Unimodal and Multimodal Human Perception of Naturalistic Non-Basic Affective States during Human-Computer Interactions

Sidney K. D’Mello, Nia Dowell, and Art Graesser

Abstract—The present study investigated unimodal and multimodal emotion perception by humans, with an eye for applying the findings towards automated affect detection. The focus was on assessing the reliability by which untrained human observers could detect naturalistic expressions of non-basic affective states (boredom, engagement/flow, confusion, frustration, and neutral) from previously recorded videos of learners interacting with a computer tutor. The experiment manipulated three modalities to produce seven conditions: face, speech, context, face+speech, face+context, speech+context, face+speech+context. Agreement between two observers (OO) and between an observer and a learner (LO) were computed and analyzed with mixed-effects logistic regression models. The results indicated that agreement was generally low (kappas ranged from .030 to .183), but, with one exception, was greater than chance. Comparisons of overall agreement (across affective states) between the unimodal and multimodal conditions supported redundancy effects between modalities, but there were superadditive, additive, redundant, and inhibitory effects when affective states were individually considered. There was both convergence and divergence of patterns in the OO and LO datasets; however, LO models yielded lower agreement but higher multimodal effects compared to OO models. Implications of the findings for automated affect detection are discussed.

Index Terms—multimodal affect detection, emotion perception, naturalistic expressions, non-basic emotions, superadditivity

1. INTRODUCTION

AFFECT detection has been, and remains, one of the most significant challenges in Affective Computing (AC) research. It is also an area that has garnered significant attention by AC researchers. Considerable progress has been made in applying machine learning techniques to detect affective states from facial expressions, acoustic-prosodic cues, body movements, gesture, contextual cues, text and discourse, and both peripheral and central physiology [1, 2, 3, 4 for reviews]. However, there are still a number of challenges that need to be overcome before fully-automated affect detectors with modest accuracy can be deployed in real-world contexts. Of course, this is to be expected, given the difficulty of the problem and the relative infancy of the field. As a point of comparison, automatic speech recognition, whose difficulty is sometimes equated with affect detection [5], has been an active research area for more than 75 years [6]. Yet, it is only recently that functional speech recognition applications are coming online.

There have also been some impressive advances in affect detection over the last decade. While earlier approaches concentrated on single modalities [3], several researchers have recently been focusing on developing affect detectors that combine multiple modalities such as face and speech [7-11], speech and text [12-14], different physiological signals [15-18], and sometimes combinations of face, speech, posture, and other information [19-23].

Given the effort and expense associated with the increased complexity of multimodal systems, there is the important question of whether the combination of multiple modalities results in substantial rather than incremental improvements over the individual modalities. There is also the possibility that multiple modalities will yield lower classification rates than the individual modalities (e.g., if the individual modalities provide conflicting information that cannot be effectively resolved). These are the questions that motivated the present paper, which begins with a description of some criteria to categorize the performance of multimodal affect classifiers.

1.1. Quantifying Multimodal Effects

We propose superadditivity, additivity, redundancy, and inhibition as four possible outcomes of multimodal affect detection [23]. We conceptualize superadditivity as occurring when the combination of modalities results in improvements in affect detection accuracies above what could be expected from an additive combination alone. Simply put, the whole is greater than the sum of the parts. Additivity occurs when the multimodal combination yields accuracies that could be expected from an additive combination of the individual modalities. Redundancy occurs when the combination of modalities produces null or negligible improvements in accuracies, while inhibition
occurs when a combination of modalities yields lower accuracies than the individual modalities.

One simple metric to quantify the performance of a multimodal classifier (multimodal effect size) is to consider improvements in accuracy over the best unimodal classifier. If \( a_1 \) and \( a_2 \) are accuracies associated with two unimodal classifiers, and \( a_{1+2} \) is the multimodal accuracy, then the multimodal effect size is computed as:

\[
MM\text{ Effect Size } = 100 \times \frac{a_{1+2} - \max(a_1, a_2)}{\max(a_1, a_2)}
\]

This metric is widely used by neuroscientists studying multisensory integration with respect to visual, audio, and tactile senses in humans and animals [24]. Though attractive in terms of its simplicity and intuitive appeal, it has two main limitations. First, the metric only considers the size of the multimodal effect, but does not consider the statistical significance of the effects. This can be misleading in cases when there is substantial variance in accuracies, because the MM effect can be large but the multimodal classifier might not be significantly different from the best unimodal classifier. Second, this metric does not afford a clear categorization of the size of an effect. For example, for a given classification task, is a 3% improvement small or is a 10% improvement small?

To address these limitations, we propose an alternate metric to categorize the performance of a multimodal classifier as being superadditive, additive, redundant, or inhibitory. The metric has two components: (1) significance testing and (2) comparison of observed multimodal accuracy with expected accuracy (\( add_{\exp} \)). The idea is that a multimodal classifier must be both significantly and substantially greater than the unimodal classifiers for it to yield superadditive effects. If it is significantly better, but not substantially better, then it has yielded additive effects. If it is statistically equivalent to any of the unimodal classifiers, then the multimodal fusion has resulted in redundancy. Inhibition occurs when the multimodal classifier results in a significantly lower classification accuracy than all of the unimodal classifiers. More specifically:

- \( \text{if } a_{1+2} \text{ significantly } > a_1 \text{ and } a_2: \)
- \( \text{if } a_{1+2} > add_{\exp} \quad \text{superadditivity} \)
- \( \text{else } \quad \text{additivity} \)
- \( \text{else if } a_{1+2} \text{ statistically } = \text{ to } a_1 \text{ or } a_2 \quad \text{redundancy} \)
- \( \text{else if } a_{1+2} \text{ statistically } < \text{ to } a_1 \text{ and } a_2 \quad \text{inhibition} \)

It is necessary to specify the expected additive threshold. One simple metric is to base this threshold on an expected additive combination of modalities (see Eq. 2 and 3 for cases with two and three modalities, respectively). This formulation is based on the formula for the probability of the union (additive combination) between two independent events: \( \Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A) \times \Pr(B) \).

\[
add_{\exp} = a_1 + a_2 - a_1 \times a_2
\]

1.2. Motivation for Present Study

How well do multimodal affect classifiers perform compared to their unimodal counterparts? D’Mello and Kory [25] recently addressed this question by performing a meta-analysis on 30 published studies that reported both multimodal and unimodal affect detection accuracies. These 30 studies had considerable variation in terms of \textit{data}, \textit{affect, modality}, and \textit{method} – details of which are beyond the scope of this paper. Most studies did not perform or report significance tests that compared multimodal accuracy scores to unimodal accuracy scores. This made it difficult to discriminate between superadditive, additive, redundant, or inhibitive effects, at least based on the criteria proposed above. Hence, multimodal improvements were quantified on the basis of MM1 effects as specified in Eq. 1.

The results indicated that multimodal accuracies were consistently (26 out of 30 studies) better than unimodal accuracies, and on average, yielded an 8.12% improvement over the best unimodal classifiers. Interestingly, the mean MM effects size for classifiers trained on acted expressions was 12.1%, which was three times greater than mean MM effects for classifiers trained on more naturalistic expressions (4.39%). Another important finding was that performance of the best unimodal classifier explained 80.6% (cross-validated) of the variance in multimodal accuracy.

The aforementioned meta-analysis focused on quantifying the improvements of \textit{automated} multimodal affect detectors over unimodal systems [25]. Taking a somewhat different approach, the present study aims to quantify multimodal improvements when \textit{humans} performed the emotion classifications. We were interested in evaluating the accuracy by which untrained humans could detect \textit{naturalistic} expressions of \textit{non-basic affective states} from unimodal and \textit{multimodal} signals involving the face, speech, and context.

The present focus on human affect detection is motivated by the assumption that human-generated affect detection accuracy scores can be used as \textit{one} (but not the only) meaningful standard to compare machine-generated detection scores. We of course do not want to rule out the fact that the machines might outperform the humans or that humans and machines might detect affect in drastically different ways. For example, on one hand, machines have advanced sensing capabilities (e.g., thermal images, EEG, etc) that are beyond the radar of humans. On the other hand, humans have fine-tuned their emotion decoding capabilities over thousands of years and their use of situational context to assist with emotion perception far exceeds what is capable by a machine. To be clear, we are not arguing in favor of either humans or machines being better affect detectors. Our claim is much more modest and simply asserts that analyzing performance of humans can provide important insights for designing and evaluating multimodal affect detectors. In line with this, we begin with a brief review of previous
work on emotion perception by humans.

1.3. Emotion Perception by Humans

The means and accuracy by which humans express and perceive emotion has been an active area of research for several decades. Many theories of how humans perceive emotions have been proposed and have fallen in and out of fashion (see [26-28] for a review). Two dominant contemporary theories include simulation- and inference-based approaches. Simulation models (or shared substrates models) posit that the emotion recognition process consists of two non-mutually exclusive processes: emotion contagion and simulation [29]. Emotion contagion involves experiencing an emotion while perceiving a similar emotion in others [30]. Simulation is a process by which the bodily and motor processes associated with experiencing emotion X are activated during the recognition of X [27].

Taking a somewhat different approach, Scherer [28] has recently proposed the dynamic Tripartite Emotional Expression and Perception Model (TEEP), which is an inference-based model. According to TEEP, externalized emotional expressions, which may not be intentional, unfold dynamically as the emotional event is sequentially appraised by the expresser. These expressions are perceived as emotional symptoms by the observer, who interprets these signals via inference and attribution mechanisms that are strongly constrained by socio-cultural context of the interaction.

There is a massive literature on emotion perception by humans and it is beyond the scope of this article to delve deeply into this literature. Hence, we focus on studies that are particularly germane to the present paper. These include: (1) accuracy and reliability of emotion perception in both human-human and human-computer interactions and (2) unimodal vs. multimodal emotion perception in human-human communication (there are no known studies for human-computer interactions).

With respect to the first point of focus, recent reviews of the emotion perception literature have indicated that accuracy (rater agreement compared to some objective ground truth) and reliability (interrater agreement) scores range from below chance to large effects depending on the task, the methodology, and other factors [26, 31-33]. Despite this considerable range in accuracies and reliabilities, one consistent finding is that human performance is low when emotions are not intentionally elicited and when there are significant cross-cultural differences.

While many of the aforementioned studies focus on human-human communication, the literature is more silent on emotion perception during human-computer interactions. One study that investigated this issue [34] found agreement scores ranging from 0.08 (Cohen’s kappa) to 0.36 depending on who provided the emotion judgments (emoters themselves, untrained observers, or trained judges). A more recent study, consisting of three trained raters labeling short 3.4 second clips of individuals engaged in computer-based learning and card-sorting tasks, yielded an overall Fleiss’s kappa of 0.35 [35].

With respect to the second point of focus, most (but not all) of the studies in the literature have focused on unimodal perception of emotion that concentrate on one channel, mainly the face and sometimes speech. The studies that have investigated multimodal combinations of the face and the speech have yielded interesting modality interactions [36, 37]. For example, using a McGurk-like paradigm, where the degree of concordance between visual and vocal expression of emotion was manipulated, de Gelder and Vroomen [36] show that individuals typically combine information from both modalities. Interestingly, explicit instructions to focus on facial information did not weaken this effect. de Gelder and colleagues [36, 37] provide a useful review of some of the emerging research on emotion perception when the face is combined with body postures, speech, and information on scenes. One general conclusion is that recognition of a facial expression is more reliable and faster when there is congruence between the facial expression and the other informational channels (e.g., a disgusting face embedded within a scene of a garbage dump). However, in contrast, Cowie and colleagues [38] conducted an interesting study that compared human perception of emotions from audio, video, and audiovisual naturalistic stimuli (TV shows and other media clips). They report kappa values of 0.54 for audio, 0.43 for video, and 0.37 for audiovisual, which is contrary to what was expected.

1.4. Overview of Present Study

We conducted an experiment in which 63 participants (called judges) judged the affective states experienced by 21 individuals (called learners) who were engaged in a learning task with a computer. The videotaped sessions were collected in a previous study involving tutorial sessions with a dialogue-based intelligent tutoring system (ITS) called AutoTutor [39]. We focused on learning sessions because previous research has indicated that learning with technology is an excellent context to investigate non-basic affective states, such as boredom, engagement/flow, confusion, engagement/flow, and frustration [40] that are the focus of this paper.

Judges in the present study judged the learner’s affective states (boredom, engagement/flow, confusion, frustration) and neutral in a seven-condition experiment consisting of unimodal, bimodal, and trimodal conditions. The unimodal conditions included: (1) the video of a learner’s face (F), (2) audio of a learner’s speech (S), and (3) the text of the dialogue history between learner and tutor representing the context (C). There were also bimodal and trimodal conditions of: (4) face plus speech (FS), (5) face plus context (FC), (6) speech plus context (SC), and (7) face, speech, and context (FSC).

One concern with this setup is that the content of the learners’ speech would be available to the judges in all of the conditions, except the F condition. In order to alleviate this potential confound, human-transcribed text of the learners’ speech was presented to the judges in all conditions. Although this was needed to facilitate meaningful between-condition differences, a downside is that, strictly speaking, we no longer have unimodal F and S conditions. That being said, it should be noted that previous research on a dialogue corpus with the same tutor has
indicated that the content of learners’ utterances primarily consists of short (3-5 word) domain-related word fragments (e.g., “Operating System writes to RAM”) that were not very diagnostic of learner affect without the overarching tutor dialogue [41]. There was also only one instance of an explicit emotional term (e.g., learners saying “I’m confused”) in over 1,500 student responses. Hence, in our view, providing transcriptions of the learners’ spoken responses does not bias the F and S conditions, so these are still referred to as unimodal conditions.

Although there is considerable literature on human perception of emotions (as briefly surveyed above), the present study is unique in that it focuses on comparing human-generated multimodal and unimodal accuracies associated with detecting context-sensitive naturalistic expressions of non-basic emotions during human-computer interactions. First, the present focus on multimodality is somewhat novel in itself because the overwhelming majority (but not all – e.g., [36, 37]) of studies have focused on unimodal signals, particularly the face and speech. Second, many (but not all) of the previous studies have focused on the perception of context-free posed expressions of the basic emotions. The present focus is on context-sensitive naturalistic expressions of non-basic affective states. Third, while most of the previous studies have focused on emotions in human-human interactions, the present study focuses on judging emotions that arise during human-computer interactions. Each of these unique features is not novel in itself; however, to the best of our knowledge, their combination in a single study is novel. Furthermore, the present study bears a close resemblance to the affect detection problem (by computers) because it involves unimodal and multimodal detection of context-sensitive naturalistic expressions of affective states that arise when humans interact with computers. Therefore, we have some confidence that any insights gleaned can be informative to multimodal affect detection.

2. Stimuli Used in Present Experiment

The stimuli for the present experiment consisted of a subset of 21 videos from 30 learners collected in a previous study [42]. The study had two phases: a learning session and an affect judgment phase. It is important to point out that this earlier tutoring study was completed before the present study was conceptualized, so we are constrained by the data available from this previous study.

2.1. Learning Session

Thirty learners (63% female, 33% Caucasian, 60% African-American, 7% “Other”) completed a 35-minute tutorial session on topics in Computer Literacy (hardware, operating systems, the Internet) with AutoTutor. AutoTutor is an intelligent tutoring system that helps students learn difficult technical material such as Newtonian physics, critical thinking, and computer literacy by holding a conversation in natural language [39]. AutoTutor’s dialogues are organized around difficult questions, such as why, how, what-if, what if not, how is X similar to Y, that require answers involving inferences, explanations, and deep reasoning. Although each question requires 3-7 sentence-like ideas in a correct answer, learners rarely give the complete answer in a single conversational turn. Therefore, the tutor scaffolds the construction of an answer by an adaptive dialogue with pumps for information, hints, prompts, assertions, summaries, and feedback. An excerpt of the dialogue between AutoTutor and the learner is presented in the Methods section.

AutoTutor delivers its dialogue moves via an animated conversational agent that speaks the content of the tutor’s turns and makes some rudimentary gestures. In this study, learners spoke their responses to the tutor via a microphone and a commercially available speech recognition system was used for speech-to-text translation. A video of the learner’s face was recorded during the interaction. The video also included all audio generated by the tutor (synthesized speech) and the learner.

2.2. Retrospective Affect Judgment

Learners self-reported judgments of their affective states after the tutorial session, following methods similar to the cued elicitation procedure of Rosenberg and Ekman [43]. Learning activities during the tutoring session were not interrupted. The judgments of a learner’s tutoring session proceeded by playing a video of the face captured during the session. The video also included the audio of the spoken dialogue between the learner and the tutor. The content of the spoken interchange between learner and tutor constituted the context of the tutorial session when retrospective judgments were collected.

The learners were instructed to make judgments on what affective states were present at three different points during the tutorial dialogue: (a) immediately before the learner started expressing his/her spoken turn (33 points), (b) after the learner’s spoken turn was completed (33 points), and (c) other randomly selected points in the dialogue (33 points). These points were selected by first obtaining all available points corresponding to (a) and (b) and randomly selecting 33 judgment points each, thereby yielding 66 affect judgment points for both (a) and (b). An additional 33 points were randomly selected from the entire tutoring session; these points did not systematically correspond to any particular event. Learners provided affective ratings at these 99 judgment points.

A list of the affective states and definitions was provided for the learners. The states were boredom, confusion, engagement/flow, frustration, delight, surprise, and neutral. These were the affective states that were most frequently experienced in previous studies of AutoTutor that investigated emotions with alternative methods: emote-aloud procedures during learning [44] and online and offline observations by trained judges and untrained peers [45, 46]. Hence, judgments were made on the basis of the learner’s facial expressions, contextual cues via the recorded audio, the definitions of the affective states, and memories of the recently completed tutorial session.

It is important to mention three points pertaining to the retrospective affect judgment methodology. This procedure was adopted because it affords monitoring learners’ affective states at multiple points, with minimal task...
interference, and without learners knowing that these states were being monitored (i.e., during the learning session). Second, this retrospective affect-judgment method has been previously validated [43, 45]. Analyses comparing these offline affect judgments with online measures encompassing self-reports and observations by judges have produced similar distributions of emotions [44, 46]. Third, the offline affect annotations obtained via this retrospective protocol correlate with online recordings of facial activity and body movements in expected directions [23]. Although no method is without its limitations, the present method appears to be a viable approach to track emotion at a relatively fine-grained temporal resolution, with no interruptions of the learning activity, and no demand characteristics of emphasizing emotions during the course of learning.

The affect judgment procedure yielded 99 learner-reported affective states for each learner’s session. Descriptive statistics (means and standard deviation) for proportional occurrence of each affective state are shown in Table 1. This distribution of affective states is consistent with other studies that tracked learner affect during interactions with AutoTutor [44, 45, 47].

<table>
<thead>
<tr>
<th>Affect</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boredom</td>
<td>.160 (.140)</td>
</tr>
<tr>
<td>Confusion</td>
<td>.188 (.098)</td>
</tr>
<tr>
<td>Engagement/flow</td>
<td>.232 (.199)</td>
</tr>
<tr>
<td>Delight</td>
<td>.051 (.068)</td>
</tr>
<tr>
<td>Frustration</td>
<td>.129 (.085)</td>
</tr>
<tr>
<td>Surprise</td>
<td>.021 (.023)</td>
</tr>
<tr>
<td>Neutral</td>
<td>.189 (.132)</td>
</tr>
</tbody>
</table>

### 3. Method

#### 3.1. Judges (Participants)

The judges were 63 undergraduate psychology students from a university in the U.S., who participated for credit in their courses. There were 16 males and 47 females. Data from four judges were discarded due to computer failure.

#### 3.2. Design

The experiment used a 7-condition within-subjects design in which judges assessed learners’ affective states from unimodal and multimodal presentations of the face, speech, and context. As will be explained in the subsequent sections, the context consists of a researcher-generated text transcript of the dialogue exchange between the learner and the computer tutor. The unimodal conditions were Face (F), Speech (S), and Context (C). The bimodal conditions included Face-Speech (FS), Face-Context (FC), and Speech-Context (SC). There was also a trimodal Face-Speech-Context (FSC) condition.

Each judge viewed video clips from 7 different learners across the 7 conditions (clips are described in the Materials subsection). However, they viewed clips from the same learner within a given condition, but the clips were randomly ordered to remove any global contextual cues. For example, a judge might have viewed 10 clips from learner X in the F condition and another set of 10 clips from learner Y in the FS condition.

The affective states of a particular learner in a particular condition were assessed by three judges. This permitted three comparisons for each learner-modality combination, Judge 1 vs. Judge 2, Judge 1 vs. Judge 3, and Judge 2 vs. Judge 3.

The assignment of learners to conditions and the ordering of conditions was counterbalanced across judges with three 7 × 7 Graeco Latin Squares; one square for 7 learners yields three squares for the 21 learners who were selected (out of 30 learners) for the present study (see Materials subsection below). Three judges were assigned to each row of the Graeco Latin Square, so 63 judges were required to complete the design (3 judges for each row × 7 rows in each Latin Square × 3 Latin Squares’ = 63).

It should be noted that this somewhat more complex counterbalancing and design was adopted in lieu of simpler designs in order to satisfy three constraints. First, there was the goal to include clips from as many learners as possible in order to maximize generalizability. Second, it was important to ensure that there were a sufficient number of rating points in each condition, while simultaneously keeping the time demands manageable. Third, there was the need to ensure that assignment of judges, learners, and conditions was balanced while minimizing interference across conditions (i.e., not permitting a judge to be presented with clips from the same learner in multiple conditions).

#### 3.3. Materials

**Video Clips.** The videos collected from the previous study with AutoTutor served as stimuli in the present experiment. Two selection criteria were adopted in deciding which of the 30 videos to include in the present experiment. First, a video was discarded if the learner’s face was not clearly visible during extended periods in the video; this was an occasional problem because the camera was fixed but learners were free to move around in their chairs. Second, only learners who reported an approximately uniform distribution of the affective states were selected. For example, a learner who reported boredom 90% of the time was not included. On the basis of these criteria, videos from 21 out of the 30 learners were selected as stimuli for the present experiment.

The actual stimuli consisted of clips extracted from the 21 videos. The clips were selected so that they corresponded to instances when learners provided judgments on their affective states. This permitted us to compare the judges’ affect judgments (collected in the present experiment) to learners’ self-reported affect judgments (obtained from the previous study). Since occurrences of delight and surprise were rare (see Table 1), we only focused on extracting clips which corresponded to judgments of boredom, engagement/flow, confusion, frustra-
tion, and neutral.

As described in the previous section, the learners provided affect judgments before speaking their responses (1/3rd of judgments), after speaking their responses (1/3rd), and at random points (1/3rd). There were 99 judgment points for each learner’s session, yielding 2079 data points for the 21 selected learners. Since one of the goals of the present study was to test the reliability of affect judgments from speech, we selected only those clips that corresponded to affect judgments after a spoken utterance (random judgment points and judgments prior to speaking were not included). This reduced the set of possible clips to 693 (1/3 × 2079).

Moreover, a subset of 269 clips was selected from this set of 693 clips. The selection was random, but we ensured that the clips corresponded to an approximately uniform distribution of learner-reported affective states both within and across learners. Specifically, 18% of the clips corresponded to boredom, 20% to confusion, 17% to engagement/flow, 22% to frustration, and 23% to neutral. The number of clips per learner ranged from 14 to 15 with a mean of 14.7 (SD = .457).

The clips ranged from 34 to 95 seconds with a mean length of 86.1 seconds (SD = 8.70 seconds). An excerpt of a clip is presented in Table 2. It begins with the tutor providing a hint (“what kind of device is a compact disk?”) to which the learner responded with an answer (“CDR”) and a metacognitive statement (“but I really don’t know”). The tutor then provided a prompt and so on. The clip ends when the learner provided the third response (“Linear storage devices, I really don't know”).

**Table 2**

<table>
<thead>
<tr>
<th>Spkr.</th>
<th>Start (mm:ss)</th>
<th>End (mm:ss)</th>
<th>Spoken Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>12:59</td>
<td>13:16</td>
<td>I think you have covered something important here, you can write data anywhere on a random access device. Let's try this together, what kind of device is a compact disk?</td>
</tr>
<tr>
<td>L1</td>
<td>13:24</td>
<td>13:31</td>
<td>CDR, but I really don't know.</td>
</tr>
<tr>
<td>T2</td>
<td>13:38</td>
<td>13:49</td>
<td>Uh, I bet you know this: floppy disk and compact disk are random access ...?</td>
</tr>
<tr>
<td>L2</td>
<td>13:53</td>
<td>13:56</td>
<td>I don't know.</td>
</tr>
<tr>
<td>T3</td>
<td>14:01</td>
<td>14:15</td>
<td>Could you at least try to give me an answer, that way we could have a starting place to work from. Unlike sequential storage devices, compact disks are</td>
</tr>
<tr>
<td>L3</td>
<td>14:19</td>
<td>14:26</td>
<td>Linear storage devices, I really don't know.</td>
</tr>
</tbody>
</table>

Note. Sprk = Speaker. T1, T2, and T3 denote the first, second, and third tutor turn, respectively. L1, L2, and L3 denote the first, second, and third learner turn, respectively.

Evident from this example, the clips consisted of relatively short but coherent dialogue exchanges between the learner and the tutor. These short video clips were presented to the judges in the present study who judged the learner’s emotions at multiple points in each clip. Judgments were elicited immediately after students’ spoken interactions, which would correspond to turns L1, L2, and L3 in the sample clip presented in Table 2. It should be noted that learner reports of affect were available for only 1/3rd of these points (see previous section), so there are more opportunities to assess agreement between observers (judge) than between an observer and a learner.

It is important to mention one additional point pertaining to the presentation of the clips. The three judgment points within a clip were ordered and this provides a form of local context (in the context conditions) for the short intra-clip exchange between learner and tutor. On the other hand, the inter-clip ordering was random, so judges could be presented with a clip spanning the 10:00 to 10:30 interval before a clip spanning the 03:00 to 04:15 interval. This random ordering process was adopted in order to eliminate the possibility of a pervasive global context; this additional information would bias the local discourse context condition.

**Computer Interface for Stimulus Presentation.** A customized software program was used to display the clips and record the affect judgments. The interface is presented in Figure 1. The video was displayed in the top window, while human transcribed text of learner and tutor speech (i.e., the context) was presented in the bottom window. Controls for playing the video and making affect judgments were also provided (see drop-down list on bottom left).

The interface was configured to appropriately render each of the seven experimental conditions. More specifically, there were five information streams from each clip: (a) the video of the learner’s face, (b) the audio of the learner’s spoken utterance, (c) a transcript of the learner’s spoken utterance, (d) the audio of the tutor’s spoken utterance, and (e) a transcript of the tutor’s spoken utterance. With the exception of the face, which was always present in the video, the other four information sources were only intermittently active over the course of the session. For example, the learner’s audio was not available when the learner was not speaking. Hence, the computer interface controlled the informational content of the video in order to eliminate confounds across conditions (e.g., showing the face throughout the clip but only playing audio when the learner spoke).

As an illustrative example, consider how the sample clip in Table 2 would be presented in the F condition. There were three spoken learner turns in this clip so judges would provide affect judgments immediately after each of these three speech segments (13:31, 13:56, and 14:26). The first segment would appear immediately after the learner spoken turn ended (13:31). This 7-second spoken turn started at (13:24), so a black screen would mask the video until the learner started speaking. The mask would disappear over the 13:24 to 13:31 interval, so the learner’s face would be visible at this time. The audio
channel would have been muted at this time, so learner speech would not be available in the F condition. Additionally, all audio generated by the tutor would also be muted because this is considered to be part of the context and should not be presented in the F condition.

After the speech ends (13:31), the mask would reappear and the video would automatically stop playing. Hence, while the face would be visible during the speech segment, it would be masked when the judge provides the affect judgment. The judge would be required to initiate an affect judgment by pressing the Rate button. The judge would then choose one of five states (boredom, engagement/flow, confusion, frustration, and neutral) from a combo box and would submit the judgments by pressing the Commit button. The masked video would commence to play after the rating was submitted and after the judge clicked the Play button. A similar procedure would be repeated for the second (13:53 to 13:56) and third (14:19 to 14:26) spoken segments.

A similar sequence of events would occur in the S condition. However, the mask would always be present so the face would never be visible, but the audio channel would be unmuted when the learner was speaking (L1, L2, and L3). Therefore, judgments would be made based on the paralinguistic features of the learner’s speech.

The context is considered to be the text of the interchange between the learner and the tutor. In the C condition, the video would be masked throughout and all audio would be muted. However, transcribed text of both the tutor’s and the learner’s speech would appear in the text box at the appropriate intervals (i.e., when these are being uttered). For example, the text of the first tutor’s utterance would appear at 12:59, the text of the students’ response would appear at 13:24, and so on. The text box would be cleared when judges were in the process of making an affect judgment in order to ensure equal exposure to the stimuli, as would be the case for the F and the C conditions. Hence, no stimulus would be available while the judges make their judgments and time on task was consistent across conditions.

As noted in the Introduction, to avoid a confound associated with the content of the learner’s spoken responses being available in all conditions except the F condition, a human-generated text transcription of the learner’s speech was presented in the text box below the video for all conditions. Hence, in the F condition the judges had access to the face and the text of the learner’s spoken utterance but not the audio of the learner’s speech nor the discourse context. In the S condition, the judges had access to both the text and audio of the learner’s speech but not the tutor’s utterance and the learner’s face. Finally, in the C condition, text transcripts of both the tutor’s and the learner’s speech were available, but there was no video of the face and no audio of the learner’s speech.

A mapping of information streams to conditions is presented in Table 3. As explained above, the text transcript of the learner’s speech was available in all conditions. The audio of the tutor’s text was always muted because the synthesized voice was somewhat monotone and there was the concern that this might have biased the judgments. Instead, human-generated (typed) transcripts of the tutor’s speech were provided when applicable. Therefore, there were always two information sources for the F, S, and C conditions, three sources for the FS, FC, and SC conditions, and four sources for the FSC condition.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>MAPPING OF INFORMATION PRESENTED BY CONDITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information Stream</td>
<td>F</td>
</tr>
<tr>
<td>Learner face</td>
<td>+</td>
</tr>
<tr>
<td>Learner speech</td>
<td>+</td>
</tr>
<tr>
<td>Learner text</td>
<td>+</td>
</tr>
<tr>
<td>Tutor speech</td>
<td>+</td>
</tr>
<tr>
<td>Tutor text</td>
<td>+</td>
</tr>
</tbody>
</table>

Note: F: Face; S: Speech; C: Context

### 3.4. Procedure

Judges were tested individually over a 2.5-hour session. After signing an informed consent, the experimenter demonstrated the interface and explained the judgment task to the judges. The judges were also provided the list of affective states with definitions. Next, judges completed seven emotion judgment tasks, one for each condition. They had the option of taking short 1-minute breaks be-
between adjacent conditions, but were required to take a 5-10 minute break after the fourth condition. Judges were fully debriefed after completing all seven conditions.

### 3.5. Data Treatment

Item-level affect judgments in the present experiment consisted of a label of one out of five affective states (boredom, engagement/flow, confusion, frustration, and neutral). Two agreement measures were computed from the present data. First, each affect judgment was assigned either a 1 (agree) or a 0 (disagree) if the two judges agreed or disagreed with each other, respectively. This measure is referred to as *observer-observer* (OO) agreement. Second, agreement scores were also computed on the basis of whether a judgment obtained in the current study aligned with the learner’s self-reported emotions from the previous study. This measure is referred to as *learner-observer* (LO) agreement. There were 13,561 valid OO agreement points across the seven conditions. There were 5,179 LO observations because learner-judgments were only available for approximately a third of OO judgment points.

### 4. Results and Discussion

#### 4.1. Overall Agreement Patterns

The judges agreed on 4,675 of the 13,561 OO observations, yielding a mean proportional agreement score of .345 (SD = .075), with a Cohen’s kappa of 0.162. Agreement for the LO observations was lower with a mean proportional agreement score of .269 (SD = .044) and a kappa of 0.085. Although agreement was generally low, it exceeded chance (kappa of 0) for both OO and LO observations.

#### 4.2. Modality Effects

Descriptive statistics on agreement scores and kappas are presented in Table 4. One-sample t-tests comparing OO kappa scores for each modality to chance (kappa = 0) were statistically significant (p < .05 for this and all subsequent analyses unless specified otherwise) with a mean effect size of 1.44 sigma (Cohen’s d). With the exception of the context (p = .107), the remaining six LO kappa scores were also significantly different from chance (kappa = 0), with a mean effect size of .486 sigma.

The analyses proceeded by investigating whether there were significant differences in agreement scores by modality. A mixed-effects modeling approach was adopted because of the repeated and nested nature of the design, where judgments were nested within clips, clips were nested within learners, modalities and learners were counterbalanced across judges, and judges were paired with other judges (OO observations) or with learners (LO observations). Mixed-effect modeling is the recommended analysis method for this type of complex data [48, 49].

Mixed-effects models include a combination of fixed and random effects. They can be used to assess the influence of the fixed effects on dependent variables after accounting for any extraneous random effects. The lme4 package in R [49] was used to perform the requisite computation with model parameters being estimated with Laplace approximation.

Two models were constructed with either OO or LO agreement scores as the dependent variables. A logistic modeling approach was adopted since these dependent variables are binary at the item level (1 if there is agreement, else 0). The fixed effect was *modality* with seven levels for F, S, C, FS, FC, SC, and FSC. The random effects that were controlled for in the present analyses were *judge-pair, clip, learner, and Latin Square row*. Judge-pair was a categorical variable denoting different pairs of judges for the OO models and combinations of judges and learners for the LO models. Learner and clip were also categorical variables with 269 (# clips) and 21 levels (# learners), respectively. Finally, *Latin Square row* was included as a 21-level categorical variable to account for the different assignments of modalities and learners to judges. This analysis procedure yields a very stringent test on the effect of modality on agreement because it controls for variance associated with differences in the judges, the learners, the clips, and elements of the experimental design (i.e., the Latin Square).

<table>
<thead>
<tr>
<th>Modality</th>
<th>Observer-Observer (OO) Mean Agreement (Stdev)</th>
<th>Learner-Observer (LO) Mean Agreement (Stdev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bimodal</td>
<td>FS .343 (.121) FC .308 (.121) SC .384 (.123)</td>
<td>FS .307 (.143) FC .280 (.142) SC .295 (.123)</td>
</tr>
<tr>
<td>Trinodal</td>
<td>FSC .365 (.133)</td>
<td>FSC .292 (.120)</td>
</tr>
</tbody>
</table>

It is important to mention two additional details pertaining to these mixed effects logistic regression models. First, in addition to constructing the models with modality as a fixed effect (*modality models*), *null models* with the random effects but no fixed effects were also constructed. Comparisons (likelihood ration tests) of the null models with the modality models allowed us to determine whether modality significantly predicts agreement above and beyond the random effects.

Second, instead of performing posthoc tests on all 21 (7 x 6) / 2 possible comparisons between modalities, 12 planned comparisons were used to probe significant main effects. These included comparisons between: (a) the three unimodal conditions (F vs. S, F vs. C, and S vs. C), (b) bimodal conditions vs. the unimodal conditions (FS vs. F, FS vs. S, FC vs. F, FC vs. C, SC vs. S, and SC vs. C), and (c) the trimodal condition vs. the unimodal condi-
tions (FSC vs. F, FSC vs. S, FSC vs. C). A Tukey correction was applied so that the familywise error rate was held constant at .05.

**Observer-Observer Agreement.** A likelihood-ratio test on the OO data indicated that the modality model yielded a significantly better fit than the null model, $\chi^2(6) = 34.7$. This implies that modality was a significant predictor of OO agreement. The planned comparisons indicated that there were no unimodal differences in agreement scores ($F = S = C$). The bimodal combinations of FS and FC did not yield significant differences over the individual modalities, i.e., $FS = F = S$ and $FC = S = C$. However, agreement associated with the SC combination was greater than $C$ but not $S$. Agreement for the FSC condition was not significantly greater than the individual modalities. Hence, with the exception of the one significant effect involving SC and C, the general pattern suggests that there was redundancy between the F, S, and C.

**Learner-Observer Agreement.** A somewhat similar pattern emerged for the LO data. The modality model yielded a significantly better fit than the null model, $\chi^2(6) = 22.5$. There were no significant differences between the unimodal conditions ($F = S = C$). Two significant differences emerged when the bimodality conditions were compared to the unimodal conditions. The FS condition was significantly greater than the $S$ but not the $F$ condition. There was a similar improvement in agreement when SC was compared to $C$ but SC and $S$ were equivalent. Finally, FSC yielded significantly higher agreement than $C$, but not $F$ or $S$. The general pattern in the LO data appears to indicate that the context was not very useful when observer judgments were compared to learner self-reports. The face and the speech appear to be more relevant, but a fusion of these modalities resulted in redundancy instead of more synergistic effects.

### 4.3. Affect × Modality Interaction

The analysis so far has focused on general agreement patterns and did not take into account the specific affective states that the judges were agreeing on. It might be the case that different modalities yield superadditive effects for particular affective states, and redundant or inhibitory effects for others. We addressed this question by performing a follow-up analysis that focused more closely on modality effects associated with individual affective states. Table 5 shows mean proportional agreement scores for each affective state by modality.

**Observer-Observer Agreement.** The analyses proceeded by segregating the 4,675 OO observations for which the judges agreed, and grouping these observations into five categories, one for each affective state.

**Boredom.** A likelihood-ratio test indicated that the modality model yielded a significantly better fit than the null model for boredom, $\chi^2(6) = 192$. The planned comparisons indicated that agreement from the face was significantly greater than agreement from both speech and context, which were on par with one another, $F > [S = C]$. Multimodal conditions involving the face were greater than the unimodal conditions that did not include the face, but were equivalent to the unimodal face condition, $[FS = F] > S; [FC = F] > C; [FSC = F] > [S = C]$. Quite clearly, the face played a critical role in the detection of boredom.

Perhaps the most intriguing finding was that the bimodal speech + context condition yielded significantly greater agreement than when these modalities were individually considered, $SC > [S = C]$. There is the question of whether this significant improvement in SC agreement was consistent with superadditivity or additivity. According to the criterion established in Eq. 2, observed multimodal agreement ($a_{i+2}$) should exceed expected additive agreement ($add_{exp}$) before superadditivity can be declared. If $a_i$ is designated as speech and $a_2$ as context, with values of .025 and .012, respectively (from Table 5), then expected additivity $a_{i+2}$ is .037 (from Eq. 2). From Table 5 we note that the proportional agreement for SC was .046, which exceeds expected additivity (.037), thereby allowing us to conclude that the combined effect of speech and context was consistent with superadditivity for boredom.

**Confusion.** There was a significant effect of modality for confusion, $\chi^2(6) = 22.7$, but the comparisons revealed only one significant difference. It appears that judges were more likely to agree that the learner was experiencing confusion when provided with the speech compared to the context. There was also a marginally significant effect ($p = .059$) in favor of the FSC combination compared to the context alone. In general, the results are suggestive of redundancy among modalities.

**Engagement/Flow.** The modality model was significant when compared to the null model for engagement/flow, $\chi^2(6) = 51.9$. The planned comparisons yielded a number of important effects. One finding was that judges were more likely to agree in the context condition...
compared to the face condition. However, of greatest significance was the fact the combined FSC model yielded significantly lower agreement than the individual F, S, and C modalities. This is consistent with an inhibitory effect for engagement/flow.

**Frustration.** A likelihood-ratio test indicated that the modality model yielded a significantly better fit than the null model for frustration, \( \chi^2(6) = 36.2 \). The planned comparisons indicated that the three individual modalities were on par (F = S = C), but the SC combination was significantly greater than the context and marginally significantly greater than the speech (\( p = .055 \), SC > [S = C]). However, SC agreement (.175 from Table 5) did not exceed expected additivity (.217, see Eq. 2), leading us to conclude that the combination of speech and context yielded an additive effect for frustration.

**Neutral.** The modality model was significant for neutral, \( \chi^2(6) = 41.2 \). Planned comparisons showed that agreement rates from the face and speech were on par with each other and were significantly lower than the context condition, C > [F = S]. The FC, SC, and FSC conditions all yielded agreement rates that were lower than the C condition. This pattern suggests that context was critical in judging neutral, a finding that is contrary to what was obtained for boredom. Specifically, while the face played a critical role in the perception of boredom, the context was very relevant for neutral.

**Learner-Observer Agreement.** The learner and the observer agreed for 1,395 of the observations. Similar to the previous analysis, separate models were run on each of the five affective states to test for a modality x affect interaction in LO agreement patterns. However, none of the models were significant.

### 4.4. Size of MM Effects

Similar to the meta-analysis on affect detection systems described in the Introduction, we computed MM effect sizes for the FSC multimodal combinations using Eq. 1. The focus was on the trimodal combination in order to facilitate comparisons with the meta-analysis on automated affect detectors, which focused on MM effects associated with combinations of all available modalities. These results are shown in Figure 2 for both OO and LO observations.

The overall FSC multimodal effect for OO observations was 3.40%, which is somewhat similar to the 4.39% effect obtained from the automated affect detectors that were trained and validated on more naturalistic data (see discussion on meta-analysis in the Introduction). There was a much larger improvement (MM Effect of 14.1%) for the LO observations.

Consistent patterns among OO and LO observations were discovered for engagement/flow, frustration, and neutral. More specifically, combining three modalities was detrimental for engagement/flow and neutral but was beneficial for frustration. OO and LO patterns somewhat differed for boredom and confusion, but the size of these effects (effects ranged from -5.26% to 10.5%) were considerably smaller than engagement/flow, frustration, and neutral (effects ranged from -42.8% to 26.4%).

### 5. General Discussion

We conducted a human emotion perception study in which 63 untrained judges provided judgments of naturalistic expressions of non-basic affective states from unimodal and multimodal combinations of face, speech, and context obtained in human-computer interaction sessions. Our focus was on quantifying agreement at multiple levels including modality, affective state, and affect judges (observer-observer vs. learner-observer), with the goal of applying our findings to fully-automated affect detection by computers. In this section we take stock of our findings, discuss limitations, and consider implications of this study to multimodal affect detection.

#### 5.1. Overview of Present Findings

The first finding was that agreement associated with perception of the non-basic affective states was low with overall kappas of 0.165 for observer-observer (OO) agreement and 0.085 for learner-observer (LO) agreement. This low agreement highlights the difficulty of detecting naturalistic expressions of affect that are sometimes muted instead of being prominently displayed [50]. These findings are within the range of the few previous studies that have investigated this issue in similar contexts [34, 35]. They are also consistent with the general finding that humans’ emotion perception abilities are modest when emotions are not intentionally elicited (e.g., by actors) and when information from multiple, and sometimes conflicting channels needs to be considered [32, 51-54]. Furthermore, while much of the previous research has focused on the basic emotions in human-human interactions, we show that this finding replicates to a different set of affective states, namely the non-basic states of confusion, frustration, boredom, and engagement/flow, and in human-computer interactions. Despite the low agreements obtained in the present study, it is important to point out that with one exception (LO agreement in the context condition), agreement scores were significantly greater than random guessing. This suggests that although there is considerable noise in the face, speech, and context signals, and their multimodal combinations, the signal is still detectable from the surrounding noise.
Our second important finding was that there were no significant unimodal effects on overall agreement rates. The lack of significant differences associated with the face and speech suggests that these modalities provide cues that are equally diagnostic of learner affect. What is significant, however, is that agreement in the context condition rivaled the face and speech conditions. Recent statements in the literature have suggested that context plays a major role in shaping emotion perceptions [28, 37, 52, 55, 56]. Situational context is often thought to impact emotion perception by providing top-down influences on bottom-up facial and vocal processing [52, 55-58]. According to the conceptual-act model of emotion [32, 56, 59], the face conveys very general information (e.g., approach or avoidance), whereas the situational context helps to shape the specific emotion being perceived. The present findings contribute to this growing literature by indicating that even impoverished contextual cues such as brief exposures to the local dialogue history between learner and tutor can yield agreement rates that are statistically indistinguishable from the face and speech.

The third finding is related to important differences between the affective states. Although there was a lack of a modality effect when overall agreement was considered, there were some informative differences when affective states were individually examined. However, the differences were only significant for the OO models. The clearest pattern emerged for boredom and neutral. It appears that boredom was more reliably detected from the face, F > [S = C], but it was the context that was more reliable for neutral, C > [F = S]. The fact that the face was quite diagnostic of boredom is particularly surprising because some of our previous research has failed to identify clear facial correlates for this state [44, 50]. Perhaps the judges were sensitive to a noticeable lack of facial activity while making boredom judgments, but this hypothesis warrants empirical support. Aside from boredom and neutral, unimodal effects were also discovered for confusion (S > C) and engagement/flow (C > F). Taken together, the results indicate that no single modality yielded consistently higher agreement, and there was considerable evidence for a modality × affect interaction. Furthermore, although the differences were not significant for the LO models, the direction of the LO effects (see Figure 2) were mostly consistent with OO effects for engagement/flow, frustration, and neutral.

The fourth finding was that, overall, unimodal and multimodal agreement was quantitatively equivalent. This finding is more in line with emotion models that espouse a loose coupling between the different bodily correlates of an emotion [26, 31, 33, 56, 59, 60]. According to these models, that there is no central affect program that coordinates the various components of an emotional episode, which is a major tenet of the theory of basic emotions [27, 28]. Instead, emotion components are loosely coupled and the specific context and situational appraisals determine which bodily systems are activated. Therefore, coordinated bodily responses associated with particular emotions are rare. These models would accommodate the prediction that with the exception of rare cases of prototypical emotions or acted emotions, a combination of modalities might conceivably yield null or small improvements in multimodal accuracy.

That being said, the discovery of a modality × affect interaction introduces some complications because an entire range of effects such as inhibition, redundancy, additivity, and even superadditivity were discovered for different affective states. For example, a fusion of speech and context in the OO models yielded superadditivity for boredom, weak additive effects for frustration, inhibition for engagement/flow, and redundancy for confusion and neutral. One general pattern is that the multimodal conditions yielded improvements in agreement when unimodal agreement was low (i.e., boredom and frustration). Multimodal combinations were rather ineffective (i.e., confusion and neutral) or even detrimental (i.e., engagement/flow) when there was considerable agreement in the unimodal conditions.

Finally, there were several points of convergence and some points of divergence between the OO and LO models. In terms of overall agreement, both models yielded equivalence across unimodal conditions and a general lack of impressive multimodal effects. However, overall agreement for the LO models was lower, but FSC multimodal effects were quantitatively higher than OO models. It should be noted that the OO and LO comparisons should be interpreted with caution since direct comparisons between observer judgments and learner-reported affect are difficult in the present study. This is because observers in the current experiment provided affect judgments from short clips in seven conditions, while the learners in the previous study had access to the videos of their entire sessions and with all five information channels listed in Table 3. Furthermore, while the observers only had access to the local context (short transcripts of tutorial dialogue), the entire dialogue history was available to the learners. The comparison is further complicated by a lack of a ground-truth model of learners' affective states. One possible solution to this issue is to obtain an estimate of ground truth by tailoring the tutorial session to induce specific emotions and then comparing unimodal and multimodal OO and LO agreements to one another and to this more objective ground truth standard (i.e., the elicited emotion). This is an important item for future work.

5.2. Limitations
There are four primary limitations with the present study. First, there are concerns that the use of a forced-choice affect judgment methodology might artificially inflate reliability [61]. We adopted a forced-choice paradigm in the present experiment in order to be consistent with the previous study where learners self-reported their emotions using a forced-choice measurement instrument. This concern can be alleviated with a modified forced-choice paradigm when a “none of these terms are correct” or an “other” option is added to the list of emotion terms [62]. We should point out that the “other” option was rarely used (< 5%) in one previous study involving learners self-reporting their affective states after interactions with Au-
to Tutor [63]. Additionally, we have some confidence that the forced-choice paradigm did not threaten the internal validity of the experiment because there is no reason to assume that it would differentially impact one condition over another.

The second limitation was that there was no pure speech condition because the content of the learner’s spoken utterance was available. More specifically, while both the linguistic and paralinguistic features of speech were naturally available in the speech conditions, the content of the speech (i.e., the linguistic information) was not naturally available in the face condition. In order to eliminate this potential confound, the content of the spoken utterance was displayed (in text form) to the judges in the face condition. As a consequence, there were no true unimodal face and speech conditions. Of course, this is an unintended consequence of using naturalistic stimuli and as indicated in the Introduction, previous research has indicated that the short 3-5 word fragments of learner speech does not provide viable cues into their affective states.

The third limitation is that the relatively short video clips of approximately 90 seconds might not have provided sufficient cues to accurately detect naturalistic affect expressions that can be quite subtle. It might be the case that a longer time frame is needed before affect can be accurately detected. Previous research has indicated that observers can accurately perceive emotions based on relatively short (thin-slice) displays of behaviors [64, 65]. However, it is unclear if these findings generalize to the present stimuli and to non-basic affective states, so it would be advisable to replicate this study with longer clips that contain more contextual cues.

The fact that only 21 learners served as stimuli constitutes the fourth limitation because there might be considerable individual differences in how non-basic states like confusion and frustration are expressed. In addition, affective responses were only collected from a single stimulus, i.e., interactions with AutoTutor. In addition, we did not measure affective intensity, so there is the unanswered question of whether the generally low agreement scores might be attributed to general difficulty of the judgment task or to a specific case of lower intensity emotional expressions. It might be the case that the results shown here might not replicate with different stimuli, especially when a different set of more intense affective expressions are at play. Hence, increasing the size and diversity of the stimuli is an important item for future work.

5.3. Implications for Multimodal Affect Detection
Our findings have two important implications for multimodal affect detection. The first implication pertains to the accuracy rates of human- and machine-generated affect labels. On average, and independent of modality, observers only agreed with other observers for approximately 1/3rd of the observations. Observer-learner agreement was even lower in that there was agreement for approximately 1/4th of the observations. It has previously been documented that humans overestimate their own abilities at perceiving emotions [66], but this study and previous studies have demonstrated that humans are not very accurate at detecting context-sensitive naturalistic expressions of emotion [59-62]. It is therefore quite conceivable that automated classifiers might be able to detect emotions with accuracies that rival, or even exceed, the humans.

The second implication of our findings is related to some of the consequences of considering multiple modalities for affect detection. On one hand, combining modalities did not have much of an effect when overall classification rates are considered. This present finding with humans replicates an earlier finding with multimodal affect detectors [25]. On the other hand, no single modality could be declared the consistent winner in the present study. Instead, affect and modality were intrinsically coupled, in that certain affective states were best detected from specific modalities. This suggests that there might be advantages to including different modalities, provided they are combined in a manner that best reflects the intrinsic alignment between affect and modality. Simply put, some affective states might best be detected from a single dominant modality, while others might require multiple modalities to be accurately detected.

An interesting next step of this research is to train affect detection systems to classify learners’ emotions in this database and repeat the present study with the automated classifier instead of the humans as the affect judges. In addition to pursuing this line of research, we plan on making the present database of naturalistic non-basic affective states during human-computer interactions available to other researchers to test their affect detection models. These two research efforts will facilitate the triangulation of different judgment sources: actor (learner), observer (judge), and computer, thereby affording interesting comparisons of unimodal and multimodal affect detection capabilities of humans and computers.

ACKNOWLEDGMENT
The authors gratefully thank Patrick Chipman, Rosaire Daigle, Brandon King, Bethany McDaniel, Evie Johnson, and Kristy Tapp for assistance with data collection.

This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, and DRL 125958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES


[49] D.M. Bates and M. Maechler, “lme4: Linear mixed-effects models using S4 classes,” Book lme4: Linear mixed-effects models using S4 classes, Series lme4: Linear mixed-effects models using S4 classes, ed., Editor ed., eds., 2010, pp.


**AUTHOR BIOS**

Sidney D’Mello is an Assistant Professor in the departments of Computer Science and Psychology at the University of Notre Dame. His research interests include affective computing, artificial intelligence, human-computer interaction, speech recognition, and natural language understanding. He has published over 150 journal papers, book chapters, and conference proceedings. He is an associate editor for IEEE Transactions on Affective Computing and an advisory editor for the Journal of Educational Psychology. D’Mello received his PhD in Computer Science from the University of Memphis in 2009.

Nia Dowell is a cognitive psychology graduate student at the University of Memphis. Her primary interests are in computational linguistics and affective sciences. More specific interest focuses on modeling discourse and social dynamics in authoritarian regimes, emotions, and terrorism research.

Art Graesser is a professor in the Department of Psychology at the University of Memphis and is a Senior Research Fellow in the Department of Education at the University of Oxford. Graesser’s primary research interests are in cognitive science, discourse processing, and the learning sciences. He has published over 500 articles in journals, books, and conference proceedings.