Decomposition of Response Time to Give Better Prediction of Children's Reading Comprehension

Zhila Aghajari *
Lehigh University
Bethlehem, PA
zha219@lehigh.edu

Deniz Sonmez Unal *
University of Pittsburgh
Pittsburgh, PA
des204@pitt.edu

Mesut Erhan Unal
University of Pittsburgh
Pittsburgh, PA
meu6@pitt.edu

Ligia Gómez
Arizona State University
Tempe, AZ
ligia.gomez@asu.edu

Erin Walker
University of Pittsburgh
Pittsburgh, PA
eawalker@pitt.edu

ABSTRACT

Response time has been used as an important predictor of student performance in various models. Much of this work is based on the hypothesis that if students respond to a problem step too quickly or too slowly, they are most likely to be unsuccessful in that step. However, something that is less explored is that students may cycle through different states within a single response time and the time spent in those states may have separate effects on students' performance. The core hypothesis of this work is that identifying the different states and estimating how much time is devoted to them in a single response time period will help us predict student performance more accurately. In this work, we decompose response time into meaningful subcategories that can be indicative of helpful or harmful cognitive states. We then show how a model that is using these subcategories as predictors instead of response time as a whole outperforms both a linear and a non-linear baseline model.

Keywords

Response time, student modeling, regression models, on-task and off-task behaviors

1. INTRODUCTION

Intelligent Tutoring Systems (ITS) help students learn a wide variety of skills from problem solving [23] to reading [22, 19]. To improve ITS designs, researchers often study students' learning patterns to identify their relationship to performance and target them for intervention. Within this context, response time has been widely used to predict student performance [39, 40] and to interpret cognitive and motivational states during ITS use [39, 3, 7].

Much of the research involving response time is based on the hypothesis that the relationship between response time and student performance is non-linear [9]. Fast or slow response times may be indicative of both helpful and harmful cognitive states. For example, a fast response time could be a result of either mastery of a skill or guessing. Likewise, a long response time could be because of struggling or being off-task. Contextual information surrounding response time is often used in identifying the correct cognitive or motivational states. For example, a long response time after reading a bug or a hint message can be linked to reflection [32], whereas a short response time after such actions can be a sign of gaming the system [4]. Thus, previous literature has focused on identifying students' cognitive states based on sequences of actions and the time spent between them [4, 2, 7, 5]. However, students may go through different cognitive states even within a time period between consecutive actions [32]. Despite a large body of research dedicated to studying students' cognitive states, little is known about the different states a student might be in during a single response time and how time spent in those states would affect learning.

We hypothesize that response time can be divided into subcategories that can be indicative of some helpful and harmful cognitive states, and that identifying time spent on these states within one response time can improve student performance prediction. In our previous work [35], we divided response times during a reading comprehension task into two: reading and thinking time. Results of a piecewise regression model revealed that thinking time could include four states: gaming, productive thinking, wheel spinning, and mind wandering. With the insight from these results, we further investigate the different states that could occur in one response time. We compare a model that is based on decomposition of response time to a linear baseline model which only uses average response time, and also to a non-linear baseline (a piecewise regression). By decomposing the response time, we show that students can go through multiple cognitive states in between log events. We also show that by identifying how much time is devoted to these states, we can improve the predictive models of student performance.

2. RELATED WORK

2.1 Cognitive and Motivational States

*These authors contributed equally to this work.

Within this section we review the cognitive states that are related to productive thinking, gaming the system, and unproductive thinking. We extract these states from the broader research literature, although it should be noted that these states were also identified by teachers as important [13].

We gather learning events that are associated with robust learning under **productive thinking** behaviors. [17] divided these events into three categories: understanding and sense-making processes, induction and refinement processes, memory and fluency-related processes. Some example behaviors that fall under understanding and sense-making processes and induction and refinement processes that could be relevant in a reading domain are self-explanation and self-reflection. These behaviors are shown to be positively related to learning [32, 8].

Gaming the system is an undesirable cognitive state wherein students try to reach the correct answers and advance in the lesson by systematically misusing the features of the system [3]. It is linked to short response times and rapid actions [3]. [29] divides gaming into two main types: systematic guessing and help abuse. Systematic guessing could be inferred from short response times between step attempts [12, 28, 2], entering the same answer in multiple contexts, and entering similar answers [29]. Help abuse was defined as searching for bottom-out hints, asking for help without any reflection on the help, and entering multiple incorrect answers despite receiving help.

Within **unproductive thinking** states, we review wheel spinning and mind wandering. **Wheel spinning** occurs when the student makes an effort but does not succeed. It is linked to long response times and many help requests. [7] illustrates that if students need help solving the first twenty problems they are in wheel spinning phase, and presenting them with more problems will not be helpful. [5] showed wheel spinning is negatively correlated with flow, positively correlated with gaming and confusion, and not correlated with boredom. **Mind wandering** occurs when students involuntarily shift their attention to task-unrelated thoughts [15, 14, 34], and is associated with distraction or boredom. This cognitive state occurs 20-40% of the time during reading [30] and causes students to fail in gaining reading comprehension skills [33, 36]. As mind wandering occurs involuntarily, it is very difficult to measure, and it is often measured using self-reported approaches [24].

2.2 Response Time in Student Modeling

Response time has been widely used in different kinds of student models, and can improve the accuracy of those models [39, 20]. For example, [4] presents a model that uses response time to detect shallow learning, and [11] predicts student performance in transfer learning using response time. [6] developed an item response theory (IRT) model to show an overall level of students engagement by analyzing response times, problem difficulty, and correctness of responses. [16] also presents an IRT-based model to estimate student proficiency and motivation level where motivation was measured based on time spent between actions and a short response time was an indicator of unmotivated behavior.

In this paper, we are inspired by work that centers response time as a non-linear predictor of students' performance. [9] suggests that the relationship between time and student success is not linear, and there is an ideal range of time for students to respond to a problem. In [10], they further support this non-linear relationship by showing that including time as a quadratic predictor instead of linear yields to a better prediction of students' performance. These studies support the intuition that accounting for the activities in different ranges of response time can give a better prediction of student performance.

Other efforts have shown success in estimating time spent on the activities that occur within a single response time. These efforts involved decomposing response time. [32] presented a model that predicts student performance relying on estimation of activities that cannot be directly observed from the log data such as thinking about hints, entering an answer, and reflecting on the hints. The preliminary results of our previous work that decomposes response times in a reading comprehension domain also revealed that students may go through multiple cognitive states during a single response time period [35].

In this work, we aim to show that identifying time spent on different cognitive states within the response time will provide better predictions of student performance.

3. CORPUS AND MEASURES

The datasets used in our work are log data collected during two studies with an iPad application called EMBRACE [38]. EMBRACE is designed to help young dual-language learners improve their reading comprehension in English. The students read interactive story books divided into chapters and they answer 3 to 9 multiple choice questions about the text at the end of each chapter. Books consisted either of narrative stories or of informational texts. Students see the text they should read in a box and they press a button labeled "Next" at the bottom of the screen to move from one sentence to another. They also see images representing what is depicted by the text.

In the full versions of EMBRACE, students are asked to either imagine the highlighted sentences or move the images on the screen to enact these sentences. They can get feedback based on how they are moving the images. Some features that are in the full versions of EMBRACE are not provided in the control version. Students still see the images in this version as well as the highlighted sentences, however, the only actions that they can perform are tapping on words to hear their pronunciations, and pressing the "Next" button to move to the next sentence. In this work, we are particularly interested in the control version as it gives us a more restricted set of student actions, which better enables us to focus on the role response time plays in reading comprehension. In the control version, we use the following measures:

1. **Student performance:** The proportion of correctly answered questions at the end of the chapters.
2. **Response time:** The time spent between when the sentence is first loaded and when the student presses the 'Next' button to proceed to the next sentence.
3. **Help requests:** The frequency of tapping on an underlined word to hear its pronunciation in a sentence.

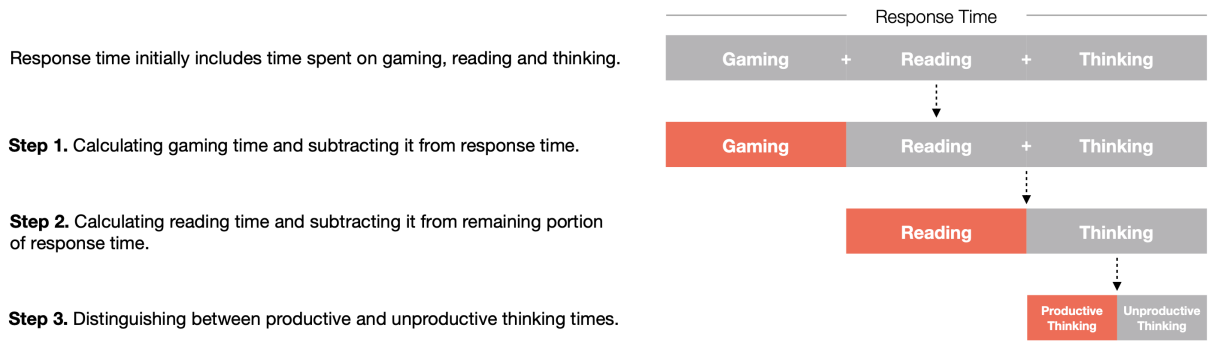

Figure 1: The protocol of designing productive vs. unproductive thinking portions in students learning.

4. **Frequency of gaming:** The "Next" button is disabled for 1 to 3 seconds depending on the length of the sentence to encourage the students to read the sentence completely. However, students might try to skip the current sentence and press this button while it is disabled. Frequency of this behavior within a sentence is our indicator of gaming since systematic and rapid actions to advance in the curriculum has been identified as gaming in previous research [3]. Note that this metric is not available in the second dataset.

5. **Decoding ability:** Decoding is defined as the ability to correctly pronounce written words. Our decoding measure is the student's score on the decoding part of Qualitative Reading Inventory (QRI) [18] that is in range $[0, 40]$.

6. **Sentence difficulty:** We used the Flesch-Kincaid readability grade level (FK) [21] to measure sentence difficulty. It is based on number of words in the sentence, and syllables in words. This measure represents the grade level required to understand a certain text. The difficulty of each chapter is calculated based on the average difficulty of the sentences in the chapter. FK is often used for long texts rather than single sentences. To confirm this measure is also appropriate for computing sentence level difficulty, we also computed chapter difficulty by applying FK on complete chapter texts. We did not observe a noticeable difference between computing sentence level difficulties per chapter($M = 4.81$, $SD = 1.39$) and applying FK on complete chapter texts ($M = 4.64$, $SD = 1.35$) as $RMSE = .37$.

In our datasets, data points are distinct student-chapter pairs as student performance can only be calculated in chapter level. The first dataset includes 22 students who are native Spanish speakers from second to fourth grade with mean QRI score 34.71 ($SD = 5.19$). One student is excluded from the dataset due to having scored less than 50% on the QRI test, and thus being unable to effectively use the application. We also excluded the first chapters of the books that were read out loud to the student by the application. Finally, some of the student-chapter pairs are excluded from the dataset due to logging errors such as unrealistic response times or not completing the chapter. In total, we had data from 21 students, and 716 distinct student-chapter pairs. The mean number of book chapters per student is 34.09 ($SD = 2.3$) with mean sentence difficulty 4.82 ($SD = 2.84$) across 7 story books.

Table 1: Descriptive statistics of time (in seconds) subcategories across student-chapter pairs in the first dataset (Spanish)

Measurements	Mean	SD	Min	Max
Reading Time	6.04	1.67	1.11	11.44
Productive Thinking	2.18	1.72	0	7.84
Unproductive Thinking	1.68	4.01	0	39.47
Gaming Time	0.21	0.44	0	5.36
Time Spent on Help	0.15	0.09	0.07	0.33
Time Spent on Sentence	10.12	6.02	2.75	51.5

Table 2: Descriptive statistics of time (in seconds) subcategories across student-chapter pairs in the second dataset (Mandarin)

Measurements	Mean	SD	Min	Max
Reading Time	4.40	0.65	2.67	6.12
Productive Thinking	2.68	1.07	0.01	4.09
Unproductive Thinking	1.45	3.08	0	27.79
Gaming Time	0.08	0.38	0	4.09
Time Spent on Help	0.84	0.60	0.06	3.61
Time Spent on Sentence	9.17	4.29	2.81	39.56

In the second dataset, collected from an earlier experiment, we had 24 native Mandarin speaker students from seventh to ninth grade with mean QRI score 37.42 ($SD = 1.79$). Only one student-chapter pair was excluded from the dataset as the student in that pair did not complete the assessment task for the chapter. In this dataset we had 479 distinct student-chapter pairs. The mean number of book chapters per student is 19.95 ($SD = 0.20$) with mean sentence difficulty 4.14 ($SD = 1.06$) across 4 story books.

4. RESPONSE TIME DECOMPOSITION

Figure 1 visualizes how we decompose response time at a high level. In the following subsections we describe how each time subcategory was computed in detail. The descriptive statistics of the time subcategories for the datasets are given in Tables 1 and 2.

4.1 Time Spent on Gaming

For each sentence, if the student never pressed "Next" when it was disabled, the gaming time on that sentence is 0. Otherwise, we first calculate how long the student waited after the last time they pressed "Next" when it was disabled until they actually passed the sentence. We calculated gaming time by subtracting this waiting time from total time spent on sentence. A student who waited for a long time to pass the sentence after pressing "Next" when it was disabled will have a low gaming time estimate. Note that, in the second dataset, since our gaming indicator was not available, we did not include gaming time in our analyses.

4.2 Time Spent on Reading

We first estimated how many words students should read per minute based on their grade according to [25]. For example, if a student is in third grade, they should be able to read between 120 to 170 words per minute. To give a more specific estimate for reading rate, instead of using only the student's grade, we include their ability in decoding English words and sentence difficulty, as students with a higher decoding ability may read more words. Similarly, in more difficult texts, students may read fewer words per minute. We first divide the normalized decoding score by the normalized sentence difficulty for each student and sentence pair. Let $[a, b]$ be the interval representing the possible values of this measure. We create another interval $[c, d]$, by getting the possible values of how many words students should read per minute within our students' grade levels from [25]. We simply map interval $[a, b]$ on interval $[c, d]$ using the linear mapping formula below:

$$f(x) = c + ((d - c)/(b - a)) * (x - a) \quad (1)$$

Here, x is one specific decoding/difficulty score for a student-sentence pair and $f(x)$ will give an estimate for how many words this student should read adjusted by the student's decoding ability and the difficulty of the sentence. Then, we simply calculated the time spent on reading for each sentence based on the student's reading rate and the word count in the sentence. For example, if a student is estimated to read 120 words per minute, their reading time estimate for a 6-word sentence is 3 seconds.

$$T_{\text{read}_{u,s}} = \frac{s_w}{u_{\text{wpm}_s}} * 60s \quad (2)$$

Here $T_{\text{read}_{u,s}}$ denotes the time estimate for student u to read sentence s, s_w denotes the number of words in sentence s and u_{wpm_s} denotes the rough estimate of the reading rate for student u while reading sentence s.

4.3 Time Spent on Help Requests

We computed the exact time it takes to play the help audios. Then we computed the time spent on help requests by multiplying the time it takes to play the tapped words by two as each word is played twice.

4.4 Time Spent on Thinking

Finally, we calculate the thinking time by simply subtracting gaming time, reading time and time spent on help requests from total time spent on one sentence.

$$T_{\text{think}_{u,s}} = T_{\text{total}_{u,s}} - (T_{\text{game}_{u,s}} + T_{\text{read}_{u,s}} + T_{\text{help}_{u,s}}) \quad (3)$$

Following this procedure, thinking time was estimated to be negative for 34% of the data points as the reading time estimate was higher than the total time spent on sentence. In that case, we simply adjust reading time estimate so that the time spent on sentence would be equal to reading time, and thinking time would be assigned to 0, which means that the time spent on sentence was devoted to reading and/or gaming. Even though zeroing-out negative thinking times seems to remove the variance that could be indicative of student performance, we did not observe any difference in terms of model performance. Moreover, doing so resulted in thinking time estimates becoming more interpretable.

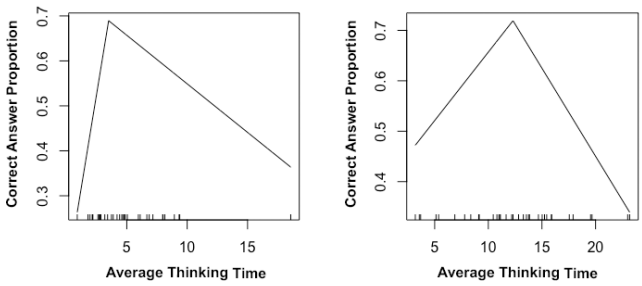

Figure 2: Example student-level thresholds for a high decoding (left) and a low decoding student (right).

4.5 Distinguishing Between Productive and Unproductive Thinking Time

To distinguish between productive and unproductive thinking, we use a data-driven method to find a threshold in thinking time for a student and chapter where spending more time on thinking after passing that threshold will be unhelpful. We first estimate that threshold at the student level and then similarly at the chapter level. We then combine the two thresholds to estimate one threshold for each student-chapter pair.

To find student level thresholds, using the `segmented` function in R [26, 27], we fit a separate piecewise regression model with our performance measure as the outcome and the mean time spent on thinking as the predictor for each student ($R^2 = 0.24$). There will be one breakpoint in thinking time for each student which will be independent of the chapter. Similarly, to estimate the thresholds at the chapter level, we fit one piecewise regression model with the performance measure as the outcome and the mean time spent on thinking as the predictor for each chapter (across all students) ($R^2 = 0.23$). The breakpoints represent the thresholds distinguishing between productive and unproductive thinking times for chapters. Figure 2 shows example thresholds returned from the piecewise regression models for a high decoding and a low decoding student, and Figure 3 shows example thresholds for an easy and a difficult chapter. High and low decoding students and easy and difficult chapters were decided based on median splits.

Although the separate thresholds we found are reasonable estimates, we do not use them directly when deriving the time spent on productive thinking, for two reasons. First, the threshold between productive and unproductive thinking time should be adjusted to both student and chapter characteristics in the same way that we adjusted time spent

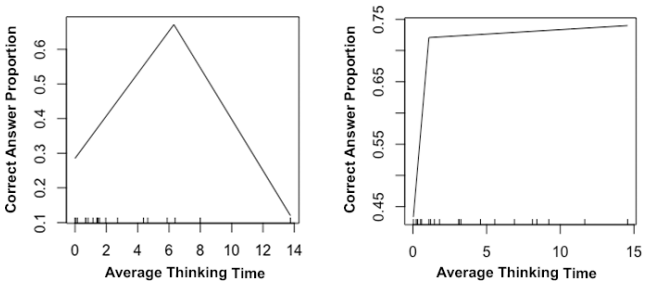

Figure 3: Example chapter-level thresholds for an easy chapter (left) and a difficult chapter (right).

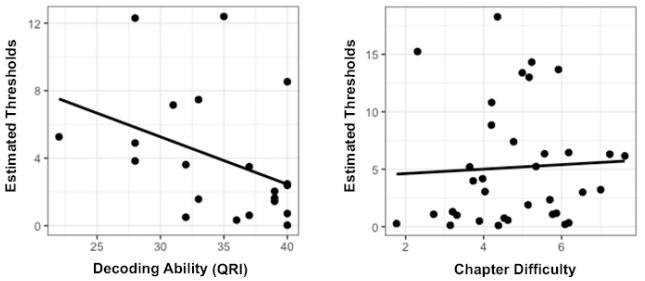

Figure 4: The relationship between decoding score and estimated student-level thresholds (left), and chapter difficulty and estimated chapter-level thresholds (right).

on reading. Second, we estimated these points by building a model which predicts the student performance, the variable that we would like to predict. Time spent on productive and unproductive thinking will be used as predictors in the model that we propose in this work. While extracting these features, leaking information from our outcome measure may cause overfitting in the final model. Therefore, we combine the thresholds found from separate regression equations. We build two separate linear regression models to predict the 'true' student and chapter level thresholds. Then we combine the two equations by taking their weighted average. This allows us to have one threshold estimate for an arbitrary student-chapter pair based on the decoding ability of the student and the difficulty of the chapter.

Figure 4 shows the relationship between QRI and 'true' student-level thresholds, and the relationship between chapter difficulty and 'true' chapter level thresholds. As seen in the figure, the estimated student-level thresholds are negatively correlated with decoding ability. This indicates that segregates productive and unproductive thinking regions occurs earlier for the students who scored better in the decoding test. The same figure also shows that chapter-level thresholds and chapter difficulties are far less correlated, which suggests that in estimating a threshold based on both student and chapter characteristics, the student characteristic (decoding score) is more important than the chapter characteristic (difficulty).

To combine these thresholds, we first find the equation for chapter difficulty and thinking time thresholds.

$$\lambda_{\text{chapter}} = \hat{B}_0 + \hat{B}_1 * \text{DIFF}_{\text{chapter}} \tag{4}$$

Here $\text{DIFF}_{\text{chapter}}$ denotes the chapter difficulty, and $\hat{B}_1$ and $\hat{B}_0$ are the estimated slope and y-intercept of the linear equation respectively. Similarly, we learn an equation for thinking time threshold in student level as follows:

$$\lambda_{\text{student}} = \hat{C}_0 + \hat{C}_1 * \text{QRI}_{\text{student}} \tag{5}$$

where $\text{QRI}_{\text{student}}$ denotes the student's decoding score, and $\hat{C}_1$ and $\hat{C}_0$ denote the estimated slope and y-intercept of this linear equation respectively. We combine these two separate thresholds by taking the weighted average of them. We weigh the equations by the correlation coefficient between the QRI score and the estimated student level thresholds (x), and the correlation coefficient between chapter difficulty and the estimated chapter level thresholds (y). We find the combined threshold as follows:

$$\lambda_{\text{combined}} = \frac{x * \lambda_{\text{chapter}} + y * \lambda_{\text{student}}}{|x| + |y|} \tag{6}$$

where $x = 0.05$ and $y = -0.39$. Using this equation, we have one estimate for thinking time threshold for a given student and chapter based on both decoding ability of the student and chapter difficulty.

Finally, productive thinking time is defined as time spent on thinking before this threshold. If the time spent on thinking is less than this threshold for a given student and chapter pair, all thinking was productive and time spent on unproductive thinking is 0. If the time spent on thinking is larger than the threshold, time spent on thinking until the threshold will be counted as productive thinking time and any time beyond the threshold will be counted as unproductive thinking time.

5. PREDICTING COMPREHENSION

The core hypothesis in our work is that dividing response time into subcategories in a way that could be indicative of some helpful and harmful cognitive states will improve predictive models of student performance. To test this hypothesis, we compared the proposed linear model (Decomposed RT) to two baselines: one that uses response time as whole (Baseline 1), and another that uses response time as a non-linear predictor (Baseline 2) to show that we are not simply accounting for non-linearity in response time but we show identifying the states within response time will help us predict comprehension more accurately. We report AIC [1] and BIC [31] to show the improvement in the model is not because of the increased number of predictors. Table 3 summarizes the feature sets we used in the 3 models we compare.We performed a cross-validation at the student level within a scheme for 50 iterations in which each time we left out a unique student pair from the whole procedure (decomposition of response time and training the models) and used their data for testing.

Table 4 shows the average RMSE, R^2, AIC and BIC values of the 50 iterations. For both datasets, we randomly flip the sign of the difference between paired model outcomes to conduct a paired-sample permutation test [37] to compare the mean of differences in evaluation metrics between our model and each baseline. We performed 1000 permutation trials in total. For the first dataset, we found significant improvements against both baselines in RMSE ($p < 0.005$), in AIC ($p < 0.001$), and in BIC ($p < 0.001$).

Table 3: Feature sets used in the models: Decomposed RT, Baseline 1, and Baseline 2. † indicates being a significant predictor ($p < 0.05$) of student performance in more than 80% of the folds.

Decomposed RT (Linear regression)	Baseline 1 (Linear regression)	Baseline 2 (Piecewise regression)
1. Frequency of gaming[1] 2. Frequency of help requests 3. Chapter difficulty † 4. Student decoding score † 5. Time spent on reading † 6. Time spent on gaming[1] 7. Time spent on productive thinking † 8. Time spent on unproductive thinking	1. Frequency of gaming[1] † 2. Frequency of help requests † 3. Chapter difficulty † 4. Student decoding score † 5. Time spent on sentence †	1. Frequency of gaming[1] † 2. Frequency of help requests 3. Chapter difficulty † 4. Student decoding score † 5. Time spent on sentence † *(as the non-linear parameter)*

Table 4: Comparison of evaluation metrics for proposed model and baselines on the first (Spanish) and second (Mandarin) dataset

	Model	RMSE	R^2	AIC	BIC
First Dataset	Decomposed RT (Linear regression)	.267 (.027)	.220 (.013)	80.242 (14.441)	124.988 (14.446)
	Baseline 1 (Linear regression)	.275 (.034)	.160 (.013)	122.399 (16.785)	153.721 (16.796)
	Baseline 2 (Piecewise regression)	.273 (.031)	.193 (.012)	100.868 (16.259)	141.139 (16.276)
Second Dataset	Decomposed RT (Linear regression)	.268 (.034)	.131 (.014)	58.350 (10.521)	91.027 (10.521)
	Baseline 1 (Linear regression)	.274 (.034)	.093 (.015)	73.090 (10.301)	97.599 (10.301)
	Baseline 2 (Piecewise regression)	.271 (.033)	.133 (.012)	57.508 (10.067)	90.185 (10.068)

For the second dataset, while decomposing response time, we made two adjustments. Firstly, since the students in this dataset were older (from 7th to 9th grade), their reading rates were adjusted for their grade level when calculating reading time. Secondly, the version of EMBRACE that was used to collect this data was not tracking when the students were pressing the "Next" button when it was disabled. Therefore, we discarded gaming time from our model. The remaining subcategories are calculated the same way as we did in the first dataset. The improvement in RMSE was significant against both Baseline 1 ($p < 0.001$) and Baseline 2 ($p < 0.005$). The improvement in AIC and BIC was significant against Baseline 1 ($p < 0.001$) while Baseline 2 was significantly better than Decomposed RT ($p < 0.05$).

Overall, Decomposed RT outperformed both baselines both in prediction error and the model fit criteria in the first dataset. However in the second dataset, although we see an improvement in prediction errors in favor of Decomposed RT, Baseline 2 had significantly better AIC and BIC values than Decomposed RT.

6. CONCLUSION

Within this paper, we proposed a new methodology to decompose response time so that time spent on gaming the system, productive thinking, and unproductive thinking states within a single response time can be accounted. Results showed that, using the time spent on these states as separate predictors rather than using response time as a whole gave better predictions of student performance. Comparison against another baseline that employs response time as a non-linear predictor also revealed that the improvement was not due to addressing the non-linearity in response time, and using the decomposition of response time to explain how much time was spent on different cognitive states indeed yielded better predictions. Moreover, comparison of AIC and BIC values supported that the improvement in the predictions were not due to introducing more predictors. However, we could not observe the same results for AIC and BIC between the proposed model and the non-linear baseline on the Mandarin dataset. A possible explanation is that we were not able to estimate the time spent on gaming in this dataset, thus time estimates for the other states were not as accurate as in the first dataset.

There are several other limitations of the work that need to be noted. Firstly, our estimation of reading time might not be the most accurate as there may be more factors influencing reading time than we addressed such as frequency of words and familiarity with the topic. Secondly, our model does not distinguish between the unproductive thinking behaviors (mind wandering and wheel-spinning) in its current stage. We plan to further explore how we can capture these different kinds of unproductive thinking.

In conclusion, we proposed a new method to use response time as a predictor in student modeling. The results show a promising improvement in predictive models of student performance when response time is decomposed into subcategories that can be indicative of the possible cognitive states students engage in. Future work should further assess this method's generalizability to different student profiles and different domains.

7. ACKNOWLEDGMENTS

We thank Arthur Glenberg and Maria Adelaida Restrepo for their helpful suggestions. This work was supported by the National Science Foundation Award No. 1324807.

[1]This measure is available only in the first dataset (Spanish).

8. REFERENCES

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[2] R. S. Baker, A. T. Corbett, and K. R. Koedinger. Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems*, pages 531–540. Springer, 2004.

[3] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner. Off-task behavior in the cognitive tutor classroom: when students" game the system". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 383–390, 2004.

[4] R. S. Baker, S. M. Gowda, A. T. Corbett, and J. Ocumpaugh. Towards automatically detecting whether student learning is shallow. In *International Conference on Intelligent Tutoring Systems*, pages 444–453. Springer, 2012.

[5] J. Beck and M. M. T. Rodrigo. Understanding wheel spinning in the context of affective factors. In *International conference on intelligent tutoring systems*, pages 162–167. Springer, 2014.

[6] J. E. Beck. Using response times to model student disengagement. In *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments*, volume 20, 2004.

[7] J. E. Beck and Y. Gong. Wheel-spinning: Students who fail to master a skill. In *International conference on artificial intelligence in education*, pages 431–440. Springer, 2013.

[8] M. T. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, 13(2):145–182, 1989.

[9] I.-A. Chounta and P. Carvalho. Will time tell? exploring the relationship between step duration and student performance. International Society of the Learning Sciences, Inc.[ISLS]., 2018.

[10] I.-A. Chounta and P. F. Carvalho. Square it up! how to model step duration when predicting student performance. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 330–334, 2019.

[11] R. S. d Baker, S. M. Gowda, and A. T. Corbett. Towards predicting future transfer of learning. In *International Conference on Artificial Intelligence in Education*, pages 23–30. Springer, 2011.

[12] Y. Gong, J. E. Beck, N. T. Heffernan, and E. Forbes-Summers. The fine-grained impact of gaming (?) on learning. In *International Conference on Intelligent Tutoring Systems*, pages 194–203. Springer, 2010.

[13] K. Holstein, G. Hong, M. Tegene, B. M. McLaren, and V. Aleven. The classroom as a dashboard: co-designing wearable cognitive augmentation for k-12 teachers. In *Proceedings of the 8th international conference on learning Analytics and knowledge*, pages 79–88, 2018.

[14] S. Hutt, J. Hardey, R. Bixler, A. Stewart, E. Risko, and S. K. D'Mello. Gaze-based detection of mind wandering during lecture viewing. *International Educational Data Mining Society*, 2017.

[15] S. Hutt, C. Mills, S. White, P. J. Donnelly, and S. K. D'Mello. The eyes have it: Gaze-based detection of mind wandering during learning with an intelligent tutoring system. *International Educational Data Mining Society*, 2016.

[16] J. Johns and B. Woolf. A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 163. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[17] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[18] L. Leslie and J. S. Caldwell. *Qualitative reading inventory*. Pearson, 2016.

[19] H. Li and W. Baer. Scaffolding adult learners' reading strategies in the intelligent tutoring system. In *Deep Comprehension*, pages 166–179. Routledge, 2018.

[20] C. Lin, S. Shen, and M. Chi. Incorporating student response time and tutor instructional interventions into student modeling. In *Proceedings of the 2016 Conference on user modeling adaptation and personalization*, pages 157–161, 2016.

[21] M. Lovric. *International Encyclopedia of Statistical Science*. Springer, 2011.

[22] K. S. McCarthy, C. Soto, C. Malbrán, L. Fonseca, M. Simian, and D. S. McNamara. istart-e: Reading comprehension strategy training for spanish speakers. In *International Conference on Artificial Intelligence in Education*, pages 215–219. Springer, 2018.

[23] E. Melis and J. Siekmann. Activemath: An intelligent tutoring system for mathematics. In *International Conference on Artificial Intelligence and Soft Computing*, pages 91–101. Springer, 2004.

[24] C. Mills, S. D'Mello, N. Bosch, and A. M. Olney. Mind wandering during learning with an intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 267–276. Springer, 2015.

[25] D. Morris. *Diagnosis and correction of reading problems*. Guilford Publications, 2013.

[26] V. M. Muggeo. Estimating regression models with unknown break-points. *Statistics in medicine*, 22(19):3055–3071, 2003.

[27] V. M. Muggeo et al. Segmented: an r package to fit regression models with broken-line relationships. *R news*, 8(1):20–25, 2008.

[28] K. Muldner, W. Burleson, B. Van de Sande, and K. VanLehn. An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts. *User modeling and user-adapted interaction*, 21(1-2):99–135, 2011.

[29] L. Paquette, A. M. de Carvalho, and R. S. Baker. Towards understanding expert coding of student disengagement in online learning. In *CogSci*, 2014.

[30] J. W. Schooler. Zoning out while reading: Evidence for dissociations between experience and metaconsciousness jonathan w. schooler, erik d. reichle, and david v. halpern. *Thinking and seeing: Visual metacognition in adults and children*, 203, 2004.

[31] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[32] B. Shih, K. R. Koedinger, and R. Scheines. A response time model for bottom-out hints as worked examples. *Handbook of educational data mining*, pages 201–212, 2011.

[33] J. Smallwood, D. J. Fishman, and J. W. Schooler. Counting the cost of an absent mind: Mind wandering as an underrecognized influence on educational performance. *Psychonomic bulletin & review*, 14(2):230–236, 2007.

[34] J. Smallwood and J. W. Schooler. The restless mind. *Psychological bulletin*, 132(6):946, 2006.

[35] Sonmez Unal, Deniz. Modeling student performance and disengagement using decomposition of response time data. In *EDM*. International Educational Data Mining Society (IEDMS), 2019.

[36] K. K. Szpunar, S. T. Moulton, and D. L. Schacter. Mind wandering and education: from the classroom to online learning. *Frontiers in psychology*, 4:495, 2013.

[37] H. van der Voet. Comparing the predictive accuracy of models using a simple randomization test. *Chemometrics and intelligent laboratory systems*, 25(2):313–323, 1994.

[38] E. Walker, A. Adams, M. A. Restrepo, S. Fialko, and A. M. Glenberg. When (and how) interacting with technology-enhanced storybooks helps dual language learners. *Translational Issues in Psychological Science*, 3(1):66, 2017.

[39] Y. Wang and N. T. Heffernan. Leveraging first response time into the knowledge tracing model. *International Educational Data Mining Society*, 2012.

[40] X. Xiong, Z. A. Pardos, et al. An analysis of response time data for improving student performance prediction. 2011.