Language and Discourse are Powerful Signals of Student Emotions during Tutoring

Sidney K. D’Mello and Art Graesser

Abstract—We explored the possibility of predicting student emotions (boredom, flow/engagement, confusion, and frustration) by analyzing the text of student and tutor dialogues during interactions with an intelligent tutoring system with conversational dialogues. After completing a learning session with the tutor, student emotions were judged by the students themselves (self judgments), untrained peers, and trained judges. Transcripts from the tutorial dialogues were analyzed with four methods that included (a) identifying direct expressions of affect, (b) aligning the semantic content of student responses to affective terms, (c) identifying psychological and linguistic terms that are predictive of affect, and (d) assessing cohesion relationships that might reveal student affect. Models constructed by regressing the proportional occurrence of each emotion on textual features derived from these methods yielded large effects ($R^2 \approx 38\%$) for the psychological, linguistic, and cohesion-based methods, but not the direct expression and semantic alignment methods. We discuss the theoretical, methodological, and applied implications of our findings towards text-based emotion detection during tutoring.

Index Terms—emotions, affect, intelligent tutoring system, affect from text

1 INTRODUCTION

It is frequently assumed that emotions are vague, diffuse, and difficult to pin down. Similar claims are periodically made about language. However, it is conceivable that both of these assumptions are overstated, if not misleading. Moreover, it may very well be the case that language and discourse have features that are quite effective in signaling emotions of conversational partners. This is precisely the question we explore in this article. We are investigating the extent to which characteristics of language and discourse diagnostically reveal the emotions of students as they attempt to learn from a conversational computer tutor. Our lens is on the textual aspects of language and discourse, not the paralinguistic cues. It is perfectly obvious and well documented that speech intonation, gestures, and facial expressions manifest the emotions of speakers (see [1, 2] for reviews) However, these paralinguistic levels of communication are not at all relevant to the goals of the present study. Our goal is to ascertain what aspects of the verbal text are diagnostic of the emotions of students while they learn with a computer tutor.

Investigating student emotions during learning is critical because learning at deeper levels of comprehension is inherently an emotionally rich experience [3, 4]. Although, advanced learning environments have come a long way towards providing individualized instruction by modeling and responding to students’ cognitive states, the link between emotions and learning suggests that they should be more than mere cognitive machines; they should be affective processors as well [5, 6]. An affect-sensitive ITS would incorporate assessments of students’ cognitive and affective states into its pedagogical and motivational strategies in order to keep students engaged, boost self-confidence, heighten interest, and presumably maximize learning.

A computer tutor can never respond to students’ emotions if it cannot detect their emotions. Hence, emotion detection is an important challenge that needs to be adequately addressed before computer tutors can respond in an affect-sensitive fashion. Text-based affect detection is particularly attractive because it does not require any specialized sensors, is non-intrusive, inexpensive, scalable, and can be readily deployed in authentic learning contexts such as classrooms and computer labs. Our previous research has already established that some conversational characteristics from tutor logs can predict the emotions of students [7]. The learning environment was an automated computer tutor (AutoTutor) that helped students learn about difficult topics, such as computer literacy, by holding a conversation in natural language [8]. The emotions that we discovered to be prominent during these challenging learning sessions were confusion, frustration, boredom, and flow (engagement); delight and surprise occurred with lower frequencies (see [9] for a review). The same set of emotions are prominent in other learning and problem solving environments (see [10] for synthesis of studies that have tracked emotions during learning with a variety of advanced learning technologies).

We found text and discourse to be particularly diagnostic for many of the emotions, a finding that motivated the present study. We have made some progress in identifying features of the discourse experience that telegraph
particular emotions, but have only begun to skim the surface of how emotions are manifested in the text features extracted from student responses. The purpose of this study is to dig deeper and explore how far we can go in determining how a broad profile of language and discourse characteristics predict the four emotions (confusion, frustration, boredom, and flow) that we already know exist during learning sessions with human and computer tutors. Our approach to conducting a more thorough analysis of text is to use recent computer tools that incorporate advances in computational linguistics and computational discourse. These include Coh-Metrix [11] and the Linguistic Inquiry Word Count (LIWC) [12]. To what extent do the metrics supplied by these computerized scaling programs predict the emotions of students while they hold conversations with AutoTutor?

To answer this question, we analyzed tutorial dialogues from a previous study [13] in which 28 students learned computer literacy topics with AutoTutor. After the tutorial interaction, the students’ affective states were rated by the learners themselves, untrained peers, and two trained judges. Our use of multiple judges is justified by the fact that there is no clear gold standard to declare what the students’ states truly are. Therefore, we considered models that generalized across judges as well as models that were specific to the affect ratings of any particular judge.

The extent to which tutorial dialogues were predictive of student emotions was assessed via four methods: direct expression, semantic alignment, LIWC, and Coh-Metrix (these are described below). We conducted analyses on a large set of measures that spanned states of cognition, emotion, tutorial dialogue, and other aspects of the session. These measures have theoretical and practical significance that need to be placed in context. Our goal is to identify the particular aspects of the tutorial session that are highly diagnostic of students’ emotional states. A measure is highly diagnostic if it predicts a particular emotion very well, it does so better than alternative measures, and it predicts over and above the contributions of other measures.

Due to the relatively small size of our data set (28 participants), we relied on relatively simple statistical models, specifically multiple regressions, rather than some of the more complex machine learning algorithms, to identify such diagnostic features. As explained in more detail in the Methods and Results sections, the models focused on predicting the proportional occurrence of a set of four affective states across a 32-minute session rather than on the presence or absence of individual emotion episodes. Although this does limit the practical utility of this work (as explained in the General Discussion), the ability to identify a diagnostic feature set of affect via this coarse-grained analysis represents an important step towards developing more fine-grained text-based affect detectors for human-computer tutorial dialogues.

We in fact were successful in identifying a small set of measures in the tutorial dialogue that predicted the proportional occurrence of each state across the entire learning session, as well as a large set of measures that were unsuccessful. We begin by discussing some existing methods to text-based affect detection, followed by a description of the text-analysis approach adopted in this paper.

2. PRIOR RESEARCH ON TEXT AND AFFECT

Considerable research has demonstrated that textual features have been useful for predicting a number of psychological phenomena ranging from personality to depression to emotion (see [14, 15] for a review). For example, in an analysis of one-on-one tutorial dialogues with a computer tutor, Ward and colleagues have shown that the extent to which the students and tutors align (or converge) at a semantic level can be predictive of a number of interesting outcomes, such as changes in motivation orientations [16], and learning gains [17]. Although these and other studies have used text-analysis techniques similar to the ones analyzed here, the present focus is on the application of text-processing methods on affective data. As such, this brief review of the literature will exclusively focus on the problem of predicting affective states from text, a field that is sometimes called sentiment analysis. It should also be noted that although sentiment analysis is a burgeoning research field [15, 18], text-based affect detection has rarely been automated in learning environments (see [19, 20] for notable exceptions). Our survey of the literature will therefore consider some of the general approaches to text-based affect detection, which can be conveniently aligned into four major research thrusts.

Some of the first attempts identified a small number of dimensions that underlie expressions of affect [21, 22]. This research was pioneered by Osgood and colleagues, who analyzed how people in different cultures rated the similarity of common words. Dimensional reduction analyses on the similarity matrices converged upon evaluation (i.e., good or bad), potency (i.e., strong or weak), and activity (i.e., active or passive) as the critical dimensions. These dimensions of affective expressions are qualitatively similar to valence and arousal, which are considered to be two of the fundamental dimensions of affective experience [23].

The second strand of research involves a lexical analysis of the text in order to identify words that are predictive of the affective states of writers or speakers [14, 24, 25]. Several of these approaches rely on the Linguistic Inquiry and Word Count (LIWC) [12], a validated computer tool that analyses bodies of text using dictionary-based categorization. LIWC based affect-detection methods attempt to identify particular words that are expected to reveal the affective content in the text. For example, first person singular pronouns in essays (e.g., “I”, “me”) have been linked to negative emotions [26, 27].

In addition to LIWC, researchers have developed lexical databases that provide affective information for common words. Two examples of this are WordNet-Affect, an extension of WordNet for affective content [28] and the Affective Norm for English Words [29]. These affective databases provide normed ratings of valence, arousal,
and dominance for common words. The affective gist of texts can then be derived by aggregating the valence, arousal, and dominance scores associated with the content words in the text. Indeed, this method can be quite efficient and effective for text-based affect detection, as recently demonstrated by [30].

The third set of text-based affect detection systems go a step beyond simple word matching by performing a semantic analysis of the text. For example, Gill and colleagues [31] analyzed 200 blogs and reported that texts judged by humans as expressing fear and joy were semantically similar to emotional words, such as phobia, terror, for fear and delight, bliss for joy. They used Latent Semantic Analysis (LSA) [32] and the Hyperspace Analogue to Language (HAL) [33] to automatically compute the semantic similarity between the texts and emotion keywords (e.g., fear, joy, anger). Although this method of semantically aligning text to emotional concept words showed some promise for fear and joy texts, it failed for texts conveying six other emotions, such as anger, disgust, and sadness. So it is an open question whether semantic alignment of texts to emotional concept terms is a useful method for emotion detection.

The fourth and perhaps most complex approach to textual affect sensing involves systems that construct affective models from large corpora of world knowledge and applying these models to identify the affective tone in texts [15, 34, 35]. For example, the word “accident” is typically associated with an undesirable event. Hence, the presence of “accident” will increase the assigned negative valence of the sentence “I was late to work because of an accident on the freeway”. These approaches are gaining traction in the computational linguistics community and are extensively discussed in a recent review [15].

3. CURRENT APPROACHES TO TEXT-BASED AFFECT DETECTION

3.1. Direct Expression and Semantic Alignment

As evident from the discussion above, the last decade has witnessed a surge of research activity aimed at automatically identifying emotions in spoken and written text. Most of the systems have been successfully applied to targeted classes of texts with notable affect, such as movie reviews, product reviews, blogs, and email messages [31, 34, 35]. Despite the methodological differences in the various approaches, one commonality that emerges is that the systems operate under the assumption that affective content is explicitly and literally articulated in the text (e.g., “I have some bad news”, “This movie is a real drag”). Another assumption is that the texts contain words that have marked affective valence (e.g., “accident”, “crash”, “smiling”). Although this may be a valid assumption for obviously emotion-rich corpora such as blogs and movie reviews, where people are directly expressing opinions, this may not generalize to student responses to computer tutors.

The question of whether students freely express affect in their responses is a critical issue. This is because relatively straightforward methods will suffice to detect student emotions if their responses resonate with a sufficient degree of affective content. For example, a bored student might say, “I am bored” or “This material is boring”. If students do openly express their emotions to the tutor in this fashion, then regular expressions for affective terms and phrases might be sufficient to monitor students’ affective states. The present paper refers to this approach as direct expression.

Alternatively, students might not express their affective states to the tutor in a direct and explicit fashion. Instead, their responses may convey affective content indirectly. For example, a confused student might say, “I really am not understanding this material”, and an exasperated student might state, “I don’t know what to do. I keep trying and trying with no luck so I’m going to give up”. In these examples, affective content is expressed in the semantics of the message, even though there are no explicit emotion words (e.g., “confused”, “stuck”). There is some research to suggest that the underlying semantics of people’s utterances do convey affective information [31]. We refer to models that detect affect by comparing the semantic content of the text to emotion terms as semantic alignment models.

The question remains unanswered as to whether direct expression or semantic alignment of texts to emotional concept terms are valid methods for monitoring affective states in learning contexts. This question is therefore explored in the current paper. Failure of these approaches would suggest that a more systematic analysis of tutorial dialogues might be necessary to uncover subtle cues that might be diagnostic of students’ emotions. Therefore, we explored two alternate methods to detect student emotions from the language and discourse in the tutorial session. These included shallow assessments, such as the usage of psychological and linguistic terms, as well as deeper natural language processing (NLP) techniques, such as assessing discourse cohesion. There of course is no sharp boundary between shallow and deep NLP techniques. We are simply contrasting NLP methods that primarily rely on simple word counting (i.e., shallow methods) and those that perform a more processing intensive linguistic analysis (i.e., deep methods). More specifically, the Linguistic Inquiry and Word Count (LIWC) tool [36] analyzed the words in order to extract a set of psychological and linguistic features from the student and tutor dialogues. The Coh-Metrix program [11] was used to derive multiple measures of discourse cohesion. The features extracted from these two NLP tools were used to predict student emotions.

3.2. Linguistic Inquiry and Word Count (LIWC)

LIWC is a validated computational tool that analyses bodies of text using dictionary based approaches. LIWC has a large lexicon that specifies how each word of a text is mapped onto one or more predefined word categories. For example, “crying”, and “grief” are words in the sad category, whereas “love” and “nice” are words that are assigned to the positive emotion category. The version of LIWC (LIWC2007), which was used in the current study, provides approximately 80 word categories. LIWC oper-
ates by analyzing a transcript of text and counting the number of words that belong to each word category. A proportion score for each word category is then computed by dividing the number of words in the text that belong to that category by the total number of words in the text.

We focused on LIWC’s psychological and linguistic features. LIWC provides 32 features that are indicative of psychological processes. We selected a subset of 15 psychological features for the current analyses. We mainly focused on affective and cognitive terms that were found to be diagnostic of emotions in previous analyses [14, 24, 25]. LIWC provides 28 linguistic features that comprise function words, various types of pronouns, common and auxiliary verbs, different tenses, adverbs, conjunctions, negations, quantifiers, numbers, and swear words [36]; 15 of these linguistic features were selected. Table 1 presents the features selected for the present analysis.

### TABLE 1. PSYCHOLOGICAL AND LINGUISTIC FEATURES OF LIWC

<table>
<thead>
<tr>
<th>Psychological Features</th>
<th>Linguistic Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1. Affect terms (happy, sad)</td>
<td>L1. Function words</td>
</tr>
<tr>
<td>P2. Positive emo. (love, nice)</td>
<td>L2. 1st prsn. sing. pronoun (I, me)</td>
</tr>
<tr>
<td>P3. Negative emo. (hurt)</td>
<td>L3. 1st prsn. plrl. pronoun (we, us)</td>
</tr>
<tr>
<td>P4. Anxiety (worried)</td>
<td>L4. 2nd prsn. pronoun (you, your)</td>
</tr>
<tr>
<td>P5. Anger (hate, kill)</td>
<td>L5. 3rd prsn. sing. pronoun (she, her)</td>
</tr>
<tr>
<td>P6. Sadness (crying, grief)</td>
<td>L6. 3rd prsn. plrl. pronoun (they)</td>
</tr>
<tr>
<td>P7. Cognitive terms (know)</td>
<td>L7. Impersonal pronoun (it, it’s)</td>
</tr>
<tr>
<td>P8. Insight (think, know)</td>
<td>L8. Auxiliary verbs (am, will)</td>
</tr>
<tr>
<td>P10. Discrepancy (should)</td>
<td>L10. Present tense (is, does)</td>
</tr>
<tr>
<td>P11. Tentative (maybe)</td>
<td>L11. Future tense (will, gonna)</td>
</tr>
<tr>
<td>P13. Inhibition (constrain)</td>
<td>L13. Prepositions (to, with)</td>
</tr>
<tr>
<td>P15. Exclusive (but, without)</td>
<td>L15. Negations (no, not)</td>
</tr>
</tbody>
</table>

Note. emo. = emotion, prsn. = person, plrl. = plural, prnou = pronoun.
Examples were obtained from the LIWC 2007 Language Manual [36]

### 3.3. Coh-Metrix

A more complex approach involves an analysis of cohesion relationships in the tutoring dialogues. Cohesion, a textual construct, is a measurable characteristic of text that is signaled by relationships between textual constituents [11]. Cohesion is related to coherence, a psychological construct that is a characteristic of the text together with the reader’s mental representation of the substantive ideas expressed in the text.

Coh-Metrix is a computer program that provides over 100 measures of various types of cohesion, including coreference, referential, causal, spatial, temporal, and structural cohesion [11]. Coh-Metrix also has hundreds of measures of linguistic complexity (e.g., syntactic complexity), characteristics of words, and readability scores. Coh-Metrix is substantially more complex than LIWC so a comprehensive description is beyond the scope of this paper. Table 2 presents the measures of Coh-Metrix that were analyzed in the present study. The selected measures have been validated in their ability to discriminate between low and high cohesion texts in a variety of corpora (see [37] for a summary).

### Table 2. Cohesion Features Derived from Coh-Metrix

<table>
<thead>
<tr>
<th>Co-Reference Cohesion</th>
<th>Connectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>C5. Pronoun ref. cohesion</td>
<td>C17. Negative additive</td>
</tr>
<tr>
<td>Semantic Cohesion</td>
<td>Other Measures</td>
</tr>
<tr>
<td>C8. LSA sen. to sen. adj. mean</td>
<td>C20. Mean hypernym nouns</td>
</tr>
<tr>
<td>C9. LSA sen. to sen. adj. stddev</td>
<td>C21. Mean hypernym verbs</td>
</tr>
<tr>
<td>C10. LSA turn to turn mean</td>
<td>C22. Content words</td>
</tr>
<tr>
<td>C11. LSA turn to turn stddev</td>
<td>C23. TTR for content words</td>
</tr>
<tr>
<td>C12. Given info. in each sen.</td>
<td>C24. Flesch Reading Ease</td>
</tr>
</tbody>
</table>

Note. adj. = adjacent, sen. = sentence, info. = information, TTR = type token ratio, ref. = referential

### Causal cohesion

Causal cohesion occurs when actions and events in the text are connected by causal connectives and other linking word particles [11]. Events and actions have main verbs that are designated as intentional or causal (e.g., “kill”, “impact”), as determined by categories in the WordNet lexicon [39]. Causal particles connect the-
Se events and actions with connectives, adverbs, and other word categories that link ideas (e.g., “because”, “consequently”, “hence”). Coh-Metrix provides measures on the incidence of causal verb categories (occurrences per 1000 words) (C6). The most significant measure of causal cohesion is the causal ratio that specifies the ratio of causal particles to events and actions (C7). A high causal ratio indicates that there are many connectives and other particles that stitch together the explicit actions and events in the text.

**Semantic cohesion.** In addition to the co-reference variables discussed earlier, Coh-Metrix assesses conceptual overlap between sentences by a statistical model of word meaning: Latent Semantic Analysis (LSA) [32]. LSA is a statistical technique for representing world knowledge, based on a large corpus of texts. The central intuition is that two words have similarity in meaning to the extent that they occur in similar contexts. For example, the word hammer will be highly associated with words of the same functional context, such as screwdriver, tool, and construction. LSA uses a statistical technique called singular value decomposition to condense a very large corpus of texts to 100-500 dimensions (Landauer et al., 2007). The conceptual similarity between any two excerpts of text (e.g., word, clause, sentence, text) is computed as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine typically varies from 0 to 1.

Coh-Metrix uses the LSA cosine scores between texts segments to assess semantic cohesion. Adjacent sentences that have higher LSA overlap scores (i.e., higher semantic similarity) are more cohesive than adjacent sentences with low LSA scores. Both mean LSA scores and standard deviations in the scores are computed (see C9-C11). A semantic cohesion gap is expected to occur for adjacent sentences with low LSA scores and also when there is a high standard deviation (because of an occasional adjacency with a very low LSA score). The given information measure (C12) computes the extent to which an incoming sentence is redundant with (i.e., LSA overlap) the previous sentences in the dialogue history for a particular problem.

**Connectives.** Connectives are words and phrases that signal cohesion relationships by explicitly linking ideas expressed in a text [11]. Coh-Metrix provides incidence scores on several types of connectives. In addition to all categories of connectives (C13), which include the causal and intentional connectives to assess causal cohesion, we segregated the temporal (e.g., “before”, “when”), additive (e.g., “also”), and conditional (e.g., “if”, “else”) connectives. Temporal and additive connectives have both negative (e.g., “however”, “in contrast”) and positive valences (e.g., “therefore”, “in addition”) (C13-C18).

**Other measures.** Coh-Metrix provides many measures of words and language in addition to the primary cohesion measures. We included a measure of the incidence of negations (C19). There were measures of the degree of abstraction of nouns and verbs in the text (C20, C21), obtained from the hypernym index in WordNet [39]; lower values of hypernym indicate the word is more abstract.

The incidence of content words (e.g., nouns, main verbs, adverbs, adjectives) was also included (C22) as a measure of the amount of substantive content in the text. The type-token ratio for content words is an index of lexical diversity, with a value of 1.0 meaning that each word was used once in the text whereas values nearer to zero mean that words were frequently repeated (C23). The last two measures assessed reading ease (C24) and verbosity (C25). Reading ease was measured by the Flesch Kincaid Reading Ease score [40] whereas verbosity was measured as the length of the longest sentence in the text (C25).

### 4. Method

**4.1. Participants**

The participants were 28 undergraduate students from a mid-south university who participated for extra course credit. They were lower division undergraduate students enrolled in an introductory psychology class.

**4.2. AutoTutor**

AutoTutor is a validated intelligent tutoring system (ITS) that helps students learn topics in Newtonian physics, computer literacy, and critical thinking via a mixed-initiative conversational dialogue between the student and the tutor [8, 41]. The impact of AutoTutor in facilitating the learning of deep conceptual knowledge has been validated in over a dozen experiments on college students [8, 41].

AutoTutor’s dialogues are organized around difficult questions and problems (called main questions) that require reasoning and explanations in the answers. AutoTutor actively monitors students’ knowledge states with LSA and weighted keyword matching and engages them in a turn-based dialogue as they attempt to answer these questions. It adaptively manages the tutorial dialogue by providing feedback (e.g., “good job”, “not quite”), prompting the student for more information (e.g., “What else”), giving hints (e.g., “What about X?”), prompts (e.g., “X is a type of what”), identifying and correcting misconceptions, answering questions, and summarizing answers. Students typed (instead of speaking) their responses in the version of AutoTutor used in this study. A sample dialogue between a college student and AutoTutor is presented in the excerpt below.

```
1. AUTOTUTOR: When you turn on the computer, how is the operating system first activated and loaded into RAM? [this is the main question]
2. STUDENT: When the computer is turned on a file is automatically booted up.
3. AUTOTUTOR: Anything else? [this is a pump for more information]
4. STUDENT: This file contains the operating system.
5. AUTOTUTOR: What about storage? [this is a hint]
6. STUDENT: The operating system is then read into RAM and activated.
7. AUTOTUTOR: Right. The CPU executes instructions in ROM, which helps the operating system to start. [after giving positive feedback, there is an assertion]
```
4.3. Procedure

Interaction with AutoTutor. Students interacted with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, Internet, or operating systems. Each of these topics had 12 questions that required about a paragraph of information (3-7 sentences) in an ideal answer. The participants never completed all 12 questions during the 32 minutes; the mean number of questions completed was 5 (SD = 2).

A video of the participant’s face and a video of their computer screen were recorded throughout the tutoring session. The video included the synthesized speech generated by the animated conversational agent.

Judging affective states. Similar to a cued-recall procedure [42] the judgments for a student’s tutoring session proceeded by playing a video of the face along with the screen capture video of the interaction with AutoTutor on a dual-monitor computer system. The screen capture included the tutor’s synthesized speech, printed text, the student’s responses, the dialogue history, and images, in essence recreating the context of the tutorial interaction.

Judges were instructed to make judgments on what affective states were present at any moment during the tutoring session by manually pausing the videos (spontaneous judgments). They were also instructed to make judgments at each 20-second interval where the video automatically stopped (fixed judgments). Judges were provided with a checklist of seven states for them to mark along with definitions of the states (see [13] for definitions of the emotions). Hence, judgments were made on the basis of the student’s facial expressions, contextual cues via the screen capture, the definitions of the cognitive-affective states, and recent memories of the interaction (for self reports only as described below).

Four sets of judgments were made for the observed affective states of each AutoTutor session. First, for the self-judgments, the student watched his or her own session with the tutor immediately after having interacted with AutoTutor. Second, for the peer judgments, each student came back a week later to watch and judge another student’s session. Finally, there were two trained judges: undergraduate research assistants who were trained extensively on AutoTutor’s dialogue characteristics (i.e., the context) and how to detect facial action units according to Ekman’s Facial Action Coding System [43]. The two trained judges scored all sessions separately.

It is important to mention three important points pertaining to the present affect judgment methodology. This procedure was adopted because it affords monitoring participants’ affective states at multiple points, with minimal task interference, and without participants knowing that these states were being monitored while completing the learning task. Second, this retrospective affect-judgment method has been previously validated [42], and analyses comparing these offline affect judgments with online measures encompassing self-reports and observations by judges have produced similar distributions of emotions (see [9] for a review). Third, the offline affect annotations obtained via this retrospective protocol correlate with online recordings of facial activity and gross body movements in expected directions [44]. Although no method is without its limitations, the present method appears to be a viable approach to track emotions at a relatively fine-grained temporal resolution.

4.4. Scoring of Data Collected in Tutorial Session

Proportions of affective states experienced. The affect judgment procedure yielded 2,967 self-judgments, 3,012 peer judgments, and 2,995 and 3,093 judgments for the two trained judges. We examined the proportion of judgments that were made for each of the affect categories, averaging over the 4 judges. A repeated measures ANOVA indicated that there was a statistically significant difference in the distribution of states, F(6, 162) = 10.81, Mse = .023, p < .001, partial eta-square = .286. Bonferroni post-hoc tests revealed that the proportional occurrence of boredom (M = .160, SD = .140), confusion (M = .180, SD = .127), flow/engagement (M = .199, SD = .161), frustration (M = .114, SD = .107), and neutral (M = .288, SD = .248) were on par and significantly greater than delight (M = .032, SD = .038) and surprise (M = .027, SD = .028), which were equivalent to each other. Delight and surprise were excluded from the subsequent analyses because they were relatively rare.

Reliability between judges. We evaluated the reliability by which the affective states were rated by the four judges. Proportional agreement scores for the six judge pairs were: self-peer (.279), self-judge1 (.364), self-judge2 (.330), peer-judge1 (.394), peer-judge2 (.368), and judge1-judge2 (.520). These scores indicate that the trained judges had the highest agreement, the self-peer pair had the lowest agreement, and the other pairs of judges were in between. Another finding is that there are actor-observer differences in the agreement scores. The average actor-observer agreement was .324 (i.e., average of self-peer, self-judge1, and self-judge2), which is lower than the average observer-observer agreement score of .427 (i.e., average of peer-judge1, peer-judge2, judge1-judge2).

Although the agreement scores appear to be low, they are on par with data reported by other researchers who have assessed the problem of measuring complex psychological constructs, such as emotions [45-48]. Agreement is low when emotions are not intentionally elicited, contextual factors play an important role, and the unit of analysis is on individual emotion events. It is unclear who provides the most accurate judgments of the learner’s affective states [13]. Is it the self, the untrained peer, the trained judges, or physiological instrumentation? A neutral, but defensible position is to independently consider ratings of the different judges, thereby allowing us to examine patterns that generalize across judges as well as patterns that are sensitive to individual judges. This strategy was adopted in the current paper.

Extracting transcripts of tutorial dialogues. At the end of each student turn, AutoTutor maintained a log file that captured the student’s response, a variety of assessments of the response, the feedback provided, and the tutor’s next move. Transcripts of the tutorial dialogue between the student and the tutor were extracted for each problem that was collaboratively solved during the tutorial ses-
The tutorial sessions yielded 164 student-tutor dialogue transcripts. The transcripts contained 1,637 student and tutor turns.

Two sets of responses were obtained from the transcripts. The first, called student responses, were obtained by only considering the student turns in each transcript. The second dialogue category, called tutor responses, consisted of the tutor’s statements. The purpose of dividing the transcripts into separate student and tutor responses is to assess the impact of each response category in predicting the student’s emotions.

Computing LIWC and Coh-Metrix features. Psychological and linguistic features were computed for the student and tutor dialogues using LIWC 2007. This resulted in 60 features (30 for each response category). Similarly, 50 cohesion features (25 for each response category) were computed using Coh-Metrix 2.0. The text submitted to each computational tool consisted of the responses generated during the solution of an individual problem. An aggregated score for each predictor was derived for each subject by averaging the scores across problems. Hence, the unit of analysis for the subsequent analyses was an individual subject.

5. DIRECT EXPRESSION AND SEMANTIC ALIGNMENT

Direct Expression. According to direct affect expression models, students directly express their affective states to the tutor. We tested this model by inspecting the transcripts and counting the frequency of direct affective expressions. Three researchers coded the 1,637 student responses for the presence (coded as 1) or absence (coded as a 0) of any emotional expression. The coding scheme for an emotional expression was quite broad. A response that contained any emotional term (anger, confusion, confused, bored, sad, etc.) was scored as a 1.

Due to the simplicity of the coding scheme, reliability between the three coders was a perfect 1.00 kappa. The results indicated that there was only one occurrence of a direct emotional expression in the 1,637 student responses. This singular instance of an emotional expression (“I’m confused”) was scored by all three human coders. Therefore, it is quite clear that students did not directly express their emotions to the tutor.

Although students almost never directly express their affective states to the tutor, there might be expressions with affective semantics in a more indirect manner. An examination of our corpus of dialogues indicated that there did appear to be some instances of these. For example, an utterance that semantically aligns with frustration is, “How am I suppose to know that if I didn't know the last question that you asked me man!!!”

Semantic Alignment. We examined whether the semantic similarity of students’ utterances to emotional concepts could be predictive of their emotional states. The semantic similarity of all 1,637 utterances to each of the emotional terms was computed using LSA. This method is similar to the procedure used by Gill and colleagues [31] in their study of the emotional content in blogs. Our analyses proceeded by considering each emotional term independently and computing the semantic similarity between the emotional term and each of the 1,637 utterances. The emotional terms were “bored”, “confused”, “engaged”, and “frustrated”. The semantic scores between the utterances and emotional terms were aggregated for each student separately, so that each student contributed six LSA scores, one for each affective state. The online LSA tool (http://lsa.colorado.edu/) was used to perform the requisite computation.

Multiple linear regression (MLR) analyses were used to predict the proportional occurrence of each emotion on the basis of the semantic scores. For example, the semantic scores between the student responses and the emotional concept term “bored” were used to predict the proportional occurrence of boredom in each response. Four regression models were constructed for each affective state, given that there were four judges of emotions (self, peer, and the two trained judges). The dependent variable for a given emotion was the proportion of occurrence for the emotion as reported by the self, peer, and two trained judges. There were 24 of these MLR analyses, given there were 4 judges and 6 emotions.

The results indicated that none of the models were statistically significant ($p > .05$) and on average explained approximately 0% of the variance. Therefore, the typed expressions of the students were not semantically aligned with the emotion terms for the relevant emotions in this study. Instead, they mainly consisted of domain related responses to the tutor’s questions (e.g., “RAM is short term memory”).

6. PSYCHOLOGICAL, LINGUISTIC, AND COHESION FEATURES

We conducted a series of multiple linear regression (MLR) analyses with the textual features as predictor variables and the proportions of the affective state as criterion variables. Note that we are not predicting the presence or absence of each emotion at each judgment point. Rather, the models are predicting the proportional occurrence of each emotion either as reported by an individual judge, averaged across all judges, or an average of a subset of judges. It should also be noted that from this point on, the term “emotion” and “proportional occurrence of emotion” are used interchangeably when we refer to the dependent variables of the MLR models. Hence, a model that “predicts boredom” is actually predicting the proportional occurrence of boredom.

In addition to constructing the most robust models for each affective state, there was also the goal of quantifying the predictive power of each feature set (i.e., psychological, linguistic, versus discourse cohesion). The MLR models were constructed in two phases to address each of these goals. First, we compared the predictive power of each feature set by considering the feature sets individually. Next, the most diagnostic predictors of each feature set were collectively included as predictors of the affective states.
6.1. Selecting Diagnostic Predictors

The fact that multiple judges were used to operationally define the emotions of the student needs to be considered in the construction of the regression models. Although it is possible to construct models on the basis of each judge’s affect judgment, this opens the door to complications when the judges have different results. Therefore, it is advantageous to select diagnostic predictors that generalize across the different judges of affect prior to the construction of the MLR models. These features were selected in a preliminary correlational analysis.

The analyses proceeded by constructing a correlational matrix for each emotion. This was a 110 × 4 (feature × judge) matrix that consisted of the correlation between the features and the proportional occurrence of the affective state as reported by the four judges (see Table 3). The matrix for each emotion was examined separately. We selected features that significantly correlated with ratings by at least two of the four judges. This procedure narrowed the landscape of potential predictors to one psychological feature, five linguistic features, and seven cohesion features. In this fashion, the large set of 110 features was effectively reduced to 13 potential predictors (see Table 3).

### TABLE 3. DIRECTION (+, -) OF STATISTIcALLY SIGNIFICANT CORRELATIONS BETWEEN FEATURES AND EMOTIONS.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Bor</th>
<th>Con</th>
<th>Flow</th>
<th>Fru</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Psychological</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P12. [S] Certainty</td>
<td></td>
<td></td>
<td></td>
<td>++</td>
</tr>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L15. [S] Negations</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4. [S] 2nd Prsn. Pnoun</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L11. [T] Future Tense</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L7. [S] Impersonal Pnoun.</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C19. [S] Negations</td>
<td>+++</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5. [T] Pnoun Ref. Coh.</td>
<td>-</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6. [S] Casual Ratio</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1. [S] Noun Overlap Adj. Sen.</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7. [S] Causal Verbs</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. [S] and [T] refer to features derived from student or tutor responses, respectively. + and - indicates that a feature is a significant (p < .05) positive or negative predictor of an affective state. Empty cells indicate that the correlations between the predictor and affective state are not statistically significant. S, P, I, and 2 = self, peer, trained judge 1, and trained judge 2, respectively. Bor. = Boredom, Con. = Confusion, Flo. = Flow/engagement, Fru. = Frustration.

The requirement that predictors have to significantly correlate with the affective states ensures that only diagnostic predictors are considered for further analyses. By ensuring that the predictors correlate with affect scores of at least half of the judges, there is some modicum of confidence that results will generalize across judges.

There are some risks in this approach to selecting diagnostic features in situations where there is a degree of ambiguity in the measurement of the criterion variable, as in our case with affect. One risk is our committing Type I errors because significance tests are conducted on a large number of correlations. This risk is mitigated to some extent by requiring a significant correlation by at least 2 of the 4 judges. It is also important to note that the feature set we examined had been reduced on the basis of prior research and theoretical considerations. A set of 25 features was selected theoretically from the large feature bank (> 600) provided by Coh-Metrix, whereas only 30 out of the 80 categories provided by LIWC were included. We could possibly narrow our focus to 2 or 3 predictors, but this is not an attractive alternative because there would be a risk of Type II errors to the extent that we overlooked important text-based cues that are diagnostic of the affective states. Yet another option would be to make the significance criteria more stringent by applying the Bonferroni correction. However, this would result in an extremely small alpha value, which is also not desirable, as this substantially increases the risk of committing Type II errors.

Nevertheless, a closer examination of the correlations depicted in Table 3 suggests that it is unlikely that our results were obtained by a mere capitalization on chance. The risk of false discoveries is mitigated by the requirement that a feature is only considered for further analysis if it significantly correlates with affective judgments by at least two judges. For example, consider the causal ratio, which significantly correlates with judgments of flow by the peer and both the trained judges. Although it is quite possible that this predictor could correlate with one of the judge’s ratings of flow by chance alone, it is highly unlikely that chance effects only can result in this predictor correlating with flow judgments by three of the affect judges. Furthermore, the causal ratio does not correlate with any of the other affective states, indicating that its predictive relationship with flow cannot be attributed to a mere capitalization on chance.

6.2. Regression Models for Individual Feature Sets

We constructed several MLR models with the 13 features listed in Table 3 as predictor variables. The dependent variable was the average proportion of each emotion averaged across the four judges. The standardized coefficients of the MLR models are listed in Table 4. It appears that with the exception of frustration, the psychological feature set was not effective in predicting the student emotions (see Table 3). In contrast to the psychological models, the linguistic and cohesion features were able to predict the emotions. These models are independently examined below.
action [11]. The construction of a situation model is essential for learning at deeper levels of comprehension [49]. Engaged students use this mental representation to produce causally cohesive responses to the tutor’s questions.

Confusion was marked by a breakdown in understanding the pronouns expressed by the tutor. This predictor measures the proportion of pronouns that have a grounded reference. Reading ease and comprehension are compromised when the students do not understand the referents of pronouns. Hence, it is no surprise that tutor responses that have a higher proportion of ungrounded pronouns are linked to heightened confusion.

Frustrated students provided responses with cohesion gaps (noun overlap across adjacent sentences). This may have occurred because it is difficult for them to compose a cohesive message or because they were allocating most of their cognitive resources to managing their frustration.

Comparing the predictive power of linguistic and cohesion features. Although the same predictor (i.e., the incidence of negatives) was diagnostic of boredom in both models, the cohesion model explained somewhat less variance ($R^2_{adj.} = .191$) than the linguistic model ($R^2_{adj.} = .136$). A one-parameter model was the most effective in predicting flow for both models. The linguistic model yielded an impressive $R^2_{adj.}$ of .527, which is quantitatively superior to the $R^2_{adj.}$ of .318 obtained from the cohesion model. Hence, flow is best predicted by the linguistic features.

The best linguistic model for confusion was a two-parameter model that yielded an $R^2_{adj.}$ of .347. A quantitatively lower $R^2_{adj.}$ of .223 was obtained from the best cohesion model, which was a one-parameter model. Although it is tempting to conclude that confusion is best predicted by the linguistic features, this would be an erroneous conclusion since the number of predictors differs across models. Therefore, we conclude that both models are equivalent in their ability to predict confusion. The best cohesion model yielded an $R^2_{adj.}$ of .258 for frustration, while the linguistic features were unable to predict this affective state.

In summary, both models have their associated strengths and weaknesses that render them on par with each other. The linguistic models have a somewhat higher precision than the cohesion models (average $R^2_{adj.} = .301$ for linguistic and .248 for cohesion). However, they have lower recall because only three of the emotions could be detected for the linguistic models, while all four could be detected for the cohesion models.

### 6.3. Regression Models that Combine Linguistic and Cohesion Predictors

The next set of MLR models were constructed by an additive combination of the most diagnostic predictors from the previous analysis (L15, L4, L11, L7, C19, C5, C6, and C1). It is important to mention some important details before we proceed with a description of these models. Both feature sets indicated that negations were the most diagnostic predictor of boredom. Therefore, we did not construct a composite model for boredom. A composite

### Table 4. Parameters of Multiple Regression Models for Linguistic and Cohesion Features

<table>
<thead>
<tr>
<th>Coefficient Weights: $\beta$ ($p$)</th>
<th>Bor</th>
<th>Con</th>
<th>Flow</th>
<th>Fru</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L15. [S] Negations</td>
<td>.412 (.03)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L4. [S] 2nd Prsn. Phn.</td>
<td>-.365 (.04)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L11. [T] Future Tense</td>
<td>.472 (.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L7. [S] Impersonal Phn.</td>
<td>.739 (&lt;.001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Cohesion</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C19. [S] Negations</td>
<td>.473 (.02)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5. [T] Pronoun Ref. Coh.</td>
<td>-.505 (.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6. [S] Causal Ratio</td>
<td>.589 (&lt;.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C1. [S] Noun Olap. Adj. Sen.</td>
<td>-.536 (.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


**Linguistic predictors.** None of the linguistic predictors correlated with frustration, so we constructed models for boredom, $F(1, 25) = 5.11$, $p = .033$, $R^2_{adj.} = .136$, confusion, $F(2, 22) = 7.38$, $p = .004$, $R^2_{adj.} = .347$, and flow, $F(1, 24) = 28.9$, $p < .001$, $R^2_{adj.} = .527$. Negations were the only significant predictor of boredom. An examination of the tutorial dialogue of highly bored students indicated that these students used a large number of negations (e.g., “no”, “never”) as well as negative frozen expressions such as “I don’t care” or incessantly repeating, “I don’t know”. In contrast to this, engaged (flow) students used many impersonal pronouns, suggesting that they provided more material-focused responses.

Confusion was best predicted by a two-parameter model that included a lower use of 2nd person pronouns by the student coupled with an increase in future tense words by the tutor. It is quite plausible that confusion was increased when the tutor used more future terms, ostensibly since these terms routinely accompany deep knowledge questions (e.g., “What would you say about X?”), “How should X effect Y?”, “What should you consider if …?“). These questions require students to think, reason, and problem solve, so they are expected to generate confusion in the students.

**Cohesion predictors.** Significant models were discovered for boredom, $F(1, 24) = 9.60$, $p = .015$, $R^2_{adj.} = .191$, flow $F(1, 22) = 11.17$, $p = .002$, $R^2_{adj.} = .318$, confusion, $F(1, 24) = 8.20$, $p = .009$, $R^2_{adj.} = .223$, and frustration $F(1, 24) = 9.68$, $p = .005$, $R^2_{adj.} = .258$. Similar to the linguistic analysis, the cohesion analysis indicated that boredom was accompanied by an increased use of negations by the student. For flow, the data suggest that causally cohesive responses of the student were predictive of this state. The ability of students to produce such responses indicates that they were able to construct a causal situation model (sometimes called a mental model) that links events and
model for frustration was also not constructed because this emotion could only be predicted from one cohesion feature.

Composite models were therefore constructed for confusion and flow. Although the best linguistic model for confusion was a two-parameter model (2nd person pronouns + future tense), we proceeded with the best one-parameter model (future tense). A two-parameter linguistic model plus a one-parameter cohesion model would yield a three-parameter model. This would result in an overfitting problem due to our relatively small sample size of 28 participants. Therefore, the composite model for confusion was constructed on the basis of future tense words by the tutor (a linguistic predictor) and pronoun referential cohesion in the tutor responses (a cohesion predictor). These predictors were not correlated, $r(23) = -.081, p = .701$. The predictors for flow were impersonal pronouns (a linguistic predictor) and the causal ratio (a cohesion predictor). These predictors were also not correlated, $r(26) = .235, p = .229$.

The composite model for confusion was statistically significant, $F(2, 21) = 9.51, p = .001$. This model yielded an $R^2$ adj. of .425, which is approximately twice the variance explained by individually considering future tense words ($R^2$ adj. = .239) and pronoun referential cohesion ($R^2$ adj. = .223). The coefficients of both features were statistically significant ($p < .01$), $\beta_{[T]}$Future Tense = .511, $\beta_{[T]}$Pronoun Referential Cohesion = -.446.

The composite model for flow was also significant, $F(2, 23) = 24.86, p < .001$. This model yielded an impressive $R^2$ adj. of .656, which is quantitatively greater than the proportion of variance explained by individually considering impersonal pronouns ($R^2$ adj. = .527) and the causal ratio ($R^2$ adj. = .330). The coefficients of both features were statistically significant ($p < .01$), $\beta_{[S]}$Impersonal Pronouns = .607, $\beta_{[S]}$Causal Ratio = .394.

The fact that combining predictor sets resulted in a 78% and 24% improvement over the best individual predictors for confusion and flow, respectively, suggests that the combination resulted in additive effects. Therefore, we can conclude that each feature set explains unique aspects of the variance. Tutor utterances that have ungrounded pronoun references as well as future tense words are related to an increase in confusion. These features are probably indicative of two different sources of confusion. The confusion associated with ungrounded pronouns is more likely to be linked to comprehension problems caused by the lack of referent grounding by the tutor. On the other hand, confusion associated with future tense words is more related to the effect of the tutor’s suggestions and questions that try to direct the student, but the student does not quite know what to do. The pattern for flow is more straightforward. Engaged students are material-focused and provide causally cohesive responses.

In summary, our results indicate that six features were successfully able to predict the four learning-centered emotions. On average, the models explained 40% of the variance, which is consistent with a large effect (for power = .8) [50].

### 6.4. Generalizability Across Affect Judges

Earlier we described a method to reduce the feature landscape to a handful of diagnostic predictors that generalized across judges. According to this feature selection strategy, only predictors that significantly correlated with affect judgments provided by at least two of the four judges were included in multiple regression models. The regression analyses subsequently reduced to a set of six predictors that were the most diagnostic of the affective states. But the feature selection method only required features to correlate with at least two judges ratings. It did not require that the features be evenly distributed across the judges.

We conducted a set of follow-up multiple regression analyses to assess the generalizability (across judges) of our feature set. These analyses utilized the six most diagnostic features, but the dependent variable for an affective state was the proportion of its occurrence as reported by the self, peer, and two trained judges.

The results yielded medium to large effects for the peers (mean $R^2$ adj. across emotions = .185) and the two trained judges mean $R^2$ adj. across emotions = .277), however, with the exception of flow ($R^2$ adj. = .246), none of the self-reported affective states could be predicted with our six-item feature set ($R^2$ adj. approximately = 0). This apparent discrepancy suggests that the different classes of judges (self versus others) are sensitive to different aspects of the tutorial dialogues. It might also be prudent to refer to the current set of six features as the Observer Feature Set.

We performed a follow-up analysis to address the question of whether it is possible to predict self-reported student affect from the textual features. In particular, we constructed regression models with an alternate set of predictors that correlate with the self-judgments, but not with judgments by the peers and trained judges.

### 6.5. Predicting Self-Reported Affective States

The dependent variables for the subsequent analyses were the proportions of the four emotions, as reported by the self. As before, we identified significant correlations between four emotions and the LIWC and Coh-Metrix features. There were three significant correlations for boredom (P10, L13, C13), four for confusion (C13, L9, P4, P13), nine for frustration (P12, P6, C5, P8, P10, P7, C25), and none for flow. We constructed three regression models that attempted to predict boredom, confusion, and frustration from this set of predictors. We used stepwise regression methods [51] to isolate individual predictors or combinations of predictors that yielded the most robust models. The stepwise procedure resulted in the selection of two diagnostic predictors for each self-reported affective state.

**Boredom.** The MLR analyses resulted in a significant model for boredom, $F(2, 23) = 8.11, p = .004$, $R^2$ adj. of .363. 

---

\[\text{Improvement over Max} = \frac{R^2_{\text{adj.}} - \max(R^2_{\text{adj.}})}{\max(R^2_{\text{adj.}})} \times 100.\]
It appears that self-reported boredom is signaled by tutor responses that lack discrepant terms (e.g., “should”, “would”, “could”) ($\beta = -0.498, p = .005$) and are highly cohesive due to a high incidence of connectives ($\beta = 0.469, p = .008$). The lack of discrepant terms in the tutor responses indicates that the tutor is directly asserting information to the student with well-crafted cohesive summaries of topics. These direct tutor moves that apparently elicit boredom can be contrasted with more indirect prompts and hints that are linked to confusion [7].

**Confusion.** The MLR analyses resulted in a highly robust model for confusion, $F(2, 22) = 11.7, p < .001$, with a $R^2_{adj}$ of .472. The predictors were students’ responses that were lacking in connectives ($\beta = -0.485, p = .004$) but with an increase in inhibitory terms ($\beta = 0.462, p = .005$). It is informative to note that confused students provide responses that are rife with inhibitory terms such as “block”, “constrain”, and “stop”. Although confused students do not directly express their confusion, their responses inevitably convey their confusion via words that imply that they feel blocked, stopped, and constrained. Furthermore, the responses of these confused students were not sufficiently cohesive as they lack connectives to bind together substantive ideas.

**Frustration.** The analyses revealed some interesting insights into self-reported episodes of frustration. Frustration was predicted by a two-parameter model $F(2, 21) = 9.14, p = .001$, with a $R^2_{adj}$ of .414. It appears that frustrated students provide responses that are verbose ($\beta = .397, p = .040$) and the tutor responds with words that lack certainty ($\beta = -.398, p = .040$), ostensibly because it cannot fully comprehend the students’ responses.

**Psychological, linguistic, or cohesion features.** Although models that compared the effectiveness of each of the predictor types were not considered in this set of analyses, an examination of the coefficients of the models can provide some insights. Each affective state was predicted by one psychological feature and one cohesion feature. There was no linguistic feature that was sufficiently diagnostic of self-reported affective states. Let us refer to this set of features as the **Self Feature Set**. The fact that psychological predictors play a critical role in the **Self Feature Set** invalidates our earlier conclusions pertaining to the lack of diagnosticity of the psychological features. These psychological features were significant predictors in the **Self Feature Set** but not the **Observer Feature Set**. This suggests that psychological cues are on the radar of the students themselves but are overlooked by the other judges. Unlike the linguistic features that were included in the **Observer Feature Set** but not the **Self Feature Set**, the cohesion features were important predictors in both feature sets. Therefore, a deeper analysis of textual dialogues that the cohesion features provide yields the most generalizable models.

### 6.6. Generalizing across Participants

The small sample size of 28 participants raises some important generalizability issues. That is, to what extent do the regression models generalize to new participants? We addressed this question by performing a bootstrapping analysis, which is the recommended validation technique to assess generalizability and overfitting of models constructed on small data sets [52]. The analysis proceeded as follows. Training data was obtained by sampling a subset of the participants; models were then fit on the training data and training $R^2$ was computed. The training models were then applied to the testing data, which consisted of the training participants plus novel participants not included in the training data. Goodness of fit for the testing data ($testing R^2$) was obtained. Overfitting was computed as the difference between testing and training $R^2$ values. This procedure was repeated for 500 iterations and average $R^2$ values were computed.

The results of the bootstrapping procedure for models constructed from the **Observer** and **Self feature sets** are presented in Table 5. Although there was some overfitting (mean $R^2$ decreased from .47 to .38 from training to testing), $R^2$ values obtained from the testing data were consistent with large effects [50] for both the for the **Observer** and **Self feature sets**. Therefore, we have considerable confidence that the regression models generalize to novel participants.

### Table 5. Average $R^2$ across 500 runs of bootstrapping

<table>
<thead>
<tr>
<th>Affect</th>
<th>Observer Features</th>
<th>Self Features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>Boredom</td>
<td>.287</td>
<td>.225</td>
</tr>
<tr>
<td>Confusion</td>
<td>.512</td>
<td>.402</td>
</tr>
<tr>
<td>Flow</td>
<td>.705</td>
<td>.640</td>
</tr>
<tr>
<td>Frustration</td>
<td>.293</td>
<td>.232</td>
</tr>
<tr>
<td>Mean</td>
<td>.449</td>
<td>.375</td>
</tr>
</tbody>
</table>

### 7. General Discussion

This study considered the possibility of predicting student affect during tutoring by casting a wide net that encompassed four affect-detection methods, the perspectives of four affect judges, and an analysis of both student and tutor responses. Our results confirmed that textual cues can be a viable channel to infer student affect and the best models included an interaction among measures (i.e., language and discourse analysis methods), affect judges, and conversational participants (students and tutor). We proceed by presenting an overview of our findings, discussing some limitations of the study, and considering potential applications of our results.

#### 7.1. Overview of Findings

Our results support a number of conclusions pertaining to text-based affect detection from naturalistic tutorial dialogues. The failure of the direct expression and semantic alignment models, coupled with our own inspection of the dialogues, suggest that tutorial dialogues are not rife with affectively charged words. The affective terms in LIWC (P1 to P6 in Table 1) were similarly not diagnostic of student emotion. The important message is that methods that detect affect in a text by identifying explicit articulations of emotion or by monitoring the valence of its
content words are not always viable solutions. They are functional when valence scores can be reasonably applied to words like "accident", "rainy", "beautiful", "sunny", and "dark" in many forms of opinion-laden discourse. However, these methods fail when the major content words are not affectively charged, as is the case in educational contexts. In these learning contexts, the major content words are affect-neutral terms such as "RAM", "ROM", "speed", "velocity", "acceleration", "quotient", "remainder", and "divisor". Although the present analysis was limited to computer literacy dialogues, we have also analyzed samples of AutoTutor dialogues for topics in Newtonian physics, and critical thinking. The results were strikingly similar in the sense that the students' responses were comprised of domain related content words; affective expressions or affectively charged words were conspicuously absent.

One might claim that deriving conclusions from studies with AutoTutor is problematic because students might be less inclined to directly express affect to a computer. However, we have also analyzed transcripts from 50 tutorial sessions between students and human tutors [53]. The findings with the computer tutor were replicated with this sample as well.

Moving beyond the direct expression and semantic alignment models, our results indicate that a combination of LIWC and Coh-Metrix features explained an impressive amount of variance in predicting the major learning-centered emotions. The Observer Feature Set, consisting of four cohesion features and two linguistic features explained 38% of the variance (from bootstrapping analysis) in predicting observer-reported affect. The Self Feature Set, consisting of three psychological predictors and three-cohesion predictors also explained 38% of the variance (from bootstrapping analysis) for self-reported affect. These results support three conclusions pertaining to the feasibility of text-based affect detection for tutorial dialogues.

The first insight is that it takes a combination of psychological, linguistic, and cohesion features to predict student affect. Although models could be derived by considering each feature set independently, the most predictive models almost always encompassed a combination of feature sets. Combining the most diagnostic predictors from each feature set yielded additive effects suggesting that the different feature sets explain unique aspects of the variance. Simply put, each feature set is differentially diagnostic of student affect.

The second important finding was that the features predicted an approximately equivalent amount of variance for self-reported versus observer-reported emotions. More interestingly, however, is the fact that the self and observer models were comprised of unique feature sets. The self-reported models consisted of a combination of cohesion and psychological predictors, whereas the observer-reported models consisted of linguistic and cohesion features. There was no overlap in features across self and observer reported models. This result supports the claim that the different judges were sensitive to different aspects of the tutorial dialogue. Hence, composite models that incorporate perspectives of both the learners themselves and the observers are perhaps most defensible.

The third informative finding is that it takes an analysis of responses from both the student and the tutor to derive the most robust models. In particular, observer-reported confusion and self-reported boredom were predicted entirely from features derived from the tutor’s responses. Self-reported frustration was predicted from one tutor feature and one student feature. It is therefore important to consider the conversational dynamics of both student and tutor to predict student affect.

7.2. Limitations and Resolutions
One potential concern with the multiple regression models is that some of the effects might be linked to the tutorial domain. Our current set of models was constructed entirely from computer literacy tutorial dialogues. It is conceivable that some of the predictors will not generalize to other domains, such as physics, critical thinking, and other sciences.

We partially addressed this concern by assessing the impact of the tutoring subtopics on the predictive power of the two feature sets. Tutorial dialogues with affect judgments were not available from domains other than computer literacy, but there was sufficient variability within the computer literacy topics (hardware, operating systems, Internet) to partially address some of the domain related concerns. More specifically, there is a graded difference in difficulty of the sub-topics and each sub-topic includes its own set of concepts, terms, and acronyms, thereby providing alternate sources of variability. Fortunately, a follow-up MLR analysis indicated that the twelve diagnostic predictors remained significant after controlling for individual subtopics. Therefore, we have some confidence that the diagnostic features generalize above and beyond differences in computer literacy sub-topics. Nevertheless, the analyses need to be replicated in other domains and with other learning environments.

Another limitation of the present study pertains to the unit of analysis. The current set of regression models were constructed at the subject level because the primary goal was to explore the possibility of deriving a set of predictors that were diagnostic of student emotions. However, subject-level emotion prediction might not be very useful for affect-sensitive systems that need to be dynamically responsive to student affect. These systems require turn-based affect detection in order for the tutor to incorporate assessments of the sensed emotion in selecting its next dialogue move (e.g., withhold negative feedback because student is currently frustrated). The temporal resolution of the emotion judgments in the present study raised some challenges towards the development of turn-based affect detection. The basic problem was that an emotion judgment was obtained every 20 seconds, not after every turn, thereby making it difficult to link a particular emotion to a particular student or tutor turn.

This limitation can be addressed in two ways. First, the study can be replicated so that affect judgments are collected after each student and tutor response. Standard
machine learning technique can be subsequently applied to classify emotions in each turn. An alternate approach would involve extending the existing regression models to student emotions as they occur by analyzing incremental windows of student and tutor dialogues that are generated as the tutoring session progresses. There is also the question of whether the current MLR models will generalize to this form of more fine-grained emotion monitoring or if a different feature set will be needed. These possibilities await future research.

7.3. Applications
The applications of this research venture into the goal of designing tutoring systems that are sensitive to students’ emotional states in addition to their cognitive states. Although the use of physiological and bodily sensors represent viable solutions to detect affect [1], one disadvantage is that these sensors require expensive customized hardware and software (e.g., Body Pressure Measurement System and automated facial feature tracking systems). This raises some scalability concerns for those who want to extend this program of research beyond the lab and into the classroom. It is in the applied context of the classroom where text-based affect detectors have a unique advantage over bodily and physiological sensors provided the ITS engages in natural language dialogue. Text-based affect sensing is advantageous because it is cost effective, requires no specialized hardware, is computationally efficient, and is available to any ITS with conversational dialogues. Although text-based affect detectors currently relinquish center stage to the more popular bodily and physiological approaches, they are expected to play a more significant role in next-generation affect detection systems, particularly when efficiency, cost-effectiveness, and scalability are important concerns.

In line with this, the present paper has shown that the student-tutor dialogues contain latent cues that are predictive of student emotions. The MLR models that we developed are currently somewhat limited for affect-sensitive ITSs because they predict the proportional occurrence of each emotion across the entire session instead of the presence or absence of individual emotional episodes. Despite this limitation, they can still be used to provide a general sense of whether the student is likely to be bored, engaged, confused, or frustrated as the session unfolds. The real utility of these models, however, lies in the fact that they demonstrate that a small set of text-based predictors can be quite diagnostic of student affect. The next step is to extend this line of research to predict emotion at a much finer temporal resolution, perhaps at the turn level or the problem level. Whether fully automated text-based affect detectors can complement, or even replace, existing systems that monitor physiological and bodily measures awaits future research and empirical testing.

ACKNOWLEDGMENT
The research on was supported by the National Science Foundation (ITR 0325428, HCC 0834847, and DRL 1108845). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES


Sidney K. D’Mello is an Assistant Professor in the departments of Computer Science and Psychology at the University of Notre Dame. His research interests include emotional processing, affective computing, artificial intelligence in education, human-computer interaction, speech recognition and natural language understanding, and computational models of human cognition. He has published over 100 journal papers, book chapters, and conference proceedings in these areas. D’Mello has edited two books on Affective Computing, is an associate editor for IEEE Transactions on Affective Computing, and serves as an advisory editor for the Journal of Educational Psychology. D’Mello received his PhD in Computer Science from the University of Memphis in 2009.

Arthur Graesser is Professor of Psychology, adjunct professor of Computer Science, and co-director of the Institute for Intelligent Systems at the University of Memphis. His specific interests include knowledge representation, question asking and answering, tutoring, text comprehension, inference generation, conversation, reading, education, memory, expert systems, artificial intelligence, and human-computer interaction. He is the current editor of *Journal of Educational Psychology*. In addition to publishing over 400 articles in journals, books, and conference proceedings, he has written two books and edited nine books. Graesser received his Ph.D. in psychology from the University of California at San Diego.