Using response times to model student disengagement

Joseph E. Beck
joseph.beck@cmu.edu
Project LISTEN. www.cs.cmu.edu/~listen
Center for Automated Learning and Discovery
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213. USA

Abstract. Time on task is an important variable for learning a skill. However, learners must be focused on the 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 a model 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. To validate our model, we analyze 231 students' interactions with the 2002-2003 version of the Reading Tutor. We show that disengagement is better modeled by simultaneously estimating student proficiency and disengagement than just estimating disengagement alone. Our best model of disengagement has a correlation of -0.25 with student learning gains. The novel aspect of this work is that it requires only data normally collected by a computer tutor, and the affective model is validated against student performance on an external measure.

Keywords: affective diagnosis, measure of emotions, student modeling, item response theory, engagement/motivation

1 Introduction

Time on task is an important predictor of how well students learn a skill. However, it is important to make sure this time is well spent and the student is actively engaged in the learning. Ensuring that time is well spent is an aspect of pedagogy, an area in which intelligent tutoring systems (ITS) excel. However, it is also important to ensure students are engaged in the learning. If students are disinterested, it does not matter how pedagogically appropriate the material is, learning will not be efficient.

ITS researchers sometimes have an implicit model of the student's motivation; such models help deal with the realities of students interacting with computer tutors. For example, the Reading Tutor [6] 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 it is unlikely that how he was taught will 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 [7].

A different use of implicit models of student attitudes is the AnimalWatch mathematics tutor [10]. From observation, some students would attempt to get through problems with the minimum work necessary (an example of "gaming the system" [1]). The path of least resistance chosen by many students was to rapidly hit the return key until the tutor gave clear instructions on how to solve the problem. 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.

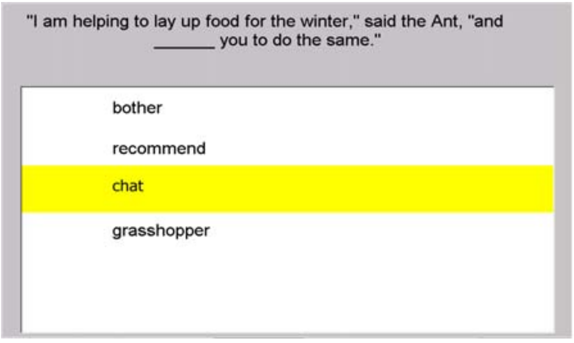
In both the cases mentioned above, the threshold for whether students were really trying was somewhat crude: a constant time threshold. Students who spent more time than the threshold required were presumed to be trying, those who spent less time were presumed to be disengaged. Differences in either the students or the questions were ignored.

This paper overcomes the shortcoming of not accounting for differences in questions or student proficiency, and addresses whether we can model how much effort students are making while solving problems in an ITS. The goal is to determine when students are disengaged with an activity, so the tutor can then change tactics by perhaps asking fewer questions, or at the very least disregard the data for the purposes of estimating the efficacy of the tutor's actions or intervening with more directed help.

2 Domain being modeled

This paper focuses on student performance on multiple-choice cloze questions. Cloze questions [4] display a sentence with one of the words deleted, and students are asked to supply the missing word. Multiple-choice cloze questions were an intervention in the 2002-2003 Reading Tutor, and were designed to assess the student's reading comprehension proficiency [7]. Figure 1 shows an example of a cloze question. Cloze questions were generated 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 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 click on the word that had been deleted from the sentence. Since the process of generating cloze questions was random, it is 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 [7] for additional details about how the cloze question intervention was instantiated in the Reading Tutor.

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. In fact, student performance on cloze question was strongly related to how much time they spent answering a question. Figure 2 shows how accurate students were at answering cloze questions based on how much time they spent before responding. Since chance performance is 25% correct, it is safe to infer that students who only spent one second answering a question were 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 a sign of disengagement or not? Rather than forcing a binary decision, we instead focus on computing the probability the student was disengaged while responding to the question.



"I am helping to lay up food for the winter," said the Ant, "and _______ you to do the same."

Figure 1. Example cloze question in the Reading Tutor

2.1 Model assumptions

We make three assumptions in our modeling: which data to include to build the model, the mathematical form of the model, and how students generate responses.

Which data are relevant. We consider four general regions in Figure 2. In region A, students perform at chance. In region B, student performance is improving as more time as spent. In region C, performance has hit a plateau. In region D, 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 did not treat the decline in student performance in region D as a sign of disengagement. Without more extensive instrumentation, such as human observers, we cannot be sure why performance decreased. However, it is more likely that students who knew the answer to a question responded relatively quickly (in 4 to 7 seconds). Students who were less sure of the answer, or who had to answer on the basis of eliminating some of the choices based on syntactic constraints, would take longer to respond. To maintain construct validity we do not consider long response times to be a sign of disengagement. Therefore, for purposes of building a model to predict the probability a student is disengaged, we only consider data in regions A, B, and C.

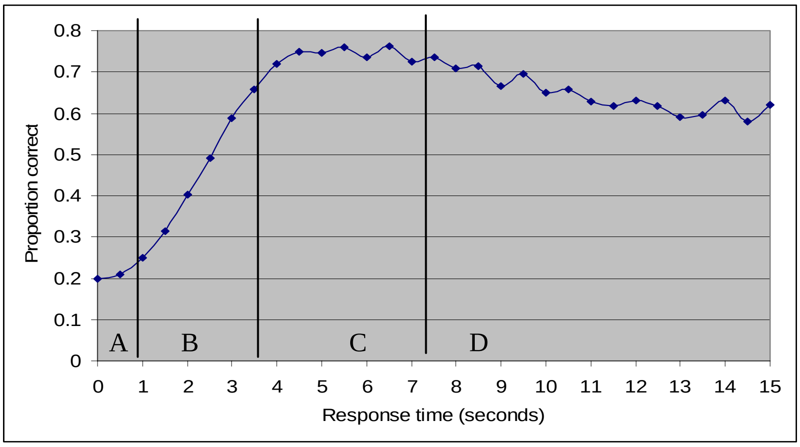


Figure 2. Student proportion correct on cloze questions plotted by response time

Form of student performance curve. Throughout regions A, B, and C, performance with respect to time is similar to a logistic curve. Therefore, we use item response theory [3] (or see http://www.uts.psu.edu/Item_Response_Theory_frame.htm for a good online introduction) 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. We need to construct a model rather than using Figure 2 directly since accuracy vs. response time will vary based on the type and length of the cloze question. We do not have sufficient data to directly estimate these values; therefore we construct a mathematical model and estimate its parameters.

Students respond to questions in one of two ways. Although we are able to estimate the probability of a correct response for a student who spends a certain amount of time responding, this measure 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 of 0.25 of being correct (since there are four response choice for cloze questions)
2. If the student is engaged, he attempts to answer the question with a probability of being correct equal to the best performance in region C.

Given these assumptions, the probability the student was disengaged is (upper bound - expected performance) / (upper bound – lower bound). For example, consider Figure 2; if a student took 3 seconds to respond to a question he had a 59% chance of being correct. The lower bound is fixed at 25%. The upper bound is the best performance in region C, in this case 76%. So the probability the student is disengaged is (76% - 59%) / (76% -

25%) = 33%, and therefore a 67% chance that he was engaged in trying to answer the question. This assumption allows us to map expected probability of correct to expected probability of being disengaged.

2.2 Model form

To estimate the probability a student's response is correct, we use item response theory. Three parameter IRT models [3] 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 (a steeper curve better discriminates between students of similar ability); *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 formula in several 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 do, since the modal number of times a particular question was seen was 1. Therefore, we estimate discrimination and item difficulty parameters for each of the four types of cloze question. 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. 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 to account for the upper bound on student performance.

The form of our modified model is $p(correct \mid rt, L_1, L_2) = c + \frac{d-c}{1+e^{-a(-rt+b(L_1+L_2)}}$. Parameters *a*, *b*, and *c* have the same meaning as in the IRT model. The *d* 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, respectively. The *d* parameter was 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 estimated the *a* (discrimination) and *b* (difficulty) parameters separately for each type of cloze question using SPSS's non-linear regression function. Table 1 shows the parameters estimates and the average length of the question and the prompt. 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 cloze question was correct, this model accounted 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[1]. 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.

[1] Although SPSS provides a straightforward mechanism for computing leave-one-out cross validation results for linear regression, such functionality appears to be missing for non-linear regressions. Anyone knowing a means of performing such a cross validation with SPSS in encouraged to contact the author.

Table 1. Model parameters and mean values for each question type

	Question type			
	Sight	Easy	Hard	Defined
Number of questions	3685	3703	2424	1080
a	-1.55	-1.34	-1.14	-0.67
b	0.038	0.037	0.039	0.036
d	0.81	0.80	0.78	0.63
L_1	45.2	45.8	47.8	67.3
L_2	17.3	23.4	26.3	33.2

Figure 3 shows our model's predictions and students' actual performance on hard word cloze questions. To determine the student's actual performance, we discretized the response time into bins of 0.5 seconds and took the mean proportion correct within the bin (there are only 28 data points for 0.5 seconds, so random variation may be why actual performance appears to drop from 0 seconds to 0.5 seconds). To determine the performance predicted by the model, we used the data in Table 1 for the *a, b,* and *d* parameters, and assumed all questions were of the mean length for hard question types (47.8 + 26.3 = 74.1 characters). As indicated by the graph, students' actual (aggregate) performance is very similar to that predicted by the model.

We now have a model that, given a cloze question type, the number of characters in the question and response choices, and time it took for the student to respond, generates a predicted probability the student was disengaged. For example, for a response time of 1.5 seconds, the predicted chance of a correct response was 34%. So a student responding in 1.5 seconds had an estimated (78% – 34%) / (78% – 25%) = 82% chance of being disengaged. However, this model does not account for individual differences in student performance. 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.

2.3 Accounting for individual differences

One approach to building a model to account for inter-student variability is to simply estimate the *a, b,* and *d* 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 $p(correct \mid rt, L_1, L_2) = c + \frac{accuracy(1-d)+d-c}{1+e^{-a(-rt+speed*b(L_1+L_2)}}$ 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.

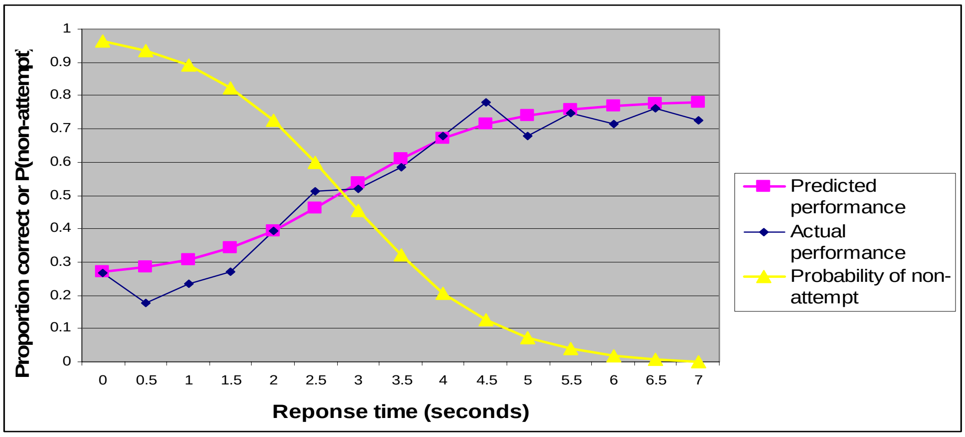


Figure 3. Predicted and actual student performance, and probability of attempting to answer

We used SPSS's non-linear regression procedure to estimate the parameters. The student-specific parameters were bounded to stop semantically nonsensical results. Specifically, the *speed* parameter was forced to be in the range [0.33, 3] (i.e. it could model that students were three times faster or slower at reading than average) and the *accuracy* parameter was in the range [-2, 1] (i.e. students could not have performance over 100%). Given the relatively small amount data per student, we wanted to avoid situations where we could obtain a good model fit by assigning a student a value that was implausible (such as reading 25 times faster than average).

To determine whether the student-specific parameters improved the model, we used the model's accuracy in predicting whether a student answered a question correctly and compared it to two other models: a model that does not use student-specific parameters, and a majority classifier (returns "correct" for sight, easy, and hard questions, and "incorrect" for defined questions). The majority class correctly predicts whether the student will be correct 65.4% of the time, the non-student specific model is correct in its predictions 71.9% of the time, and the student-specific model is correct 77.0% of the time. Therefore, the student specific model appears to best fit the student performance data.

3 Experimental design and results

Although the student specific model best fits the performance data, we needed to determine whether it is best for modeling the student's affective state. Our experimental design was to, for each cloze question a student encountered, take the question characteristics (e.g. the type of cloze question and length of the question and response choices) and apply them to a model that predicted the probability a student would generate the correct response. We then transformed this result into the probability the student had attempted to solve the question. For each student, we took 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 was built with data from questions where students responded in fewer than 7 seconds, to estimate overall disengagement we used student performance data from all cloze questions, even those with longer response times. Our belief was that students taking longer than 7 seconds to response were engaged, even if their performance decreased. As shown in Figure 3, questions on which the student spent considerable time were predicted to have a low probability of disengagement (as time increases the predicted performance would get closer to the upper bound, and the probability of disengagement would approach 0).

We compared three models: a model with student-specific parameters, a model without student-specific parameters, and simply counting how many questions students responded to in less than 2.5 seconds, which approximately corresponds to a 50% chance of engagement.

Students saw a mean of 88.7 cloze questions and a median of 69. The mean probability of disengagement (for the student-specific model) was 0.093 and the median was 0.041. The probability of disengagement was 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.

Our hypothesis was that a measure of student disengagement would correlate negatively with student gains in reading over the course of the year. This hypothesis came from [1] as well as the intuition that an active, engaged learner is likely to make more progress than one who takes less initiative. We measured reading gains as the difference between the student's pretest score on the Woodcock Reading Mastery Test's [9] Total Reading Composite (TRC) subtest and the posttest score. Students were (generally) pretested in October before using the Reading Tutor and posttested in May. The TRC is human administered and scored. We were also curious about how our measure of engagement correlated with the student's attitude towards reading as measured by the Elementary Reading Attitude Survey (ERAS) recreational reading subscale. We had data and test scores for 231 students who were in grades one through six (approximately five through twelve year olds).

Table 2 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 measures listed correlated with student gains in TRC at $p<0.05$, with the per-student model of disengagement producing the strongest results. All correlations were in the intuitive direction: disengaged students had smaller learning gains while students with a positive attitude towards reading had higher gains.

Table 2. Partial correlations between disengagement, learning gains and reading attitude

	Measures of disengagement			Reading attitude
	Per-student model	Basic model	Response < 2.5 sec	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 correlated 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 Conclusions and contributions

We have presented a means for analyzing the response times and correctness of the student's responses to model his 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 tutors. We have found that simultaneously modeling the student's proficiency allows us to better estimate his level of engagement than a model that ignores such individual differences. In the short-term, modeling a student's level of engagement enables predictions about how much students 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.

Although by focusing on a single type of affect, namely disengagement, this work is narrower in scope than most prior work (e.g. [2, 5, 8]), 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 [8]) or measurement with biological sensors (as in [2]).

REFERENCES

1. 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.
2. Conati, C., *Probabilistic Assessment of User's Emotions in Educational Games.* Journal of Applied Artificial Intelligence, 2002. **16**(7-8): p. 555-575.
3. Embretson, S.E. and S.P. Reise, *Item Response Theory for Psychologists*. Multivariate Applications, ed. L.L. Harlow. 2000, Mahwah: Lawrence Erlbaum Associates. 371.
4. Entin, E.B., *Using the cloze procedure to assess program reading comprehension.* SIGCSE Bulletin, 1984. **16**(1): p. 448.

5. 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.

6. 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.

7. 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**.

8. 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.

9. Woodcock, R.W., *Woodcock Reading Mastery Tests - Revised (WRMT-R/NU)*. 1998, Circle Pines, Minnesota: American Guidance Service.

10. 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.