Engagement tracing: using response times to model student disengagement

Joseph E. BECK
Center for Automated Learning and Discovery
Project LISTEN (www.cs.cmu.edu/~listen), Carnegie Mellon University
RI-NSH 4215, 5000 Forbes Avenue, Pittsburgh, PA. USA 15213-3890

Abstract. Time on task is an important predictor for how much students learn. However, students must be focused on their learning for the time invested to be productive. Unfortunately, students do not always try their hardest to solve problems presented by computer tutors. This paper explores student disengagement and proposes an approach, engagement tracing, for detecting whether a student is engaged in answering questions. This model is based on item response theory, and uses as input the difficulty of the question, how long the student took to respond, and whether the response was correct. From these data, the model determines the probability a student was actively engaged in trying to answer the question. The model has a reliability of 0.95, and its estimate of student engagement correlates at 0.25 with student gains on external tests. We demonstrate that simultaneously modeling student proficiency in the domain enables us to better model student engagement. Our model is sensitive enough to detect variations in student engagement within a single tutoring session. The novel aspect of this work is that it requires only data normally collected by a computer tutor, and the affective model is statistically validated against student performance on an external measure.

1. Introduction

Time on task is an important predictor for how much students learn. However, it is also important to ensure students are engaged in learning. If students are disinterested, learning will not be efficient.

Intelligent tutoring system (ITS) researchers sometimes have an implicit model of the student's engagement; such models help deal with the realities of students interacting with computer tutors. For example, the Reading Tutor [1] asks multiple-choice questions for the purpose of evaluating the efficacy of its teaching interventions. Unfortunately, if students are not taking the assessments seriously, it can be difficult to determine which intervention is actually most effective. If a student hastily responds to a question after just 0.5 seconds, then how he was taught is unlikely to have much impact on his response. Screening out hasty student responses, where students are presumably not taking the question seriously, has resulted in clearer differences between the effectiveness of teaching actions compared to using unfiltered data [2].

A different use of implicit models of student attitudes is the AnimalWatch mathematics tutor [3]. From observation, some students would attempt to get through problems with the minimum work necessary (an example of "gaming the system" [4]). The path of least resistance chosen by many students was to rapidly and repeatedly ask for more specific help until the tutor provided the answer. Setting a minimum threshold for time spent on the current problem, below which the tutor would not give help beyond "Try again" or "Check your work," did much to curtail this phenomenon.

In both the cases mentioned above, a somewhat crude model was added to an ITS to account for not all students being actively engaged: students who spent more time than the threshold were presumed to be trying, those who spent less time were presumed to be

disengaged. These ad hoc approaches have drawbacks: differences among students and questions were ignored. Furthermore these approaches are unable to detect changes in student engagement over time in order to provide better tutoring.

This paper introduces a new technique, *engagement tracing*, to overcome these shortcomings. If the tutor can detect when students are disengaged with an activity it can then change tactics by perhaps asking fewer questions or at the very least disregarding the data for the purposes of estimating the efficacy of the tutor's actions.

2. Domain being modeled

This paper focuses on modeling disengagement by examining student performance on multiple-choice cloze questions [5]. The 2002-2003 Reading Tutor generated cloze questions by deleting a word (semi) randomly from the next sentence in the story the student was reading. The distractors were chosen to be words of similar frequency in English as the deleted word. The Reading Tutor read the sentence aloud (skipping over the deleted word) to the student and then read each response choice. The student's task was to select the word that had been deleted from the sentence. Since the process of generating cloze questions was random, it was uncommon to see repeats of questions and response choices, even when considering hundreds of students using the tutor. There are four types of cloze questions: sight, easy, hard, and defined. The cloze question's type is based on the word that was deleted; sight word questions were for very common words, hard questions were for rarer words, and defined word questions were for words a human annotated as probably requiring explanation. See [2] for additional details about how the cloze question intervention was instantiated in the Reading Tutor.

Which cloze data are relevant? One concern was whether students would take cloze questions seriously. Project LISTEN member Joe Valeri suggested that if students weren't really trying to get the question correct, they would probably respond very quickly. As seen in Figure 1**,** student performance on cloze questions was strongly related to how much time they spent answering a question. Since chance performance is 25% correct, it is safe to infer that students who only spent one second before responding were not trying to answer the question and were probably disengaged. Similarly, a student who spent 7 seconds was probably engaged. But what of a student who spent 3 seconds? Students responding after 3 seconds were correct 59% of the time, much better than baseline of 25% but not nearly as high as the 75% correct attained by students who spent 5 seconds. Should we consider such a response time as a sign of disengagement or not?

We consider four general regions in Figure 1. In region R1, students perform at chance. In region R2, student performance is improving as more time as spent. In region R3, performance has hit a plateau. In region R4, performance is gradually declining as student spend more time before responding to the question.

Although there is certainly a correlation between student performance and student engagement, we do not treat the decline in student performance in region R4 as a sign of disengagement. Without more extensive instrumentation, such as human observers, we cannot be sure why performance decreases. However, it is likely that students who know the answer to a question respond relatively quickly (in 4 to 7 seconds). Students who are less sure of the answer, or who have to answer on the basis of eliminating some of the choices based on syntactic constraints, would take longer to respond. This delay is not a sign of disengagement; therefore, to maintain construct validity, we do not consider long response times to be a sign of disengagement. For purposes of building a model to predict the probability a student is disengaged, we only consider data in regions R1, R2, and R3.

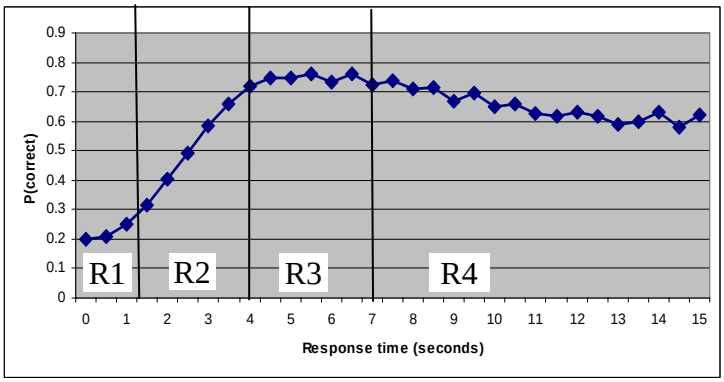

Figure 1. Student proportion correct on cloze questions plotted by response time

Describing the relation between response time and performance. Throughout regions R1, R2, and R3, performance with respect to time is similar to a logistic curve. Therefore, we use item response theory [6] as a starting point for our modeling. Item response theory (IRT) provides a framework for predicting the probability a student with a particular proficiency will answer a question correctly.

Three parameter IRT models [6] are of the form $p(correct \mid \theta) = c + \frac{1-c}{1+e^{-a(\theta-b)}}$. In this equation, θ represents the student's proficiency. The other three parameters control the shape of the logistic curve: a is the discrimination parameter, and determines the steepness of the logistic curve; b is the item difficulty parameter, and controls how far left or right the curve is shifted, and c is the "guessing" parameter and provides a lower bound for the curve. Since our items are multiple choice questions with four responses, we set c to be 0.25.

For our work, we need to modify the standard IRT formula in three ways. First, rather than taking student proficiency as input, our model uses response time as an input. Second, we cannot estimate item parameters for every cloze question, as a pure IRT model would, since most questions were only seen once. Therefore, we estimate discrimination and item difficulty parameters for each of the four types of cloze questions. Since the difficulty parameter cannot capture the differences between questions of a particular type, we also include the length of the cloze question and response choices (as the number of characters). Longer questions are probably harder than shorter ones, and at the very least should take more time to answer. Finally, in IRT models, as students become more proficient the chances of a correct response increase to 100%. For our model, the upper bound on performance is considerably less than 100%. If a student does not know the answer, giving him additional time (unless he has resources such as a dictionary to help him) is unlikely to be helpful. Therefore we introduce an additional parameter, u, to account for the upper bound on student performance.

The form of our modified model is $p(correct \mid rt, L_1, L_2) = c + \frac{u-c}{1+e^{-a(-rt+b(L_1+L_2))}}$. Parameters a, b, and c have the same meaning as in the IRT model. The u parameter represents the upper bound on performance, and L_1 and L_2 are the number of characters in the question and in all of the response choices combined, respectively. The u parameter is equal to the maximum performance (found by binning response times at a grain size of 0.5 seconds, and selecting the highest average percent correct).

We estimate the a (discrimination) and b (difficulty) parameters separately for each type of cloze question using SPSS's non-linear regression function. All question types have a similar difficulty parameter; the difference in difficulty of the questions is largely accounted for by the longer question and prompts for more difficult question types. For

predicting whether a student would answer a cloze question correctly, this model accounts for 5.1% of the variance for defined word questions, 12.3% for hard words, 14.5% for easy words, and 14.3% for sight words. These results are for testing and training on the same data set. However, the regression model is fitting only two free parameters (*a* and *b*) for each question type, and there are 1080 to 3703 questions per question type. Given the ratio of training data to free parameters, the risk of overfitting is slight, and these results should be representative of performance on an unseen test set.

Determining student engagement. Although our model can estimate the probability of a correct response given a specific response time, this model is not sufficient to detect disengagement. To enable us to make this calculation, we assume that students have two methods of generating responses:

1. If the student is disengaged, then he guesses blindly with a probability *c* of being correct.
2. If the student is engaged, then he attempts to answer the question with a probability *u* of being correct.

Given these assumptions, we can compute the probability a student is disengaged in answering a question as $\frac{u - p(correct \mid rt, L_1, L_2)}{u - c}$. For example, consider Figure 1; if a student took 3 seconds to respond to a question he had a 59% chance of being correct. The lower bound, *c*, is fixed at 25%. The upper bound, *u*, is the best performance in region R3, in this case 76%. So the probability the student is disengaged is (76% - 59%) / (76% - 25%) = 33%, and therefore a 67% chance that he is engaged in trying to answer the question.

This model form is similar to knowledge tracing [7], in that both are two-state probabilistic models attempting to estimate an underlying student property from noisy observations. Since this model concerns student engagement rather than knowledge, we call it *engagement tracing*.

To illustrate the above process, Figure 2 shows our model's predictions and students' actual performance on hard word cloze questions. To determine the student's actual performance, we discretize the response time into bins of 0.5 seconds and took the mean proportion correct within the bin. To determine the performance predicted by the model, we use the estimates for the *a*, *b*, and *u* parameters, and assume all questions are of the mean length for hard question types (47.8 character prompt + 26.3 character response choices = 74.1 characters). As indicated by the graph, students' actual (aggregate) performance is very similar to that predicted by the model; the r^2 for the model on the aggregate data is 0.954, indicating that the model form is appropriate for these data.

However, this model does not account for individual differences in students. For example, a very fast reader may be able to read the question and response choices, and consistently give correct answers after only 1.5 seconds. Is it fair to assert that this student is not engaged in answering the question simply because he reads faster than his peers? Therefore, to better model student engagement, we add parameters to account for the variability in student proficiency.

Accounting for individual differences. One approach to building a model to account for inter-student variability is to simply estimate the *a*, *b*, and *u* parameters for each student for each question type (12 total parameters). Unfortunately, we do not have enough data for each student to perform this procedure. Students saw a mean of 33.5 and a median of 22 cloze questions in which they responded in less than 7 seconds. Therefore, we first estimate the parameters for each question type (as described above), and then estimate two additional parameters for each student that apply across all question types. The new model form becomes $p(correct \mid rt, L_1, L_2) = c + \frac{accuracy(1-u) + u - c}{1 + e^{-a(-rt + speed * b(L_1 + L_2))}}$ where *accuracy* and

speed are the student-specific parameters. The first additional parameter, *speed*, accounts for differences in the student's reading speed by adjusting the impact of the length of the question and response choices. The second parameter, *accuracy*, is the student's level of knowledge. Students who know more words, or who are better at eliminating distractors from the response choices will have higher asymptotic performance.

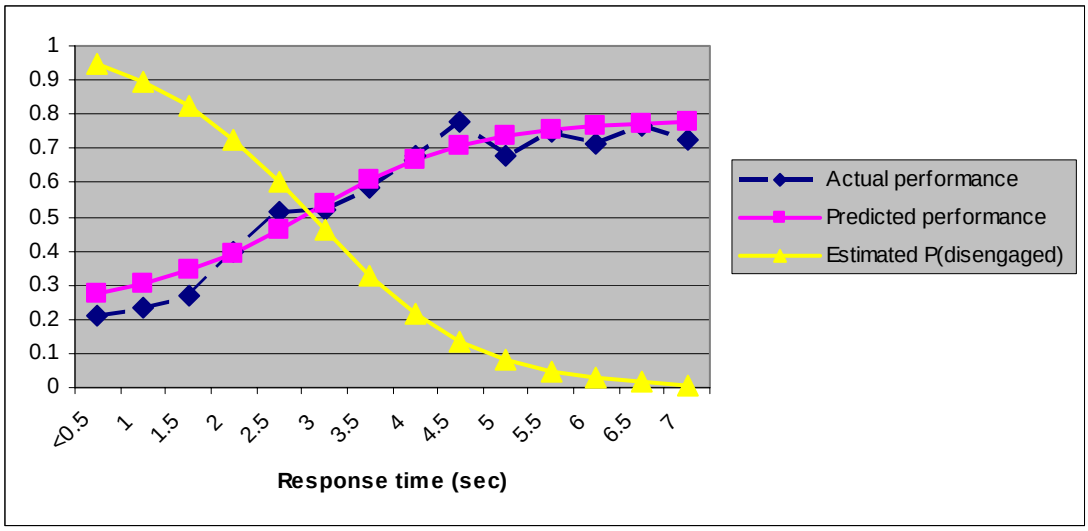

Figure 2. Empirical and predicted student behavior for hard word cloze questions

We estimate the student parameters with SPSS's non-linear regression procedure. The student-specific parameters are bounded to stop semantically nonsensical results. The *speed* parameter is forced to be in the range [0.33, 3] (i.e. it can model that students are three times faster or slower at reading than average) and the *accuracy* parameter is in the range [-2, 1] (i.e. students can not have performance over 100%). Thus we avoid obtaining a good model fit by assigning a student parameter a value that is implausible (such as reading 25 times faster than average).

3. Psychometric properties of model

There are two major psychometric properties: reliability, whether the measure is consistent, and validity, whether the model measures what it is supposed to measure. In our experimental design, for each cloze question a student encountered, we use our engagement tracing model to estimate the probability a student is engaged in answering the question. For each student, we take the mean probability of disengagement across all of the questions as a measure of the student's overall disengagement with the tutor.

Although our model's parameters are estimated from questions where students respond in fewer than 7 seconds, to estimate overall disengagement we use data from all cloze questions, even those with longer response times. Our belief is that students taking longer than 7 seconds to respond are engaged. As seen in Table 2, as response time increases the estimated probability of disengagement decreases, so including longer response times led the model to believe students were more engaged.

Students saw a mean of 88.7 and a median of 69 cloze questions. The mean probability of disengagement (for the student-specific model) is 0.093 and the median is 0.041. The probability of disengagement is positively skewed, with one student having a value of 0.671. This student saw 171 cloze items, so the high average disengagement is not a statistical fluke from seeing few items. Four students had disengagement scores over 0.5.

Reliability. To determine whether our engagement measure is psychometrically reliable, we use a split-halves approach by ordering each student's cloze data by time and

assigning each cloze item to alternating groups (i.e. observation #1 is in group A, observation 2 is in group B, observation 3 is in group A, …). For each student, we then estimate the overall disengagement for A and for B. The corrected-split halves reliability is 0.95, comparable to the best psychometric instruments. Thus, our measure of disengagement is highly reliable.

Validity. To measure validity, we relate our measure of disengagement to existing tests of student performance and interest in the domain. Our hypothesis is that a measure of student disengagement should correlate negatively with student gains in reading over the course of the year. This hypothesis came from [4] as well as the intuition that an active, engaged learner is likely to make more progress than one who takes less initiative. We measure reading gains as the difference between the student's pretest and posttest on the (human-administered and scored) Woodcock Reading Mastery Test's [8] Total Reading Composite (TRC) subtest. We also examine how our measure of engagement correlates with the student's attitude towards reading as measured by the Elementary Reading Attitude Survey (ERAS) recreational reading subscale [9]. We have data and test scores for 231 students who were in grades one through six (approximately five- through twelve-year olds) during the 2002-2003 school year.

We compare three models of student engagement: a model with student-specific parameters (*speed* and *accuracy*), a model without the two student-specific parameters, and the percentage of questions to which a student responds to in less than 2.5 seconds, which corresponds to a ≈50% chance of engagement. Table 1 shows how the measures of disengagement, student attitude towards reading, and learning gains interrelate. These partial correlations hold constant student TRC pretest scores and student gender. All of the disengagement measures correlate with student gains in TRC at $p<0.05$, with the per-student model producing the strongest results. All correlations are in the intuitive direction: disengaged students have smaller learning gains while students with a positive attitude towards reading have higher gains.

Table 1. Partial correlations between disengagement, learning gains and reading attitude

	Measures of disengagement			Reading attitude
	Per-student model	Basic model	Response < 2.5 s	ERAS
TRC gain	-0.25 (p<0.001)	-0.16 (p=0.013)	-0.15 (p=0.023)	0.18 (p=0.007)
ERAS	-0.03	0.04	0.03	-

Somewhat surprisingly, none of the measures correlate with the student's attitude towards reading. Perhaps the measures of disengagement are unrelated to the student's overall attitude, but instead measure the student's specific feelings about working with the Reading Tutor or with its multiple choice questions.

4. Temporal properties of model

Although engagement tracing is psychometrically reliable, that does not mean student engagement is stable across time. We investigate two ways in which engagement can vary. Systematic change refers to students becoming consistently more or less engaged over the course of the year. Ephemeral change investigates whether our approach is sensitive enough to detect waxing and waning student engagement. For both investigations we focus on when cloze questions occur.

Systematic properties. To find systematic trends in student engagement, for each cloze question we compute how long the student has been using the Reading Tutor before encountering the cloze question, and then bin questions based on how many months the student has been using the tutor. During the first month, students have a mean

disengagement of 6%. For each successive month the amount of disengagement increases until reach a plateau at the 4th month: 10.3%, 10.9%, 16.5%, 15.3%, and finally 16.5% during the 6th month of usage. Whether this result means students are becoming less engaged with the Reading Tutor or just bored with the questions is unclear.

Ephemeral properties. Presumably, student engagement should be similar across a small time interval, and vary more widely over a larger window. Can engagement tracing detect such transient effects? To answer this question, for a cloze question Q1, we pair Q1 with every successive cloze question seen by that student and compute the amount of intervening time between the questions. We then examine two models: the first correlates student engagement on Q1 and Q2; the second model computes a partial correlation between Q1 and Q2, holding constant the student's average level of disengagement throughout the year. Table 2 shows the results of this procedure.

Table 2. Detecting ephemeral properties of disengagement

Time between Q1 and Q2	Overall correlation	Partial correlation
< 1 minute	0.69	0.45
1 to 5 minutes	0.66	0.35
Later that day	0.63	0.21
Later that week	0.67	0.15
More than a week later	0.53	0.00

Overall, student performance on Q1 is strongly correlated with later performance on Q2. This result is not surprising, since a student presumably has an underlying level of engagement; thus we expect a strong autocorrelation. The partial correlation shows ephemeral trends in engagement. Specifically, student engagement on one question accounts for 19.8% of the variance in each measurement of engagement within a one-minute window, *even after controlling for the student's overall level of engagement throughout the year*. In contrast, a particular question only accounts for 2.3% of the variance of each measurement of student engagement later that week. This result both points to temporal trends in students using the Reading Tutor: engagement is much more consistent within a one- or five-minute interval than across successive days, and to the ability of engagement tracing to detect such differences.

5. Contributions, conclusions, and future work

Although by focusing on a single type of affect, namely disengagement, this work is narrower in scope than most prior work (e.g. [10-12]), it differs from that work by providing an empirical evaluation of whether the affective model relates to externally meaningful measures of real students. Also, the approach described in this paper does not require humans to rate user interactions (as in [12]) or measurements with biological sensors (as in [11]).

We have presented a means for analyzing the response times and correctness of the student responses to model overall level of engagement while using a computer tutor. This result is general as both response time and correctness are easily measurable by an ITS, do not require investing in new equipment, and are common across a wide variety of computer tutors.

The psychometric properties of the model include very strong reliability, and external validity to the extent of a moderate correlation with paper test scores. The model is sensitive enough to detect temporal changes in the student's level of engagement within a single session of using the tutor.

Future work with engagement tracing includes adding a temporal component to the model. Currently we simply take the mean of all student observations. Given the temporal

nature of engagement, some means of discounting older observations is needed. To compare with knowledge tracing [7], this paper develops a framework comparable to the performance parameters (slip and guess), but does not yet have an equivalent to the learning parameters to account for initial student state and transitions between states.

This paper demonstrates that simultaneously modeling the student's proficiency and engagement allows us to better estimate his level of engagement than a model that ignores individual differences in proficiency. In the short-term, modeling a student's level of engagement enables predictions about how much an individual student will benefit from using a computer tutor. In the longer term, adapting the tutor's interactions to keep the learner happy and engaged—while not sacrificing pedagogy—is a fascinating problem.

Acknowledgements

This work was supported by the National Science Foundation, under ITR/IERI Grant No. REC-0326153. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government. The author also thanks Ryan Baker and Cecily Heiner for providing useful comments on this work, and Jack Mostow for coining the term "engagement tracing" to describe the work.

References

1. Mostow, J. and G. Aist, *Evaluating tutors that listen: An overview of Project LISTEN*, in *Smart Machines in Education*, K. Forbus and P. Feltovich, Editors. 2001, MIT/AAAI Press: Menlo Park, CA. p. 169-234.
2. Mostow, J., J. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri, *Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions.* Technology, Instruction, Cognition and Learning, to appear. **2**.
3. Woolf, B.P., J.E. Beck, C. Eliot, and M.K. Stern, *Growth and Maturity of Intelligent Tutoring Systems: A Status Report*, in *Smart Machines in Education: The coming revolution in educational technology*, K. Forbus and P. Feltovich, Editors. 2001, AAAI Press. p. 99-144.
4. Baker, R.S., A.T. Corbett, K.R. Koedinger, and A.Z. Wagner. *Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System."* in *ACM CHI*. 2004.p. 383-390.
5. Entin, E.B., *Using the cloze procedure to assess program reading comprehension.* SIGCSE Bulletin, 1984. **16**(1): p. 448.
6. Embretson, S.E. and S.P. Reise, *Item Response Theory for Psychologists*. Multivariate Applications, ed. L.L. Harlow. 2000, Mahwah: Lawrence Erlbaum Associates. 371.
7. Corbett, A.T. and J.R. Anderson, *Knowledge tracing: Modeling the acquisition of procedural knowledge.* User Modeling and User-Adapted Interaction, 1995. **4**: p. 253-278.
8. Woodcock, R.W., *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU).* 1998, Circle Pines, Minnesota: American Guidance Service.
9. Kush, J.C. and M.W. Watkins, *Long-term stability of children's attitudes toward reading.* The Journal of Educational Research, 1996. **89**: p. 315-319.
10. Kopecek, I., *Constructing Personality Model from Observed Communication*. 2003: Proceedings of Workshop on Assessing and Adaptive to User Atitudes and Effect: Why, When, and How? at the Ninth International Conference on User Modeling. p. 28-31.
11. Conati, C., *Probabilistic Assessment of User's Emotions in Educational Games.* Journal of Applied Artificial Intelligence, 2002. **16**(7-8): p. 555-575.
12. Vicente, A.d. and H. Pain. *Informing the Detection of the Students' Motivational State: an Empirical Study*. in *Sixth International Conference on Intelligent Tutoring Systems*. 2002.p. 933-943 Biarritz, France.