# When Does Disengagement Correlate with Performance in Spoken Dialog Computer Tutoring?

**Kate Forbes-Riley and Diane Litman,** *Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA 15260*
*forbesk,litman@cs.pitt.edu*

**Abstract.** In this paper we investigate how student disengagement relates to two performance metrics in a spoken dialog computer tutoring corpus, both when disengagement is measured through manual annotation by a trained human judge, and also when disengagement is measured through automatic annotation by the system based on a machine learning model. First, we investigate whether manually labeled overall disengagement and six different disengagement types are predictive of learning and user satisfaction in the corpus. Our results show that although students' percentage of overall disengaged turns negatively correlates both with the amount they learn and their user satisfaction, the individual types of disengagement correlate differently: some negatively correlate with learning and user satisfaction, while others don't correlate with either metric at all. Moreover, these relationships change somewhat depending on student prerequisite knowledge level. Furthermore, using multiple disengagement types to predict learning improves predictive power. Overall, these manual label-based results suggest that although adapting to disengagement should improve both student learning and user satisfaction in computer tutoring, maximizing performance requires the system to detect and respond differently based on disengagement type. Next, we present an approach to automatically detecting and responding to user disengagement types based on their differing correlations with correctness. Investigation of our machine learning model of user disengagement shows that its automatic labels negatively correlate with both performance metrics in the same way as the manual labels. The similarity of the correlations across the manual and automatic labels suggests that the automatic labels are a reasonable substitute for the manual labels. Moreover, the significant negative correlations themselves suggest that redesigning ITSPOKE to automatically detect and respond to disengagement has the potential to remediate disengagement and thereby improve performance, even in the presence of noise introduced by the automatic detection process.

**Keywords.** types of disengagement, learning, correlations, spoken dialog computer tutors, manual and automatic annotation, machine learning, natural language processing

## BACKGROUND AND PROBLEM

The last decade has seen a significant increase in computer tutoring research aimed at improving student learning and other performance metrics by tailoring system responses to changing student affect and attitudes, over and above correctness. Student (dis)engagement behaviors have been of particular interest in this research, including displays of gaming, boredom, indifference, (lack of) interest, (low) motivation, curiosity, and flow (e.g., (Forbes-Riley et al., 2011; Baker et al., 2008; Lehman et al., 2008; Porayska-Pomsta et al., 2008; de Vicente & Pain, 2002)). Correlational analyses of student (dis)engagement behaviors in tutoring system corpora have indicated that these behaviors are predictive of learning. For example, gaming (Baker et al., 2008; Aleven et al., 2004) and boredom (Lehman et al., 2008) have been associated with decreased learning during computer tutoring, while flow (Lehman et al., 2008) and engagement (Beck, 2005) have been associated with increased learning. In addition, a number of automatic gaming detectors have been implemented and evaluated in computer tutors, with results indicating that gaming behaviors can be reliably detected in real-time using features of the tutoring interaction (cf. (Baker et al., 2008)). Moreover, controlled experiments using gaming-adaptive computer tutors - i.e., tutors enhanced with interventions that target student gaming - have shown that adapting to gaming can improve student learning (Arroyo et al., 2007; Baker et al., 2008) or other performance metrics, such as reducing gaming (Walonoski & Heffernan, 2006; Aleven et al., 2004).

Our own research builds on this prior computer tutoring research, with the larger goal of enhancing our spoken dialog computer tutor to automatically detect and respond to student disengagement over and above correctness and uncertainty,[1] and thereby improve learning and other performance metrics. However, our work is novel in that it focuses on spoken language-based displays of disengagement. Moreover, in contrast to prior computer tutoring research, which has focused on detecting and adapting to only one disengagement behavior (typically gaming), our work addresses the problem of detecting and responding to a wider range of student disengagement behaviors, with the system interventions depending on the *type* of disengagement behavior detected. Working towards our end goal, in prior work we developed and evaluated an annotation scheme for manually labeling an overall measure of disengagement, as well as different types of disengagement behavior, in our spoken dialog computer tutoring corpora. The six types of disengagement behavior we empirically identify can be framed within a wider theoretical context. In particular, recent psychological analyses of a primary indicator of student disengagement, boredom, have not only shown the negative impact of boredom on academic performance, but have also elucidated primary precursors of boredom, with the larger goal of enabling teachers to eliminate boredom by targeting its different precursors (Pekrun et al., 2010; Acee et al., 2010; Daschmann et al., 2011). Our disengagement behavior types can be aligned with these theoretical precursors.

In this paper, we first focus on extending the results of others' prior work correlating disengagement behaviors and performance, targeting two performance metrics of primary interest in spoken dialogue computer tutors: student learning and user satisfaction. In particular, we investigate the range and performance relationships of student disengagement behaviors in a spoken dialog computer tutoring corpus. This corpus was previously manually labeled by trained human judges for the disengagement behaviors. We show that although our overall measure of manually labeled disengagement is predictive both of decreased student learning and of decreased user satisfaction, different types of manually labeled disengagement behaviors correlate *differently* with these performance metrics: some negatively correlate, while others don't correlate at all. Furthermore, the amount of prerequisite knowledge a student has

[1] As discussed elsewhere, our current system already adapts to student uncertainty over and above correctness; our goal is thus to enhance this system to adapt to multiple affective states (disengagement and uncertainty) (Forbes-Riley et al., 2011).

changes these relationships somewhat. We also show that using multiple disengagement types to predict learning improves predictive power. Importantly, these results suggest that while responding to an overall measure of student disengagement can improve both learning and user satisfaction in computer tutoring, maximizing performance requires the system to respond differently based on the type of disengagement behavior.

In the second part of this paper, we then address the problem of transitioning from human-labeled disengagement to the real world task of automatically detecting and responding to multiple student disengagement types during real-time spoken dialogue computer tutoring. In particular, we show that the noise introduced by automatic detection errors can be minimized by categorizing student disengagement types based on their differing correlations with correctness. Investigation of our previously developed machine learning model of student disengagement shows that its automatic labels correlate with both of our performance metrics in the same way as the manual labels. The similarity of the correlations across our manually and automatically detected disengagement labels suggests that the automatic labels are a reasonable substitute for the manual labels. Moreover, the significant negative correlations themselves suggest that our approach to automatically detecting and responding to disengagement has the potential to remediate disengagement and thereby improve performance, even in the presence of the noise introduced by the automatic detection process. We discuss how our automatic disengagement adaptations were developed based on these results and prior computer tutoring adaptations to student gaming, and conclude by summarizing our current progress in experimentally evaluating our spoken dialogue computer tutor that automatically detects and adapts to student disengagement.

## METHODOLOGY

Our research is performed on a corpus of spoken dialogs from a prior controlled experiment evaluating an uncertainty-adaptive version of our tutoring system, ITSPOKE (**I**ntelligent **T**utoring **SPOKE**n dialog system), which is a speech-enhanced and otherwise modified version of the Why2-Atlas qualitative physics tutor (VanLehn et al., 2002).

Briefly, ITSPOKE tutors 5 physics problems (one per dialog), using a Tutor Question - Student Answer - Tutor Response format. After each tutor question, the student speech is sent to the Sphinx2 recognizer, which yields an automatic transcript. This answer's (in)correctness is then automatically classified based on this transcript, using the TuTalk semantic analyzer (Jordan et al., 2007), and the answer's (un)certainty is automatically classified by inputting features of the speech signal, the automatic transcript, and the dialog context into a logistic regression model. The appropriate tutor response is determined based on the answer's (in)correctness and (un)certainty and then sent to the Cepstral text-to-speech system, whose audio output is played through the student headphones and is also displayed on a web-based interface. See Forbes-Riley & Litman (2011) for details.

### Data Collection

College students from the Pittsburgh area who had never taken college-level physics were recruited for the experiment using flyers posted on the University of Pittsburgh campus. Accepted subjects were

**Question**: A boy tosses a rock off a cliff with an initial velocity, Vi, in the horizontal direction. Assuming air resistance is negligible, what is true of the horizontal component of the velocity of the rock while it is falling?

Answer 1: it will increase

Answer 2: it will decrease

Answer 3: it will remain the same

Answer 4: there is not enough information to answer

**Question**: Suppose that a kangaroo maintains a constant horizontal velocity despite the fact that it runs by bouncing along. Suppose you are driving your LandRover and pull alongside a kangaroo that is bouncing along in a straight line. Just then you get a call on your cell phone. You maintain your speed, ignoring the kangaroo (which also ignores you and keeps bouncing along). At the end of your call, you look out the window, what should you see?

Answer 1: the kangaroo has pulled ahead of you

Answer 2: the kangaroo is bouncing along beside you

Answer 3: the kangaroo has fallen behind

Answer 4: other

**Question**: A motorcycle is driving west on a flat road at 10 m/s. A car is driving west down a hill sloped at 45 degrees. The western component of the car's velocity is 10 m/s. How does the horizontal displacement of the car compare to the horizontal displacement of the motorcycle at any time?

Answer 1: they are the same

Answer 2: the horizontal displacement of the car is greater than that of the motorcycle

Answer 3: the horizontal displacement of the motorcycle is greater that that of the car

Answer 4: there is not enough information to answer

Fig.1. ITSPOKE Knowledge Assessment Test Example Questions

required to be novices to both the tutoring domain and the technology application; i.e., they were required to have never before interacted with a physics dialogue tutoring system. They were also required to be native (American) English speakers to enable the highest performance by the spoken language processing components of ITSPOKE (which were trained on native American English speakers).[2]

The experimental procedure was as follows: students (1) read a short physics text, (2) took a multiple choice pretest, (3) worked 5 problems (dialogues) with ITSPOKE, (4) took a user satisfaction survey, and (5) took an isomorphic posttest. The entire procedure took approximately 3 hours, with 1-2 hours comprised of student-system interaction, depending on the subject's skill level and verbosity. The entire procedure was web-based (i.e., the reading and assessment instruments were all web-based as well as the system interaction). The resulting corpus contains 360 spoken dialogs (5 per student) from 72 students (6044 student turns), 47 female and 25 male. Figure 3 (discussed below) shows disengagement-annotated

[2] As a reviewer pointed out, subject factors such as cultural background and prior experience with the domain and technology can influence the types of disengagement observed.

corpus examples.

## Assessment Instruments

The knowledge assessment tests and user satisfaction survey are used to compute our performance metrics, and are the same as those used in multiple prior ITSPOKE experiments (c.f. (Forbes-Riley & Litman, 2011)). Our knowledge assessment tests include an isomorphic pretest and posttest, each containing 26 questions with 3-5 multiple choice options. Examples of these questions are shown in Figure 1. The tests were not counter-balanced across subjects during the experiment. These tests were originally developed by the Why2-Atlas system developers, who worked with physics teachers to author one or more multiple-choice questions for each main expectation and misconception covered in the tutoring dialogues, in order to probe them in different situations, gauge the generality of student learning and elicit situation-specific misconceptions (VanLehn et al., 2007). Average pretest and posttest scores in our corpus were 51.0% and 73.1% (out of 100%) with standard deviations of 14.5% and 13.8%, respectively.

User satisfaction refers to students' perceptions of system likability, ease of use, effectiveness, etc. Our user satisfaction survey contains 16 statements rated on a 5-point Likert scale. Average total survey score was 60.9 (out of 80), with a standard deviation of 8.5. As shown in Figure 2, 9 statements concern the tutoring domain (e.g., The tutor was effective), 7 of which were taken from Baylor et al. (2003) and 2 of which were created for our system. 3 statements concern user uncertainty and were created for our system. 4 statements concern the spoken dialogue interaction (e.g., It was easy to understand the tutor's speech) and were taken from Walker et al. (2002). Our survey has also been incorporated into other recent work exploring user satisfaction in spoken dialogue computer tutors (Dzikovska et al., 2011).

## The Disengagement Annotation Scheme

Our disengagement annotation scheme is empirically derived from observations in our data but draws on prior work, including appraisal theory-based emotion models, which also distinguish emotional behaviors from their underlying causes (e.g., (Conati & Maclaren, 2009))[3], as well as prior approaches to manually annotating disengagement or related states in tutoring corpora (Lehman et al., 2008; Porayska-Pomsta et al., 2008; de Vicente & Pain, 2002). Elsewhere we discuss the development and evaluation of our annotation scheme in detail (Forbes-Riley et al., 2011); here we summarize these results.

Our disengagement annotation scheme distinguishes seven labels: one overall Disengaged label, and six Disengagement Type labels. As noted above we took an empirical approach to developing and defining our disengagement labels. First we identified the range of *behavioral* evidence of student disengagement in our corpus, and then *contextual* evidence was used to distinguish different (inferred) underlying triggers or causes of the disengagement behaviors. Our labels are summarized below.

An **overall Disengaged label (DISE)** was used for all turns expressing moderate to strong disengagement behavior in the tutoring process, i.e., answers given without much effort or without caring about correctness. Answers might also be accompanied by signs of inattention, boredom, or irritation.

[3] Appraisal theories argue that one's appraisal of a situation causes emotion; i.e., emotions result from (and don't occur without) an evaluation of a context (e.g., (Conati & Maclaren, 2009)).

**SQ1**: It was easy to learn from the tutor.

**SQ2**: The tutor didn't interfere with my understanding of the content.

**SQ3**: The tutor believed I was knowledgeable.

**SQ4**: The tutor was useful.

**SQ5**: The tutor was effective on conveying ideas.

**SQ6**: The tutor was precise in providing advice.

**SQ7**: The tutor helped me to concentrate.

**SQ8**: The tutor responded effectively after I was incorrect about the answer to a question.

**SQ9**: The tutor responded effectively after I was correct about the answer to a question.

**SQ10**: The tutor responded effectively after I was uncertain about the answer to a question.

**SQ11**: The tutor responded effectively after I was certain about the answer to a question.

**SQ12**: The tutor's responses decreased my uncertainty about my understanding of the content.

**SQ13**: It was easy to understand the tutor speech.

**SQ14**: I knew what I could say or do at each point in the conversations with the tutor.

**SQ15**: The tutor worked the way I expected it to.

**SQ16**: Based on my experience using the tutor to learn physics, I would like to use such a tutor regularly.

ALMOST ALWAYS (5), OFTEN (4), SOMETIMES (3), RARELY (2), ALMOST NEVER (1)

Fig.2. ITSPOKE User Satisfaction Survey

Clear examples include answers spoken quickly in leaden monotone or with sarcastic or playful tones, or with off-task sounds such as rhythmic tapping or electronics usage.[4]

One of six **Disengagement Type labels** accompanied each DISE label. These types distinguish different student reactions to the system's limited natural language processing abilities (NLP-Distracted and NLP-Gaming), different student perceptions of the tutoring material (Easy, Hard and Presentation), and a "catch-all" category for other student reactions as the session progresses (Done).

**Hard**: Student lost interest because this tutor question was too hard (e.g., presupposes too much prior knowledge).

**Easy**: Student lost interest because this tutor question was too easy (e.g., a similar question was asked and answered earlier in the session).

**Presentation**: Student didn't pay attention to this tutor question because the turn presentation was

[4] Affective systems research has found that total disengagement is rare in laboratory settings (e.g., (Forbes-Riley et al., 2011; Lehman et al., 2008)). As in that research, we thus equate the "disengagement" label with either no or low engagement. Since total disengagement is common in real-world unobserved human-computer interactions (e.g., deleting unsatisfactory software), it remains an open question as to how well laboratory-based findings generalize.

too long and complex; his/her answer reflects unawareness of the fact that the tutor turn strongly hinted at the correct answer. This Type can be perceived as a subset of the Hard Type, but was reserved only for tutor turns of multiple sentences in length, with the idea that these particular tutor turns could be broken down into smaller parts when they trigger disengagement.

**NLP-Gaming**: Student didn't try to work out the answer to this tutor question; s/he instead deliberately gave a vague or incorrect answer or a guess to try and fool the system's limited natural language processing capabilities. This Type can be viewed as a partially overlapping subset of other Types (e.g., Hard/Easy), but was reserved only for student turns perceived as deliberate gaming instances (regardless of the underlying cause). This type represents a subset of the gaming behaviors addressed in prior (largely non-dialogue) work (see Background, above) which focuses on hint abuse and systematic guessing.[5] ITSPOKE does not provide hints upon request, and the dialog is the only recorded behavior, thus all detectable gaming behavior in ITSPOKE is linguistic.

**NLP-Distracted**: Student became distracted and hyperarticulated[6] this answer because the system misunderstood an immediately prior answer due to its limited natural language processing capabilities. This Disengagement Type can be seen as differing from the other Types in that although students do lose the tutoring flow, this is not of their own (un)conscious volition.

**Done**: Student just wants the interaction to be over (typically later in the dialogs) - s/he is bored, tired, and/or not interested in continuing at this moment (or no other label fits).

Figure 3 provides a corpus example for each Disengagement Type. In the first example, **Student**$_{12}$ is labeled Hard because the student gave up immediately and with irritation when too much prior knowledge was required (and the tutor turn was not overly long, nor was the student demonstrating NLP-Distracted or NLP-Gaming behavior). In the second example, **Student**$_{03}$ is labeled Easy because the student sang her answer and seemed wholly disinterested in its larger purpose in the dialogue (which was to prepare her to sum the net forces). In the third example, **Student**$_{08}$ is labeled Presentation because **ITSPOKE**$_{8}$ informs the student that the car is at rest vertically, and in the prior discussion the student demonstrated his understanding that when objects are at rest it means their net force is zero. Here, however, the student answers quickly and incorrectly, and was thus judged to have lost focus due to the tutor turn's length. In the fourth example, **Student**$_{09}$ is labeled NLP-Gaming because the student avoided giving a specific numerical value, offering instead a vague (and incorrect) answer. In the fifth example, **Student**$_{15}$ is labeled NLP-Distracted because the student hyperarticulates his answer with irritation after his answer in **Student**$_{14}$ is misrecognized. In the sixth example, **Student**$_{21}$ is labeled Done because the student answered 'No' semi-jokingly in regards to continuing the experiment.

This annotation scheme was developed over several corpus annotation rounds. The first round was performed by a single annotator. Two more rounds were performed by the first annotator and a second annotator who was trained on annotations from the first annotator. Each round was followed by discussion of the (dis)agreements between the two annotators and subsequent modification of the annotation

[5] This prior work defines gaming as attempting to succeed by exploiting the system rather than learning the material and using that knowledge to answer correctly (Baker et al., 2008).

[6] That is, gives the answer with unnatural pitch, cadence, stress, or loudness in an attempt to make the computer better understand him/her. For clarify, this label was renamed from "Language" in our prior work (Forbes-Riley et al., 2011); "Types" were also renamed from "Sources".

**ITSPOKE**$_{12}$: What is the definition of Newton's Second Law?

**STUDENT**$_{12}$: I have no idea <*sigh*> **(DISE: HARD)**

**ITSPOKE**$_{03}$: What vertical force is always exerted on an object near the surface of the earth?

**STUDENT**$_{03}$: graa-vi-tyyy <*sings answer*> **(DISE: EASY)**

**ITSPOKE**$_{08}$: Because the car is at rest in the vertical direction, the normal force on the car must have equal magnitude and opposite direction to the force of gravity on the car. Otherwise the car would not be at rest vertically. Now that we know the individual forces exerted on the car during the collision, we can discuss the net force. We'll discuss the vertical and horizontal components separately. First, is there a non zero net force on the car in the vertical direction?

**STUDENT**$_{08}$: yes **(DISE: PRESENTATION)**

**ITSPOKE**$_{09}$: What's the numerical value of the man's acceleration? Please specify the units too.

**STUDENT**$_{09}$: The speed of the elevator. Meters per second. **(DISE: NLP-GAMING)**

**ITSPOKE**$_{14}$: What's the numerical value of the keys' acceleration? Please specify the units too.

**STUDENT**$_{14}$: nine point eight meters per second per second *(but ITSPOKE heard: nine point eight need there's per second per second)*

**ITSPOKE**$_{15}$: Well... The problem statement says that the man and his keys are in freefall. All objects in freefall have the same acceleration, due to the force of gravity acting upon them. It's called Gravitational Acceleration. As we discussed earlier, what's its magnitude? Please specify the units too.

**STUDENT**$_{15}$: nine point eight meters per second per second! <*hyperarticulated*> **(DISE: NLP-DISTRACTED)**

**ITSPOKE**$_{21}$: Based on our discussion, we conclude that the keys will remain in front of the man's face during the entire fall. [...] Would you like to do another problem?

**STUDENT**$_{21}$: No <*laugh*> **(DISE: DONE)**

Fig.3. Corpus Examples Illustrating Disengagement Annotation Scheme.

scheme based on these discussions. During each round, each student turn was either labeled as Disengaged with a single Disengagement Type, or was otherwise labeled non-Disengaged by default. When multiple Type labels were possible, the most specifically applicable label was selected. The goal of the Type labeling was to be as specific as possible at associating student disengagement behaviors with system triggers (to enable highly targeted system adaptations), while at the same time allowing different disengagement behaviors to possibly trace back to a common underlying cause. For example, as discussed in prior computer tutoring research (see Background, above), students may game the system for multiple reasons, including because the material is too hard, or too easy, or they lack motivation.

Our final inter-annotator reliability evaluation between the two trained annotators on a new corpus subset (393 student turns) showed that our overall Disengagement label (0.55 Kappa) and Disengagement Type labels (0.43 Kappa) can be annotated with moderate reliability.[7] Note however that our Done (catch-all) label was the most frequently occurring type in this reliability study, and can likely be further broken down (see (Forbes-Riley et al., 2011) for discussion). Note also that our Disengagement Kappas

[7] Although interpreting Kappa values is somewhat controversial and depends on the application, we find the Landis & Koch (1977) standard to be a useful guideline: 0.21-0.40 = "Fair"; 0.41-0.60 = "Moderate"; 0.61-0.80 = "Substantial"; 0.81-1.00 = "Almost Perfect".

are on par with prior emotion annotation work. For example, two studies that have compared self-reports, peer labelers, trained labelers, and combinations of labelers (Afzal & Robinson, 2011; D'Mello et al., 2008) both illustrate the common finding that human annotators display low to moderate interannotator reliability for affect annotation, and both studies further show that trained labelers yield the highest reliability on this task. Despite the lack of high interannotator reliability, responding to affect detected both by human labelers, and by automatic detectors trained on human labels, has repeatedly been shown to improve system performance (see Background, above).

After the inter-annotator reliability evaluation, all of the student turns in the corpus were manually annotated by the first trained annotator using the final Disengagement Annotation Scheme shown above. Note that although it was developed by observation of our ITSPOKE corpus, this scheme should generalize to other learning environments, including analogs of NLP-Gaming and NLP-Distracted in non-dialogue based systems, since these two types represent two disengagement behaviors triggered by the system's inherent interaction processing inflexibility, which exists regardless of the communication medium. Altogether our Disengagement Types label a range of student behaviors and complementary affective displays associated with disengagement, including off-task, bored, and frustration displays, as well as low-motivated actions that don't attempt to exploit the system. Moreover, our labels capture the fact that these behaviors can be associated with different underlying causes. For example, a student who disengages because a question is too hard may exhibit any of these behaviors. Note finally that this turn-level annotation scheme captures both fleeting disengagement states as well as long-term disengagement escalation across turns.

## Wider Theoretical Context of Disengagement Types

A number of psychological theories concerning a primary disengagement behavior, boredom, have emerged in recent years. The APA Dictionary of Psychology defines boredom as "a state of weariness or ennui resulting from a lack of engagement with stimuli in the environment" (Acee et al., 2010, p.17). Pekrun et al. (2010), Acee et al. (2010), and Daschmann et al. (2011) provide comprehensive summaries of this prior research and propose theoretical advances based on empirical studies of boredom, which help elucidate the primary precursors, or causes, of boredom, from the point of view of their different theoretical perspectives.

Pekrun et al. (2010) provide a detailed theoretical analysis of boredom that distinguishes overlapping terms (e.g. lack of interest) and defines boredom as an "achievement emotion", describing its effects on academic performance in terms of two achievement-related determinants: subjective task value appraisal and control appraisal. In this analysis, the extent to which students subjectively value the learning material and perceive control over the situation can facilitate the occurrence of different emotions, including boredom. Control-value theory posits a curvilinear relationship between control and boredom, with more boredom being experienced under conditions of high or low control, as compared with moderate control. Control-value theory further proposes that a lack of perceived intrinsic value of achievement activities, rather than a lack of extrinsic, instrumental utility, is critical for the instigation of boredom. This theoretical analysis is validated across 5 empirical studies using boredom self-reporting in different cultural and learning settings. The studies all found that perceived levels of control and task value negatively predicted boredom. Moreover, boredom was related to increased attention problems and decreased motivation, effort, use of elaboration strategies, self-regulation, and performance.

Acee et al. (2010) provide a clearer picture of the dimensionality and situational dependency of boredom, by investigating how students relate their feelings of boredom to under-challenging and over-challenging situations and whether they differentiate task-focused and self-focused boredom in these situations. Their results suggest that students were not differentiating between task and self-focused boredom in situations they recalled as being under-challenging, but were differentiating task- and self-focused boredom in remembered over-challenging situations. Moreover, task-focused boredom was characterized by the tediousness and meaninglessness of a task, while self-focused boredom was characterized by feeling dissatisfied and frustrated.

Daschmann et al. (2011) synthesize eight distinct precursors of boredom identified in these and other prior studies and empirically validate their usefulness for predicting student boredom during academic activities: 1) monotony, 2) lack of meaning, 3) opportunity costs, 4) being over-challenged, 5) being under-challenged, 6) lack of student involvement, 7) teacher dislike, 8) generalized boredom. These results support their hypothesis that student boredom in academic settings can be due to distinguishable precursors. Note that Acee et al. (2010)'s self-focused boredom appears to correspond to "generalized boredom", while their task-focused boredom appears to correspond to "monotony" and "lack of meaning". Pekrun et al. (2010)'s control and task value correspond to "lack of student involvement" and "opportunity costs", respectively.

Although our own six types of disengagement are behaviorally driven, by first observing different student disengagement behaviors and then inferring their underlying triggers or causes, it is useful to situate our six types within this wider theoretical context. Indeed there is substantial concordance between our behaviorally driven types and Daschmann et al. (2011)'s eight theoretical precursors. In particular, our Hard and Easy types largely correspond to the "over-challenging" and "under-challenging" precursors. Prior tutoring research (see Background, above) suggests that our NLP-Gaming type could potentially overlap with the any of the eight isolated precursors, though often gaming is triggered by challenge-related, control-related, or task value-related causes. Our Presentation type aligns with the "monotony" precursor, in that students disengage because they find the presentation of the material in a tutor turn too monotonous. Our Done type appears to largely correspond to the "generalized boredom" precursor, in that it accounts for habitual disengagement, and disengagement that is not due to students' perception of the tutoring but rather to their personality. Finally, our NLP-Distracted type appears to be most closely linked to the "teacher dislike" precursor, in that the disengagement is due to perceived flaws in the quality of the student-tutor natural language interaction.

## RESULTS: MANUALLY LABELED DISENGAGEMENT

To investigate how our overall disengagement manual labels relate to performance in our corpus, we first computed the percentage of each label's occurrence (total label occurrences/total turns) for each student. We then computed correlations between each of these percentages and our two performance metrics: student learning and user satisfaction. Finally, we used stepwise linear regressions to examine the relative usefulness of the disengagement metrics in more complex performance models.

## Correlation Results for Manually Labeled Disengagement

To measure the relationship between student disengagement and learning, we compute the partial Pearson's correlation between each disengagement percentage and posttest score, controlling for pretest to account for learning gain. Table 1 shows first the mean percentage (Mn%) and its standard deviation (sd) over all students, the Pearson's Correlation coefficient (R) and significance (p) with significant results bolded ($p \leq 0.05$), and the total number of occurrences (Tot) for each label in the entire dataset. These statistics are also provided for students with low and high pretest scores (see below). The last two rows show test scores for each group (Mn% and sd).

Table 1

Partial Correlation Results between Manually Labeled Percent Disengagement or Disengagement Types and Posttest Controlled for Pretest in the ITSPOKE Corpus (N=72; Low Pretests N=40; High Pretests N=32)

| | All Students | | | Low Pretests | | | High Pretests | | |
|---|---|---|---|---|---|---|---|---|---|
| Measure | Mn%(sd) | R(p) | Tot | Mn%(sd) | R(p) | Tot | Mn%(sd) | R(p) | Tot |
| **DISE** | 14.5(8.2) | **-.33(.01)** | 886 | 16.2(8.3) | **-.37(.02)** | 555 | 12.2(7.7) | -.26(.15) | 331 |
| NLPDistracted | 0.4(1.4) | -.03(.78) | 28 | 0.6(1.8) | -.07(.68) | 22 | 0.2(0.8) | .04(.81) | 6 |
| **Hard** | 2.8(2.9) | **-.36(.01)** | 172 | 3.6(3.3) | **-.35(.03)** | 124 | 1.7(2.0) | **-.46(.01)** | 48 |
| **NLPGaming** | 3.0(3.0) | **-.34(.01)** | 186 | 3.2(2.8) | **-.31(.05)** | 108 | 2.9(3.2) | **-.39(.03)** | 78 |
| Easy | 1.4(2.6) | .12(.33) | 83 | 1.1(2.0) | -.02(.92) | 36 | 1.8(3.2) | .30(.11) | 47 |
| **Presentation** | 3.0(2.2) | **-.27(.02)** | 182 | 3.6(2.1) | -.22(.17) | 124 | 2.1(2.0) | **-.35(.05)** | 58 |
| Done | 3.9(3.2) | -.08(.52) | 235 | 4.2(3.2) | -.11(.53) | 141 | 3.5(3.3) | -.04(.85) | 94 |
| Pretest | 51.0(14.5) | | | 40.5(7.8) | | | 64.1(9.2) | | |
| Posttest | 73.1(13.8) | | | 66.9(12.8) | | | 80.8(10.9) | | |

Considering the results over all students, comparison of means shows that of the 14.5% overall disengaged turns on average per student, Done is the most frequent type of disengagement, followed by NLP-Gaming and Presentation, Hard, Easy, and NLP-Distracted. Since Done is defined as a "catch-all" category, it is not surprising that it is the most frequent; that it occurs only slightly more than three of the other types suggests that our six categories are reasonably representative of the range of disengagement behaviors (and underlying causes) in our data. The high standard deviations suggest that the amount of overall DISE, and the disengagement types, are highly student-dependent.

The correlation results over all students show that overall DISE is significantly correlated with decreased learning. This supports prior work showing negative relationships between learning and boredom or gaming (see Background, above). Our results also show significant negative correlations between learning and the Hard, NLP-Gaming, and Presentation Types. Prior work suggests that gaming behaviors associated with poorer learning often occur when students lack the knowledge to answer the question (Baker et al., 2008; Arroyo et al., 2007).[8] Similarly, we hypothesize that students often exhibited linguistic (NLP) gaming in our corpus because the system's limited natural language processing abilities prevented them from eliciting information they needed to answer the question. Together, the results for

[8] Other suggested reasons for gaming in this prior work include a performance-based mentality (as opposed to learning-based) and low motivation to learn.

the NLP-Gaming and Hard Types suggest that if not remediated, disengagement can negatively impact learning when caused by questions presupposing knowledge the student doesn't have. Relatedly, the negative Presentation correlation suggests that if not remediated, disengagement can also negatively impact learning when caused by the inflexibility of the system's half of the dialog.

There are no significant correlations over all students for the NLP-Distracted, Easy, or Done Types. These null results may reflect the fact that not all student disengagement during tutoring is negatively related to learning. For example, some students may get distracted and irritated by system misunderstandings (NLP-Distracted Type) irrespective of how much they learn. Similarly, some students may temporarily lose interest when a tutor question is too easy (Easy Type), irrespective of how much they learn. Moreover, some temporary losses of student engagement that occur as the tutoring dialog or session nears its end (or for other murkier reasons such as personality) (Done Type) may occur irrespective of learning. However, further analysis is needed before such conclusions can be drawn from these null results. For example, the NLP-Distracted and Easy Types were rare in our corpus, and more data may reveal a stronger relationship to learning. Because the Done Type is a "catch-all" category, it may contain sub-types with different relationships to learning.

To further investigate how students' prerequisite knowledge level changes the relationship between disengagement behavior and learning in our data, we split students into high (N=32) and low (N=40) groups based on their mean pretest score,[9] and then reran the partial correlations on each group individually. Note that this approach substantially reduces the sample size, so the results can only be taken as suggestive, but it also affords a finer-grained view within groups with high and low domain knowledge of how prerequisite knowledge level changes the partial correlations between disengagement and posttest.[10] Comparison of means in Table 1 shows similar relative frequencies of the types across both groups. Done, Presentation and NLP-Gaming occur most often, and NLP-Distracted least often. Not surprisingly, the frequencies of Hard and Easy differ depending on knowledge level. However, additional one-way ANOVAs suggest that only overall DISE, Hard, and Presentation differ significantly across the two groups ($p<.05$), occurring more for low pretesters.

Regarding the correlations, neither group patterned exactly like the combined group. The low pretest group did not show the significant negative correlation between learning and Presentation, while the high pretest group did not show it for overall DISE. However, further research with larger sample sizes are necessary before conclusions can be drawn from these null results. The most substantial difference between high and low pretesters in Table 1 pertains to the correlations between Easy and learning for high and low knowledge students, respectively. In this case, both the correlation coefficients and significance values are substantially different across the two groups. This result suggests that the proportion of easy questions a subject receives is more likely to be a signal of the amount s/he will learn for subjects with a high level of prerequisite knowledge. Finally, our results suggest that NLP-Gaming reflects decreased

[9] We didn't use a median split because it placed the same score in both groups. A T-test showed the two groups represent different populations ($p<.001$). Also note that while a repeated test-measure ANOVA has indicated that all students learned during the tutoring ($F(1,69) = 225.688$, $p<0.001$) (Forbes-Riley & Litman, 2011), a one-way ANOVA showed no difference in normalized learning gain between the high and low pretest groups.

[10] Moreover, it enables consistency with the results in Table 2, which compare high and low pretesters using a bivariate correlation between disengagement and user satisfaction scores.

learning regardless of prerequisite knowledge. In contrast, prior work suggests that gaming behaviors in highly knowledgeable students may have little relation to learning, while the same behavior in students with low prerequisite knowledge is associated with poorer learning (Baker et al., 2008). The difference may be due in part to the fact that prior work focused on hint abuse and systematic guessing, which are gaming methods targeted at manipulating the system into giving the correct answer. In contrast, ITSPOKE students can't predict beforehand whether (linguistic) NLP-Gaming will result in the correct answer.

Table 2
Bivariate Correlation Results between Manually Labeled Percent Disengagement or Disengagement Types and User Satisfaction in the ITSPOKE Corpus (N=72; Low Pretests N=40; High Pretests N=32)

| | All Students | Low Pretests | High Pretests |
|---|---|---|---|
| Measure | R(p) | R(p) | R(p) |
| **DISE** | -.20(.10) | **-.36(.02)** | .25(.18) |
| NLPDistracted | -.21(.08) | -.29(.07) | .22(.23) |
| **Hard** | **-.38(.01)** | **-.44(.01)** | -.09(.63) |
| NLPGaming | .00(.98) | -.08(.64) | .15(.40) |
| Easy | .14(.24) | .02(.92) | .31(.09) |
| Presentation | -.08(.51) | -.10(.54) | .10(.59) |
| Done | -.12(.32) | -.20(.22) | .07(.71) |
| User Satisfaction Total | 60.9(8.5) | 59.9(10.1) | 62.2(6.1) |

To measure the relationship between student disengagement and user satisfaction, we computed the bivariate Pearson's correlation between each disengagement percentage and total user satisfaction survey score (out of a possible score of 80). Table 2 shows these results. Considering our overall DISE label and the NLP-Distracted type, the correlation results show that user satisfaction patterns similarly to learning. These disengagement metrics are negatively correlated to decreased user satisfaction (significantly or as a trend) over all students and within the low pretest students, but not within the high pretest students. The difference between high and low pretesters is most substantial for the Hard and Easy types of disengagement. In these cases, both the correlation coefficients and significance values are substantially different across the two groups. These differences suggest that the proportion of easy questions received is more likely to signal how satisfied a student will be with the tutoring experience if s/he has a high level of prerequisite knowledge. Similarly, the proportion of hard questions received is more likely to signal how dissatisfied a student will be with the tutoring experience if s/he has a low level of prerequisite knowledge. However, further research is necessary before we can conclude that these relationships between our disengagement metrics and user satisfaction differ substantially depending on prerequisite knowledge level.

## Regression Results for Manually Labeled Disengagement

Having examined how each disengagement metric predicts performance in isolation, we next investigated their relative usefulness in more complex performance models using stepwise linear regression. To model learning, we predicted posttest, allowing the model to select its inputs from pretest and our seven disengagement metrics (overall disengagement and the six types).

The following model yielded the best significant training fit to our data ($R^2$=.49, p<.001). As shown, two disengagement types were incorporated along with pretest. The (standardized) feature weights indicate relative predictive power in accounting for posttest variance. As shown, the Hard Type (p<.01) is more predictive of decreased posttest than the Presentation Type (p=.03), but both work together to significantly increase the model's predictive power over pretest alone. Note that although NLPGaming was the second strongest correlation in isolation (after Hard and before Presentation, Table 1), it is not selected in the equation after Hard has accounted for its proportion of the posttest variance.

**Posttest** = .41***Pretest** - .28***%Hard** - .21***%Presentation**

To model user satisfaction, we predicted total survey score, allowing the model to select its inputs from the seven disengagement metrics. The following model yielded the best significant training fit to our data ($R^2$=.15, p=.001). Unlike the learning model, here we see that decreased user satisfaction is best predicted by the Hard Type in isolation, rather than by a combination of disengagement metrics.

**User Satisfaction Score** = - .38***%Hard**

Note finally that our goal for all the regression analyses presented in this paper is not to produce absolute best models of performance with maximized $R^2$; our goal here is rather to examine the relative usefulness of specific metrics for predicting performance. In similar types of stepwise regressions on prior ITSPOKE corpora, we've shown that more complete models of system performance incorporating many predictors of student learning (i.e., affective states in conjunction with other dialogue tutoring features) can yield $R^2$ values of over .5 (Forbes-Riley et al., 2008). Interestingly, however, even with such combinations of features, we still find that our models of user satisfaction are much less powerful than our learning models (Forbes-Riley & Litman, 2006).

## RESULTS: AUTOMATICALLY LABELED DISENGAGEMENT

There has been substantial prior work focused on detecting disengagement behaviors during human-computer interactions across domains. In particular, as discussed above (see Background), a number of automatic gaming detectors have been developed and evaluated in computer tutors, with results indicating that gaming behaviors can be reliably detected in real-time using various features of the tutoring interaction, such as correctness and difficulty-based features drawn from the tutoring logs (cf. (Baker et al., 2008)). Researchers have also modeled user disengagement levels using other more generic interaction features, including eye gaze and turn-taking features in human-robot interactions (Sidner & Lee, 2003), manually coded spoken dialogue acts based on the SWBDL-DAMSL scheme (Stolcke et al., 2000) during human interactions with an embodied medical agent (Martalo et al., 2008), facial sensors during dynamic, multi-party dialogues in open-world settings (Bohus & Horvitz, 2009), as well as a wide variety of acoustic-prosodic, lexical and contextual features in spoken dialogue interactions (Schuller et al., 2010; Wang & Hirschberg, 2011; Jeon et al., 2010). Most of this research focuses on detecting only one specific disengagement behavior (gaming) or a few generic disengagement levels (e.g., high, low), because it is well-known that automatic detection errors multiply with every additional class distinguished.

Our own binary disengagement (DISE) detector was built using features and methods similar to prior work. In particular, we used WEKA machine learning software and 10-fold cross-validation with the J48 decision tree algorithm with a variety of features extracted from the turns in our corpus, including linguistic features (e.g., acoustic-prosodic, lexical and dialog) previously used to predict affect in speech (cf. (Forbes-Riley & Litman, 2011)), and system-specific features (e.g., correctness, timing, knowledge level, and question difficulty) previously used to predict gaming (e.g., (Baker et al., 2008; Arroyo et al., 2007; Walonoski & Heffernan, 2006; Beck, 2005)). Details of our machine learning experiments are described elsewhere (Forbes-Riley & Litman, 2012b); here we summarize our results.

First, the cross-validation evaluation on the binary overall disengagement (DISE) label yielded the averaged results shown in Table 3. As shown, in addition to accuracy, we use Unweighted Average (UA) Precision[11], Recall, and F-measure because they are the standard measures used to evaluate current affect recognition technology, particularly for unbalanced two-class problems (Schuller et al., 2009). Our results are on par with the best results of the other prior research. For example, Martalo et al. (2008) report average precision of 75% and recall of 74% (detecting three levels of disengagement), while Kapoor & Picard (2005) report an accuracy of 86% for detecting binary (dis)interest. For comparison, majority class baseline performance is also shown (i.e., always predicting the class that occurs most frequently, "nonDISE").

Table 3
Forbes-Riley & Litman (2012b) Results of 10-fold Cross-Validation Experiment on ITSPOKE Corpus Detecting the Binary DISE Label

| Algorithm | Accuracy | UA Precision | UA Recall | UA Fmeasure |
|---|---|---|---|---|
| Decision Tree | 83.1% | 68.9% | 68.7% | 68.8% |
| Majority Label | 83.8% | 41.9% | 50.0% | 45.6% |

Although our automatic disengagement detector predicts the presence or absence of overall disengagement with sufficient performance, a variety of further machine learning experiments showed that all metrics degraded unacceptably when we automatically distinguish more than two disengagement classes. As illustration, Table 4 shows the results of one machine learning experiment predicting the six Disengagement Types. This experiment used 10-fold cross-validation on a corpus subset containing *only* manually labeled disengaged turns (1170 turns). In other words, perfect binary automatic disengagement classification was assumed, with only student turns already labeled DISE being sent to the six-way Disengagement Type classifier. This experiment yielded an average accuracy of 49.1% at predicting the six Disengagement Types, with the individual precision, recall and Fmeasure results for each Type shown in Table 4.

These results represent the best possible case. Since our binary automatic DISE detector does not yield perfect binary DISE classification (Table 3), we expect that the actual performance of the six-way DISE Type detector would be substantially lower for all DISE Types. Because our prior system eval-

[11] simply ((Precision(DISE) + Precision(nonDISE))/2)

Table 4
Results of 10-fold Cross-Validation Pilot Experiment Distinguishing the Six Disengagement Types on 1170 Manually Labeled DISE Turns in ITSPOKE Corpus with Perfect Automatic DISE Detection Assumed

| DISE TYPE | Precision | Recall | Fmeasure |
|---|---|---|---|
| Presentation | 55.0% | 60.9% | 57.8% |
| Done | 46.9% | 55.5% | 50.8% |
| NLPGaming | 43.7% | 47.0% | 45.3% |
| Hard | 56.5% | 37.4% | 45.0% |
| Easy | 52.5% | 43.1% | 47.3% |
| NLPDistracted | 41.2% | 32.9% | 36.6% |

uations have shown that low precision and recall for automatic affect detection can have a significant negative impact on global system performance, by substantially decreasing the number of true positive affect instances receiving the adaptation (Forbes-Riley & Litman, 2011), we determined that proceeding with highly error-prone automatic labeling of our six Disengagement Types was unlikely to achieve our goal of producing an effective computer tutor that automatically detects and adapts to multiple disengagement types.

## Correlation Results for Automatically Labeled Disengagement

Next we sought evidence beyond our intrinsic cross-validation evaluation that our automatic (binary) DISE detector would be both useful and a reasonable substitute for our manual labels in our computer tutor. In particular, having shown in Tables 1-2 that the manual disengagement labels are negatively correlated with two measures of system performance, we next sought to verify the adequacy of our current level of automatic disengagement detection by demonstrating that even after replacing the manual disengagement labels with the automatic disengagement labels, we still see similar negative correlations with our two performance metrics.

Table 5 compares the correlations of our two performance metrics with our automatic disengagement labels to the correlations of those same metrics with our manual disengagement labels. As shown, our automatic disengagement labels are significantly related to performance in the same way as our manual labels, regardless of whether we measure performance as user satisfaction or learning gain. Moreover, in both cases the correlations are nearly identical. While our automatic disengagement detector has not yet been experimentally evaluated, the similarity of these correlations across our manually and automatically detected disengagement labels suggests that the automatic labels are a reasonable substitute for the manual labels. Moreover, the significant negative correlations themselves suggest that redesigning ITSPOKE to automatically detect and respond to disengagement has the potential to remediate disengagement and thereby improve performance, even in the presence of the noise introduced by the automatic detection process.[12]

[12] Although correlations do not necessary reflect causal relationships, spoken dialogue research has shown that redesigning a system in light of such correlational analysis can indeed yield performance improvements (Rotaru & Litman, 2009).

Table 5
Correlation Results between Automatic Disengagement and Two Performance Metrics in the ITSPOKE Corpus (N=72)

| Measure | Mn%(sd) | Total | Learning: R(p) | User Satisfaction: R(p) |
|---|---|---|---|---|
| **% Automatic DISE** | 14.8(8.1) | 908 | -.30(.01) | -.20(.09) |
| **% Manual DISE** | 14.5(8.2) | 886 | -.33(.01) | -.20(.10) |

## RESULTS: DEVELOPING AUTOMATIC ADAPTATIONS FOR DISENGAGEMENT

Despite the fact that our automatic disengagement detector could not distinguish our six different disengagement types with reasonable accuracy, further correlational analyses reveals that it was nevertheless possible for our computer tutor to respond differently to a subset of our six disengagement types. In particular, correlational analyses showed that our six manually labeled disengagement types are themselves correlated with correctness, which itself is a performance metric in computer tutoring systems. As shown in Table 6, the manually labeled disengagement types can be grouped into two classes: those which significantly negatively correlate with correctness (bolded), and those which do not (unbolded). The table further suggests that low and high pretest subjects behave very much the same with respect to all of these correlations. Finally, further analysis of our corpus showed that 98% of the negatively correlated disengaged turns were incorrect, while 77% of the non-correlated disengaged turns were correct.

Table 6
Correlation Results between Manually Labeled Percent Disengagement Types and Percent Correctness in the ITSPOKE Corpus (N=72; Low Pretests N=40; High Pretests N=32)

| | All Students | Low Pretests | High Pretests |
|---|---|---|---|
| Measure | R(p) | R(p) | R(p) |
| DISE | **-.54(.01)** | **-.51(.01)** | **-.51(.01)** |
| NLPDistracted | -.02(.84) | .09(.58) | -.12(.50) |
| Hard | **-.58(.01)** | **-.64(.01)** | **-.43(.01)** |
| NLPGaming | **-.55(.01)** | **-.45(.01)** | **-.66(.01)** |
| Easy | .28(.02) | .22(.18) | .27(.13) |
| Presentation | **-.57(.01)** | **-.45(.01)** | **-.60(.01)** |
| Done | -.18(.13) | -.15(.36) | -.16(.39) |
| Correctness | 72.4(9.4) | 69.6(8.0) | 76.0(9.9) |

Based on these results, we decided to develop two automatic disengagement adaptations for our computer tutor: one for disengaged+incorrect turns and another for disengaged+correct turns. In this way, our automatic disengagement detector could focus on predicting only the binary overall disengagement label, thus maximizing its accuracy, while our system still could better maximize student performance by responding differently to different types of disengagement behaviors. Our disengagement adaptations are described in detail elsewhere (Forbes-Riley & Litman, 2012a); here we summarize their development.

First, we developed a substantive system response for disengaged+incorrect answers, which correspond almost entirely to the Hard, NLPGaming, and Presentation types. The substantive response was intended to remediate the negative learning correlation and target learning improvement. Second, we developed a minimal, non-invasive system response for disengaged+correct answers, which correspond largely to the NLPDistracted, Easy, and Done types. This minimal response was intended to reduce disengagement without upsetting the existing learning balance (since no negative correlation was demonstrated). In this way (assuming the same proportions as our training corpus), all of the disengaged turns would receive some disengagement adaptation, and 98% of the negatively correlated disengaged turns would receive substantive adaptation targeting learning improvement. Although 23% of the non-correlated turns would also receive this substantive adaptation, we hypothesized that because they were incorrect the more substantive adaptation wouldn't have a detrimental impact on learning in these cases.

Our disengagement adaptations build on prior evaluations of gaming adaptations in computer tutors that involved preventing gaming (e.g., (Walonoski & Heffernan, 2006; Beck, 2005; Murray & vanLehn, 2005; Aleven et al., 2004)), metacognitive feedback about better ways to learn (Arroyo et al., 2007; Walonoski & Heffernan, 2006; Aleven et al., 2004), easier exercises focusing on the gamed material (Baker et al., 2008), and performance feedback reminding students of task value (Arroyo et al., 2007; Walonoski & Heffernan, 2006). In particular, our current results suggest that disengaged+incorrect turns require more substantial intervention, because they negatively correlate with learning and involve a lack of understanding of the tutor question. Our disengaged+incorrect system response thus builds on the prior finding that supplementary information can help reduce some types of disengagement for highly disengaged users (Baker et al., 2006). We hypothesized that ITSPOKE's existing response to incorrectness (a Bottom Out or Remediation Subdialogue) was insufficient for a disengaged+incorrect turn because the user had already disengaged. To benefit from this supplementary knowledge, the user first had to reengage. Thus, our system would respond with "productive interaction feedback"[13] followed by an easier "fill in the blank" version of the original question. The purpose of this two-pronged response is to regain the user's attention with the feedback and then provide a path through the learning impasse with the easier question, thereby keeping the user engaged. An example is shown in Figure 4, where **STUDENT-1** was manually labeled disengaged+incorrect because the student gives an irrelevant (and obviously incorrect) answer.

Our current results suggest that disengaged+correct turns should receive minimal, non-invasive interventions because they display no negative correlation with learning (at least at current levels). Thus their adaptation should aim to reduce disengagement without upsetting any learning that might already be occurring. Our disengaged+correct system response builds on the prior findings in the computer tutoring literature that progress reports and productive learning tips can positively impact multiple performance metrics when used without specifically targeting disengagement (Arroyo et al., 2007), but not when used after every user turn (Walonoski & Heffernan, 2006). We hypothesized that these responses might be most beneficial if they targeted disengaged turns that were already correct, by reminding these students of the larger interaction goal. Thus, our system would respond with "productive interaction feedback" followed by a progress report graphing the user's correctness both in the current dialogue and over all prior dialogues. Examples are shown in Figure 5, where **STUDENT-1** is labeled disengaged+correct

[13] This is our generalization of the concept of "productive learning tip" used in prior work (Arroyo et al., 2007).

because the user unnecessarily repeats himself, signaling his lack of interest. As shown, we distinguish two classes of productive interaction feedback. That in “2a” shows the feedback given when the progress report indicates improvement on the current dialogue relative to the prior ones, while “2b” shows the feedback given when there is a decline.

Finally, note that although we do not directly adapt to every disengagement type, our productive learning tips were authored with reference to the relevant types (e.g., ’That was an easy one to make sure you’re paying attention!’ refers to the Easy Type), and thus we are still targeting these types for adaptation at a more global level. However, while globally targeting disengagement behaviors is likely to yield some improvement within the individual disengagement types, accurately detecting and adapting to all the different dimensions of student disengagement would likely yield the greatest performance improvement during computer tutoring. For example, as the psychological studies of the theoretical precursors of boredom discussed above suggest, the disengagement interventions we present here are unlikely to be maximally effective at remediating disengagement that steeps from learners appraising the task as having low value.

---

**ITSPOKE-1**: Let’s begin by looking at the motion of the man and his keys while he’s holding them. How does his velocity compare to that of his keys?

**STUDENT-1: vertical (INCORRECT, DISENGAGED)**

**ITSPOKE-2**: That doesn’t sound right. Let’s think about this a little more. *[ Productive Interaction Feedback]* Since the man is holding his keys, they aren’t moving relative to each other. So their velocities must be WHAT? *[Supplementary Question].*

---

Fig.4. Dialogue Example Showing Adaptation for Disengaged+Incorrect Turns

---

**ITSPOKE-1**: Let’s begin by looking at the motion of the man and his keys while he’s holding them. How does his velocity compare to that of his keys?

**STUDENT-1: same same same (CORRECT, DISENGAGED)**

..............................

**ITSPOKE-2a**: Well done. Here’s a quick progress update. Good effort so far!!*[Productive Interaction Feedback When Progress Report Shows Improvement]* Now let’s see what happens after the man releases his keys. [...] *[Next Topic Question]*

..............................

**ITSPOKE-2b**: Right. Here’s a quick progress update. It might help to remember we will build on the topics we’re discussing now.*[Productive Interaction Feedback When Progress Report Shows Decline]* Now let’s see what happens after the man releases his keys. [...] *[Next Topic Question]*

---

Fig.5. Dialogue Example Showing Adaptation for Disengaged+Correct Turns

## CONCLUSIONS AND CURRENT DIRECTIONS

In this paper, we extended prior research by investigating how overall student disengagement (DISE) and its subtypes relate to two metrics of performance in spoken dialog computer tutoring. In the first part of the paper, we investigate the range and performance relationships of different manually labeled student disengagement behaviors. We showed that overall disengagement negatively correlates with student learning, as do the Hard, Presentation, and NLP-Gaming Types, but the NLP-Distracted, Easy and Done Types do not. While overall disengagement also negatively correlates with user satisfaction, only the Hard and NLP-Distracted Types also do. We further showed that prerequisite knowledge level changes these relationships. For learning, only high pretesters exhibit the Presentation correlation, while only low pretesters exhibit the overall DISE correlation. For user satisfaction only the low pretesters mirror the correlations over all students; high pretesters show only a positive correlation with the Easy type (as a trend). We then showed that using both the Hard and Presentation Types improves predictive power for modeling learning, but only the Hard Type is required to model user satisfaction. These results based on manually labeled disengagement suggest that while responding to an overall measure of student disengagement can improve both learning and user satisfaction in computer tutoring, maximizing performance requires the system to respond differently based on disengagement type.

In the second part of the paper, we addressed the real world task of automatically detecting and responding to multiple student disengagement types. First, investigation of our machine learning model of student disengagement showed that its automatic overall disengagement (DISE) label correlates with both of our performance metrics in the same way as the manual label. Although our automatic disengagement detector has not yet been extrinsically evaluated in an experiment with real subjects, these correlational results are important because they suggest that our automatic DISE label is both useful and is a reasonable substitute for the manual DISE label. In particular, the similarity of the automatic and manual correlations suggest that the automatic DISE labels are sufficiently accurate, and the significant negative correlations themselves suggest that our approach to automatically detecting and responding to disengagement has the potential to remediate disengagement and thereby improve performance, even in the presence of the noise introduced by the automatic detection process. Second, further correlational analysis showed that the substantial errors introduced when automatically detecting our six different disengagement types could be minimized by categorizing our disengagement types into only two categories, based on their differing correlations with correctness. This enabled us to develop two sets of system responses to disengagement that build on the results of prior work: one set targeting disengaged+correct turns, and another set targeting disengaged+incorrect turns.

We recently evaluated our disengagement adaptation in the "ideal" environment of a Wizard of Oz experiment, where user disengagement, uncertainty, and correctness are labeled by a hidden human during user interactions with ITSPOKE. Our results show that under these wizard conditions, our approach to adapting to disengagement can improve multiple metrics of system performance, including increasing student motivation, reducing uncertainty levels, and reducing the likelihood of continued disengagement, while also breaking the negative correlations between overall disengagement and student learning (Forbes-Riley & Litman, 2012a). Our next step is to see how disengagement detection and adaptation impact performance in the "real" environment of a fully automated system. To this end, we are currently implementing our disengagement detector in ITSPOKE. We will then evaluate the resulting spoken dialogue system for automatically detecting and adapting to multiple affective states in an upcoming controlled experiment with real users.

## ACKNOWLEDGMENTS

This work is funded by National Science Foundation (NSF) award #0914615 and #0631930. We thank Art Ward and our anonymous reviewers for comments. We thank Scott Silliman for systems support.

## REFERENCES

Acee, T. W., Kim, H., Kim, H. J., Kim, J. I., Chu, H. N. R., Kim, M., Cho, Y. J., & Wicker, F. W. (2010). Academic boredom in under-and over-challenging situations. *Contemporary Educational Psychology*, *35*(1), 17–27.

Afzal, S., & Robinson, P. (2011). Natural affect data: Collection and annotation. In S. D'Mello, & R. Calvo (Eds.) *Affect and Learning Technologies*. Springer.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In *Proc. 7th International Intelligent Tutoring Systems Conference (ITS)*, (pp. 227–239). Maceio, Brazil.

Arroyo, I., Ferguson, K., Johns, J., Dragon, T., Merheranian, H., Fisher, D., Barto, A., Mahadevan, S., & Woolf, B. (2007). Repairing disengagement with non-invasive interventions. In *Proc. Artificial Intelligence in Education (AIED)*, (pp. 195–202).

Baker, R. S., Corbett, A., Koedinger, K., Evenson, S., Roll, I., Wagner, A., Naim, M., Raspat, J., Baker, D., & Beck, J. (2006). Adapting to when students game an intelligent tutoring system. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, (pp. 392–401).

Baker, R. S., Corbett, A., Roll, I., & Koedinger, K. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction (UMUAI)*, *18*(3), 287–314.

Baylor, A. L., Ryu, J., & Shen, E. (2003). The effect of pedagogical agent voice and animation on learning, motivation, and perceived persona. In *Proceedings of the ED-MEDIA Conference*. Honolulu, Hawaii.

Beck, J. (2005). Engagement tracking: using response times to model student disengagement. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED)*, (pp. 88–95). Amsterdam.

Bohus, D., & Horvitz, E. (2009). Models for multiparty engagement in open-world dialog. In *Proceedings of SIGdial*. London, UK.

Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, *19*(3), 267–303.

Daschmann, E. C., Goetz, T., & Stupnisky, R. H. (2011). Testing the predictors of boredom at school: Development and validation of the precursors to boredom scales. *British Journal of Educational Psychology*.

de Vicente, A., & Pain, H. (2002). Informing the detection of the students' motivational state: An empirical study. In *Proceedings of the Intelligent Tutoring Systems Conference (ITS)*, (pp. 933–943).

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

Dzikovska, M., Moore, J., Steinhauser, N., & Campbell, G. (2011). Exploring user satisfaction in a tutorial dialogue system. In *Proc. 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, (pp. 162–172). Portland, Oregon.

Forbes-Riley, K., & Litman, D. (2006). Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings of the Human Language Technology Conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)*, (p. 2). New York, NY.

Forbes-Riley, K., & Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, *53*(9–10), 1115–1136.

Forbes-Riley, K., & Litman, D. (2012a). Adapting to multiple affective states in spoken dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on on Discourse and Dialogue (SIGDIAL)*, (pp. 217–226). Seoul, South Korea.

Forbes-Riley, K., & Litman, D. (2012b). Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proceedings Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, (pp. 91–102). Montreal, Canada.

Forbes-Riley, K., Litman, D., & Friedberg, H. (2011). Annotating disengagement for spoken dialogue computer tutoring. In S. D'Mello, & R. Calvo (Eds.) *New Perspectives on Affect and Learning Technologies (Explorations in the Learning Sciences, Instructional Systems and Performance Technologies 3)*. Springer.

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

Jeon, J. H., Xia, R., & Liu, Y. (2010). Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In *INTERSPEECH'10*, (pp. 2802–2805).

Jordan, P., Hall, B., Ringenberg, M., Cui, Y., & Rose, C. (2007). Tools for authoring a dialogue agent that participates in learning studies. In *Proc. Artificial Intelligence in Education*.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environments. In *13th Annual ACM International Conference on Multimedia*, (pp. 677–682). Singapore.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What are you feeling? Investigating student affective states during expert human tutoring sessions. In *Intelligent Tutoring Systems Conference (ITS)*, (pp. 50–59). Montreal, Canada.

Martalo, A., Novielli, N., & de Rosis, F. (2008). Attitude display in dialogue patterns. In *Proceedings of AISB 2008 AISB 2008 Symposium on Affective Language in Human and Machine*, (pp. 1–8). Aberdeen, Scotland.

Murray, R. C., & vanLehn, K. (2005). Effects of dissuading unnecessary help requests while providing proactive help. In *Proc. of the International Conference on Artificial Intelligence in Education*, (pp. 887–889).

Pekrun, R., Goetz, T., Daniels, L., Stupinsky, R., & Perry, R. (2010). Boredom in achievement settings: Exploring control-value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology*, *102*(3), 521–549.

Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, *18*, 125–173.

Rotaru, M., & Litman, D. (2009). Discourse structure and performance analysis: Beyond the correlation. In *Proceedings 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. London, UK.

Schuller, B., Steidl, S., & Batliner, A. (2009). The Interspeech 2009 Emotion Challenge. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech)*. ISCA, Brighton, UK.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Muller, C., & Narayanan, S. (2010). The Interspeech 2010 Paralinguistic Challenge. In *Proceedings of the 11th Annual Conference of the International Speech Communication Assocation (Interspeech)*, (pp. 2794–2797). Chiba, Japan.

Sidner, C., & Lee, C. (2003). An architecture for engagement in collaborative conversations between a robot and a human. Tech. Rep. TR2003-12, MERL.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C. V., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3).

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & RosÃĺ, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3–62.

VanLehn, K., Jordan, P., Rosé, C., Bhembe, D., Böttner, M., Gaydos, A., Makatchev, M., Pappuswamy, U., Ringenberg, M., Roque, A., Siler, S., Srivastava, R., & Wilson, R. (2002). The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. Intelligent Tutoring Systems Conference*.

Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., & Stallard, D. (2002). DARPA communicator: Cross-system results for the 2001 evaluation. In *Proc. ICSLP*.

Walonoski, J., & Heffernan, N. (2006). Prevention of off-task gaming behavior in intelligent tutoring systems. In *Proc. Intelligent Tutoring Systems (ITS)*, (pp. 722–724).

Wang, W., & Hirschberg, J. (2011). Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning. In *Proc. 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, (pp. 152–161). Portland, Oregon.