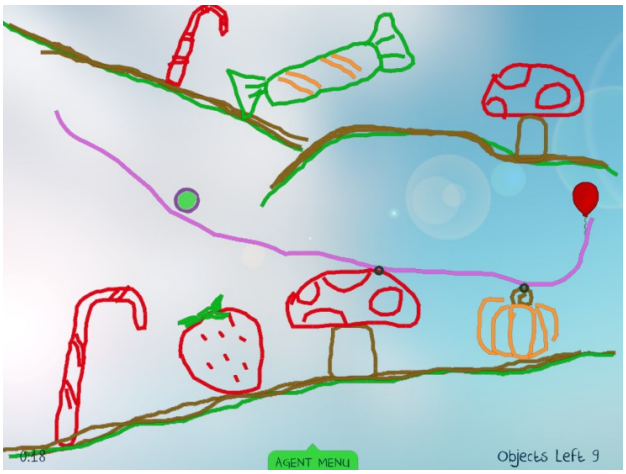

Figure 1. Ramp solution for a simple Physics Playground problem.

2.1.2 Procedure

The study took place in one of the school's computer-enabled classrooms, which was equipped with 30 desktop computers. Each computer was equipped with a monitor, mouse, keyboard, and headphones. Inexpensive webcams ($30) were affixed at the top of the monitor of each computer. At the beginning of each session, the data collection software displayed an interface that allowed students to position their faces in the center of the camera's view by adjusting the camera angle up or down. This process was guided by on-screen instructions and verbal instructions given by the experimenters, who were also available to answer any additional questions and to troubleshoot any problems.

We administered a qualitative physics pretest during the first day and a posttest at the end of the fourth day (both online). In this study we considered data from the second and third days (roughly two hours total) when students were only playing the game and not being tested.

Students' affective states were "live" (real-time) annotated during their interactions with PP using the Baker-Rodrigo Observation Method Protocol (BROMP) [27]. In BROMP, trained observers perform live affect annotations (real-time observations made while students played PP) by observing students one at a time using a round-robin technique (observing one student until visible affect is detected or 20 seconds have elapsed and moving on to the next student). Observers use side glances to make a holistic judgment of the students' affect based on facial expressions, speech, body posture, gestures, and student interaction with the computer program (e.g., whether a student is progressing or struggling). Observers record students in a pre-determined order to maintain a representative sampling of students' affect, rather than focusing on the most interesting (but not most prevalent) things occurring in the classroom. Every BROMP observer was trained and tested on the protocol and achieved sufficient agreement (kappa ≥ .6) with a certified BROMP observer before coding the data.

The coding process was implemented using the HART application for Android devices [27], which enforces the protocol while facilitating data collection. Observation-codes recorded in HART were synchronized with the videos recorded on the individual computers using Internet time servers. Observers averaged 3.2 observations per minute (approximately 19 seconds per observation). Observers moved on to the next student when they were confident in their rating so that timestamps coincide with affective manifestations.

It should be noted that there are many possible affect annotation schemes, each with their strengths and weaknesses, as recently reviewed in [30]. BROMP was selected for this study because it has been shown to achieve adequate reliability (among over 70 coders in over a dozen studies with a variety of learning environments [26]) in annotating affective states of a large number of students occurring in the "heat of the moment" and without interrupting or biasing students by asking them to self-report affect.

The affective states of interest were boredom, confusion, delight, engaged concentration, and frustration. This list of states was selected based on previous research [21] and from observing students during the first day of data collection (these data were not used in the current detectors). BROMP-coded observations of these states served as the ground truth labels for affect detection.

2.1.3 Instances of Affect Observed

Data from the two main days of game-play were collated and jointly analyzed in this study. We obtained 1,838 observations of affective states across the two days of data used in this study. Recording errors eliminated 44 face-based instances from consideration due to occasional computer crashes and performance issues. The most common affective state observed was engaged concentration (77.6%), followed by frustrated (13.5%), bored (4.3%), delighted (2.3%), and confused (2.3%).

2.2 Feature Engineering

We computed features from video recordings of students' faces as they played the game (facial features) and features obtained from logs of their interactions in the learning environment (interaction features). These features were then used to develop affect

detectors for each individual modality as well as for feature- and decision-level multimodal fusion models.

2.2.1 Facial Features

We used FACET for facial feature extraction. FACET is a commercialized version of the Computer Expression Recognition Toolbox (CERT) computer vision software [19]. Like CERT, FACET provides likelihood estimates for the presence of 19 Action Units (AUs) as well as head pose (orientation) and position information detected from video. Data from FACET was temporally aligned with affect observations in small windows. We tested five different window sizes (3, 6, 9, or 12 seconds). Features were created by aggregating (using maximum, median, standard deviation) values obtained from FACET (AUs, orientation, and position of the face) in a window of time leading up to each observation. For example, for a six-second window we created three features from the AU4 channel (brow lower). Those features were the maximum, median, and standard deviation of AU4 likelihood within the six seconds leading up to an affect observation. In all there were 78 facial features for each window.

We also used features computed from gross body movement estimated in the videos ([7]). Body movement was calculated by measuring the proportion of pixels in each video frame that differed from a continuously updated estimate of the background image generated from the four previous frames. Previous work has shown that features derived using this technique correlate with relevant affective states including boredom, confusion, and frustration [20]. We created three body movement features using the maximum, median, and standard deviation of the proportion of "motion" pixels within the window of time leading up to an observation, similar to the method used to create FACET features. In all, there were 3 body movement features, thereby yielding a total of 81 video-based features.

Facial features were only available for 65% of the instances due to face-registration errors. The unregisterable instances were treated as missing data and were addressed during multimodal fusion.

2.2.2 Interaction Features

Interaction features capture key aspects of students' actions as they played Physics Playground. The basic set of features considered the number of specific objects drawn as well as actions and events occurring during gameplay. Some examples include the number of springboard structures created in a level, the number of freeform objects drawn in a level, the time between start to end of a level, the number of gold trophies obtained in a level, and the number of stacking events (cheating behavior) in a level. In addition, time-based features focused on the amount of time elapsed between specific student actions, such as starting and pausing a level, as well as the time it took for a variety of events to occur within each level. In all, there were 113 interaction features.

Features were aggregated within a 20 second window leading up to each affect observation. A larger window size was required for interaction features than facial features because interaction events occur less frequently. Windows of all sizes for both facial and interaction features were synchronized to end at the same point in time. Differing window sizes thus did not affect the number of instances or alignment of instances between face and interaction modalities.

Occasional missing values were present at certain points in the dataset when a particular interaction was not logged. For example, a feature specifying the amount of time between the student beginning a level and his/her first restart of the level would contain a missing value if the student manages to complete a level without having to restart it. Zero imputations were performed where the missing values were replaced by the value 0. However, if all interaction features were missing for a particular instance, values were not replaced with 0. These instances occurred infrequently in situations where students did not interact with the game at all or when they were not interacting with a level, as was the case when selecting a new level or watching instructional videos on how to play the game. In all, interaction data was available for 94% of the instances compared to 65% for facial features.

2.2.3 Feature Selection

Tolerance analysis was used to eliminate features with high multicollinearity (variance inflation factor > 5) within each modality [1]. Feature selection was used to obtain a sparser and potentially more diagnostic set of features for classification. A common feature selection technique RELIEF-F [18] was run on the *training* data in order to rank features. A proportion of the highest ranked features were then used in the detectors (proportions of .1, .2, .3, .4, .5, and .75 were tested).

2.3 Supervised Learning

We built detectors using 14 different classifiers including support vector machines, C4.5 trees, Bayesian classifiers, and others in the Waikato Environment for Knowledge Analysis (WEKA), a machine learning tool [13]. A large number of classifiers were considered since we are unaware of the best approach for this type of data.

A two-class approach was used for each affective state, where that affective state was discriminated from all others. For example, engaged concentration was discriminated from all frustrated, bored, delighted, and confused instances combined (referred to as "all other"). Building individual detectors for each state allows the parameters (e.g., window size, features used) to be optimized for that particular affective state

The affective distributions lead to large class imbalances (e.g. .04 vs. .96 class priors in the bored vs. all other classification). Two different sampling techniques were used (on training data only) to compensate for class imbalance. These included downsampling (removal of random instances from the majority class) and synthetic oversampling (with SMOTE; [9]) to create equal class sizes. SMOTE creates synthetic training data by interpolating feature values between an instance and randomly chosen nearest neighbors. The distributions in the testing data were not changed, to preserve the validity of the results.

Detectors were cross-validated at the student level using leave-one-student-out cross validation. A detector was trained on data from all but one holdout student, then tested on data from the holdout student. Feature selection was performed using nested cross-validation on training data only (see below). Ten iterations of feature selection were run on the training data, using data from a randomly chosen 67% of students within the training set for each iteration. The entire cross-validation process was repeated for every student to produce student-independent predictions for every instance. This helps detectors generalize to new students since training and testing data sets are student-independent.

2.4 Multimodal Fusion

Basic feature-level fusion and decision-level fusion techniques were used in the present paper. Feature level fusion involved combining features from both modalities into a single dataset and then building a detector. The disadvantage of this simplistic approach is that the data must be present in every modality for

training and testing. Decision-level fusion, on the other hand, involved creating a detector separately with each modality, thus exploiting all the training data available for that modality. Individual detector outputs were then combined by training an additional classifier on the outputs of the individual classifiers to form a final classification. Both approaches used leave-one-student-out cross validation.

3. RESULTS

The results are organized with respect to three aspects of multimodal classification. First, how does multimodal accuracy compare to unimodal accuracy; second, does a multimodal approach improve detector availability; and third, what are the benefits of a more sophisticated multimodal approach compared to a simplistic one.

The Area Under the ROC Curve (AUC) was used to measure classification accuracy. AUC measures the tradeoff between the true positive and false positive rates of a detector. AUC ranges from 0 (completely incorrect classification) to 1 (perfect classification), while an AUC of .5 represents chance-level classification. AUC is a recommended accuracy metric when dealing with imbalanced data [14], which is the case in the current study (e.g., boredom comprised only 4.23% of data).

3.1 Fusion with Nonmissing Data only

The applicability of the feature-level fusion detector was limited because a significant number of instances were missing either one modality or the other. To make a fair comparison with identical instances, training and testing datasets contained only instances when both face and interaction data were valid. Table 1 shows a comparison of classification accuracy of the face-only, interaction-only, feature-level fusion, and decision-level fusion detectors on identical instances. The number of instances (N) varies between affective states because the window sizes varied. On average, 1,124 instances out of 1,838 total instances were used in these analyses, so a considerable amount of data were being unused.

The results indicated that the multimodal detectors yielded similar accuracies compared to the best unimodal detectors for boredom, delight, engagement, and frustration. Detection accuracy for confusion was higher for the feature-level fusion detector compared to the unimodal detectors.

Table 1. AUC comparison of modalities and fusion methods using identical instances.

	Face	Interaction	Feature-Level Fusion	Decision-Level Fusion	N
Boredom	**.594**	.546	.520	**.598**	1229
Confusion	.573	**.600**	**.663**	.596	1229
Delight	**.869**	.662	.848	**.865**	943
Engagement	**.670**	.505	.669	**.673**	1153
Frustration	**.627**	.556	**.634**	.623	1064
Mean	**.667**	.574	.667	**.671**	

Note: Bold indicates best unimodal and multimodal models.

3.2 Fusion with All Available Data

The comparison in Section 3.1 shows that the feature-level and decision-level fusion approaches are at least comparable to the best individual modalities, but did not offer a large benefit in terms of accuracy with the minor exception of confusion. This finding was not unexpected, given the null to modest multimodal improvements typically obtained for naturalistic affective expressions [25]. In the Introduction, we argued that there may be additional benefits to multimodal affect detection beyond accuracy. Here, we consider the potential of multimodal affect detection as a means to improve availability (i.e., handling of missing data).

The analyses preceded as follows. Multimodal detectors were built to operate on as many instances as possible by predicting affect for every instance as long as it contained either face or interaction data (or both). Decision-level fusion was used because it affords training individual base detectors for each modality. Each base detector can then provide a prediction (decision) for an instance if data are available for that modality. The decisions from each modality's detector can then be combined. This procedure maximizes the use of training data because the detector for each modality can be trained using all of the data from that modality. It also maximizes the availability of the detector by making a prediction whenever any of the individual modalities has available data.

There are many possible methods of combining the decisions of individual modalities to provide an overall prediction for an instance. We built three types of decision-level fusion detectors. The first type used the face-only detector whenever possible and the interaction-only detector only when face data were unavailable (face-biased). The second type was the complement, using interaction-only detection whenever possible and using the face-only detector as a backup (interaction-biased). The third type used a classifier to make decisions with the output of each of the individual modalities as input features, thereby essentially learning how to weigh each modality (learned-bias). When a channel was not available, its prediction was taken to be 0.5 (even odds for a two-way classification problem). We also experimented with 0 and 1 replacement, but found 0.5 to produce better results, which is expected since it is unlikely to bias the classifier's decision significantly. Table 2 contains an overview of the modalities used for each of the fusion methods. The AUC accuracy of these three detection-schemes after averaging across the affective states is shown in Figure 2, where we note a small advantage of the learned-bias detector compared to the face-biased and interaction-biased detectors.

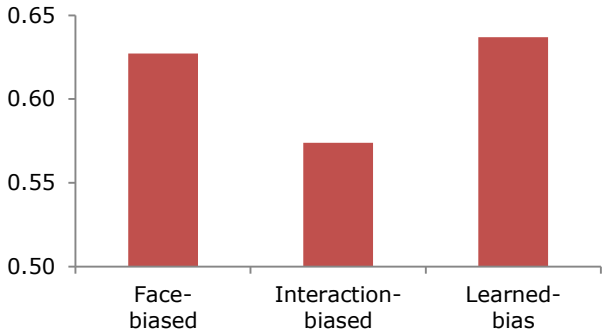

Figure 2. Comparison of decision-level fusion methods.

3.3 Accuracy vs. Availability

Section 3.1 indicated that the face-only model yielded the best accuracy when only considering cases with nonmissing instances. This lead to large data loss (high accuracy and low availability). In Section 3.2, using the interaction-biased weighting scheme yielded the best results while using all available instances. Here, we attempt to reconcile these two results.

Table 2. Modality used in multimodal fusion methods.

Available		Modality Used For Decision			
Face	**Interaction**	**Feature-level**	**Face-biased**	**Interaction-biased**	**Learned-bias**
Yes	Yes	Both	Face	Interaction	Both
Yes	No	None	Face	Face	Face
No	Yes	None	Interaction	Interaction	Interaction

Figure 3 displays and compares the available percentage of total instances and the accuracy of individual modalities versus the learned-bias multimodal detectors. We note that the proportion of total instances available for classification increased notably with the use of decision-level multimodal fusion. Thus, availability was dramatically improved over the face-only detector with little reduction in classification accuracy. Availability also marginally improved over the interaction-only detector, but accuracy was dramatically improved in this case.

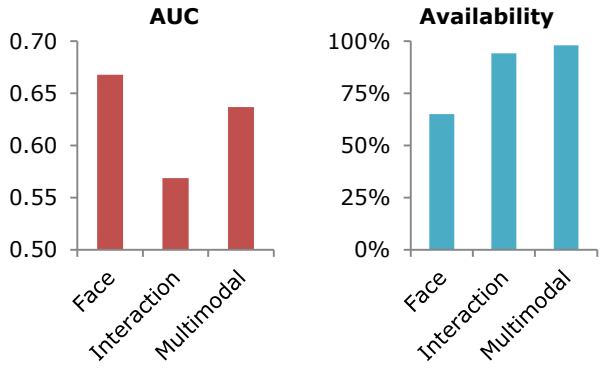

Figure 3. Comparison of accuracy versus availability for unimodal and multimodal detection.

3.4 Analysis of Learned-bias Multimodal Detector

Figure 3 showed that using multimodal fusion increased the percentage of situations in which the detectors could be applied to nearly 100% of the cases. However, the average multimodal detector accuracy was slightly lower than the face-only detectors. Thus, further analysis of the accuracy of this combined detector is needed to more clearly understand why accuracy was diminished. We examined the accuracy of the learned-bias detector in different situations (face data available vs. missing) to determine why this was the case.

There was an average of 595 instances where interaction data were available but face data were missing. In this situation the learned-bias detectors had an average accuracy of AUC = .583. In contrast, the average accuracy when the face data were present was AUC = .694. Thus, the detectors performed better when the face was present, which could be expected since the face generally outperformed the interaction-only detectors. It might be reasoned from this that most of the learned-bias detectors' predictive power stems from the face modality. However, the learned-bias detectors were more slightly accurate than simply using the face-only detector whenever possible (see Figure 2). It appears that the learned-bias detector made productive use of the information contained in the less accurate interaction modality even when the more accurate face modality was available. Thus, there was some improvement to be gained through a weighted fusion rather than simply using the face-only detector when available and reverting to the interaction-only detector otherwise. Table 3 shows the details of how individual modalities were weighted in the learned-bias detectors for the best classifiers (from those listed in section 2.3).

Table 3. Details of learned-bias detectors.

Affect	**AUC**	**Classifier**	**Face Weight**	**Interaction Weight**
Boredom	.590	Logistic Regression	0.415	0.157
Confusion	.585	Logistic Regression	1.20	20.7
Delight	.775	Logistic Regression	0.043	0.137
Engagement	.621	Updateable Naïve Bayes	NA	NA
Frustration	.614	Logistic Regression	0.314	0.064

Note: Weights denote odds ratios in logistic regression. No weights were available from the Updateable Naïve Bayes classifier.

As can be seen in Table 3, the confusion detectors are heavily biased toward using the interaction modality. That finding is consistent with Table 1, which showed that confusion was the one modality where interaction-only detectors out-performed face-only detectors, despite using instances where the face data was available. Surprisingly, the delight classifier also weighted interaction more highly than face, though not to the same extent. One possible reason for that finding is that the face data were so frequently missing for delight (48.7% missing) due to the short, 3-second window size. Thus the learned-bias detector learned to rely more heavily on the interaction detector instead.

4. GENERAL DISCUSSION

We explored multimodal affect detection in a classroom environment where students interacted with a game-based learning environment called Physics Playground [31]. Specifically, we focus on the use of multimodal approaches to affect detection as a means of improving the accuracy and availability of detectors in this noisy environment where missing data are prevalent. We employed face-only and interaction-only affect detection methods as two modalities with a tradeoff between accuracy and availability (applicability to a large number of situations and instances). In this section we review our main findings, point out limitations and opportunities for future work, and discuss implications for affect-sensitive learning environments.

4.1 Main Findings

We investigated multimodal classification of affect in several ways. First, we compared multimodal detectors with unimodal detectors. We did not expect large improvements based on previous research [25]. Indeed we found no major improvement in average accuracy with multimodal classification, but established

that accuracy was at least not diminished. Face-only affect detection was more accurate but there were a larger number of missing instances, while interaction-only detection was applicable to a larger number of instances but was less accurate. Thus, our primary goal was to improve the availability of detectors without significantly reducing accuracy. We used decision-level fusion to develop detectors that worked for 98% of the instances in our data—a notable improvement over the 65% availability of the face modality. Accuracy of these multimodal detectors was close to that of the face-only detectors. Thus, we appear to have struck an appropriate balance between accuracy and availability in this context.

We further investigated the accuracy of the improved detectors in two ways. First, we examined accuracy in situations where the face (most frequently missing modality) was missing versus those where it was not. We found that accuracy was notably better when face data were available, but that the accuracy of learned-bias multimodal detectors improved through the use of interaction data even when face data were available. This finding is significant because it shows there was benefit to the interaction modality despite its lower accuracy. A deeper analysis of the logistic regression weights of individual modalities in the final decision-level fusion detectors revealed that the interaction modality was indeed influencing the final classification decision (Table 3).

4.2 Limitations and Future Work

Some limitations of this study should be noted. First, additional modalities such as acoustic features might have proven useful, but were not collected. Future work should address this by considering other common affect detection modalities and determining which are suitable for use given the constraints of having multiple students in a computer-enabled classroom. Second, though the students in this study varied widely across some demographic variables, they were all approximately the same age and in the same location. A further study testing detectors on data with more variability in age and geographic distribution would be useful for determining the level to which results might generalize to different group of students. Third, the observation method used (BROMP) requires observers to be in the room, which could influence students displays of affect akin to a Hawthorne effect [10]. These limitations could be addressed by an additional study comparing the incidence of affective states experienced by students with different demographics using a variety of different sensors and affect annotation methodologies.

4.3 Implications for Affect-Sensitive Interfaces

Our results were particularly relevant to affect detection efforts in intelligent learning interfaces for computer-enabled classrooms. For example, the detectors we developed will be used to improve intelligent instructional strategies towards developing an affect-sensitive version of Physics Playground. Separate strategies will be used for each affective state and off-task behavior. For example, when the detectors determine that a student is engaged or delighted, Physics Playground may not intervene at all. Confusion and frustration offer intervention opportunities in the form of hints or revisiting introductory material related to the concepts in the current level. If the student has recently been frustrated and unable to complete levels, an easier level might be suggested. Conversely, a more difficult level might be appropriate if the student has not been challenged by recently completed levels. Boredom might be addressed by suggesting that the student attempt a new level or by calibrating difficulty.

4.4 Concluding Remarks

Affect interventions require real-time affect detection, and this work has extended the amount of cases where an affect detector can operate while maintaining accuracy. In particular, face-only affect detection systems seem likely to suffer from considerable missing data. Interaction-only detectors suffered less from missing data, but accuracy was significantly worse than for face-only detectors in this context. Our multimodal fusion solution offered 98% availability in the wild with better accuracy than the interaction-only method and equivalent to face-only detection. The next step is to embed our multimodal affect detection into intelligent, adaptive educational interfaces for use in computer-enabled classrooms.

5. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) (DRL 1235958) and the Bill & Melinda Gates Foundation. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF or the Bill & Melinda Gates Foundation.

6. REFERENCES

1. Paul D. Allison. 1999. *Multiple regression: A primer*. Pine Forge Press.
2. Ivon Arroyo, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. 2009. Emotion sensors go to school. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, IOS Press, 17–24.
3. Ryan Baker, Sujith M. Gowda, Michael Wixon, et al. 2012. Towards sensor-free affect detection in cognitive tutor algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126–133.
4. Ryan Baker and Jaclyn Ocumpaugh. 2015. Interaction-based affect detection in educational software. In *The Oxford Handbook of Affective Computing*, Rafael Calvo, Sidney D'Mello, J. Gratch and A. Kappas (eds.). New York: Oxford University Press, 233–245.
5. Ryan Baker, Jaclyn Ocumpaugh, Sujith M. Gowda, Amy M. Kamarainen, and Shari J. Metcalf. 2014. Extending log-based affect detection to a multi-user virtual environment for science. *22nd Conference on User Modeling, Adaptation and Personalization (UMAP 2014)*, Springer, 290–300.
6. Ryan Baker and Kalina Yacef. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining* 1, 1: 3–16.
7. Nigel Bosch, Sidney D'Mello, Ryan Baker, et al. 2015. Automatic detection of learning-centered affective states in the wild. *Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015)*, New York, NY: ACM, 379–388.
8. R.A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1: 18–37.
9. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2011. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16: 321–357.
10. Desmond L. Cook. 1962. The Hawthorne effect in educational research. *Phi Delta Kappan*: 116–122.
11. Joseph F. Grafsgaard, Joseph B. Wiggins, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2013.

Automatically recognizing facial expression: Predicting engagement and frustration. *Proceedings of the 6th International Conference on Educational Data Mining*.

12. Joseph F. Grafsgaard, Joseph B. Wiggins, Alexandria Katarina Vail, Kristy Elizabeth Boyer, Eric N. Wiebe, and James C. Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM, 42–49.
13. G. Holmes, A. Donkin, and I.H. Witten. 1994. WEKA: a machine learning workbench. *Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems*, 357–361.
14. L.A Jeni, J.F. Cohn, and F. de la Torre. 2013. Facing imbalanced data–Recommendations for the use of performance metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 245–251.
15. Shiming Kai, Luc Paquette, Ryan Baker, et al. 2015. Comparison of face-based and interaction-based affect detectors in physics playground. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*, International Educational Data Mining Society, 77–84.
16. Ashish Kapoor, Winslow Burleson, and Rosalind W. Picard. 2007. Automatic prediction of frustration. *International Journal of Human-Computer Studies* 65, 8: 724–736.
17. Ashish Kapoor and Rosalind W. Picard. 2005. Multimodal affect recognition in learning environments. *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ACM, 677–682.
18. Igor Kononenko. 1994. Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning: ECML-94*, Francesco Bergadano and Luc De Raedt (eds.). Springer, Berlin Heidelberg, 171–182.
19. Gwen Littlewort, J. Whitehill, Tingfan Wu, et al. 2011. The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011)*, 298–305.
20. Sidney D'Mello. 2011. Dynamical emotions: Bodily dynamics of affect during problem solving. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
21. Sidney D'Mello. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology* 105, 4: 1082–1099.
22. Sidney D'Mello, Nathan Blanchard, Ryan Baker, Jaclyn Ocumpaugh, and Keith Brawner. 2014. I feel your pain: A selective review of affect-sensitive instructional strategies. In *Design Recommendations for Intelligent Tutoring Systems - Volume 2: Instructional Management*, Robert Sottilare, Art Graesser, Xiangen Hu and Benjamin Goldberg (eds.). 35–48.
23. Sidney D'Mello and Art Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, 2: 147–187.
24. Sidney D'Mello, Tanner Jackson, Scotty Craig, et al. 2008. AutoTutor detects and responds to learners affective and cognitive states. *Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems*.
25. Sidney D'Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys* 47, 3: 43:1–43:36.
26. Jaclyn Ocumpaugh, Ryan Baker, Amy Kamarainen, and Shari Metcalf. 2014. Modifying field observation methods on the fly: Creative metanarrative and disgust in an environmental MUVE. *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP 2014)*, 49–54.
27. Jaclyn Ocumpaugh, Ryan Baker, and Ma Mercedes T. Rodrigo. 2012. *Baker-Rodrigo observation method protocol (BROMP) 1.0. Training manual version 1.0*. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.
28. Luc Paquette, Ryan S.J.d. Baker, Michael Sao Pedro, et al. 2014. Sensor-Free Affect Detection for a Simulation-Based Science Inquiry Learning Environment. *Proceedings of the 12th International Conference on Intelligent Tutoring Systems, ITS 2014*, 1–10.
29. Zachary a. Pardos, Ryan S.J.d. Baker, Maria O. C. Z. San Pedro, Sujith M. Gowda, and Supreeth M. Gowda. 2014. Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics* 1, 1: 107–128.
30. Kaśka Porayska-Pomsta, Manolis Mavrikis, Sidney D'Mello, Cristina Conati, and Ryan Baker. 2013. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education* 22, 3: 107–140.
31. Valerie Shute and Matthew Ventura. 2013. *Measuring and supporting learning in games: Stealth assessment*. The MIT Press, Cambridge, MA.
32. J. Whitehill, Z. Serpell, Yi-Ching Lin, A Foster, and J.R. Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1: 86–98.
33. Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 1: 39–58.