Will time tell? Exploring the relationship between step duration and student performance

Abstract: In this paper, we explore how the time students take to solve a problem is related to their success. Even though prior research indicates that students' response times can provide some indication regarding correctness, time is not consistently and broadly used when modeling students' performance. We aim to clarify the relationship between the step duration – that is, the time a student takes to carry out a step of a learning task – and the outcome of this step with respect to correctness for STEM-related courses. Then, we discuss our early findings, how they can be used to enhance student modeling and to provide meaningful and timely feedback to students.

Introduction

The multifaceted nature of the relationship between time to complete a task and performance make using time as a predictor variable for success challenging. This, in turn, makes providing just-in-time hints and scaffold in Intelligent Tutoring Systems (ITSs) a complex and unsolved question. The focus of this ongoing work is to clarify the relation between time-on-task and student performance for various types of learning activities, domains and student characteristics. The main contribution of this work is the development of statistical tools that allow us to predict student performance based on response time. As a first step, we explore how step duration (that is, the total length of time spent on a step of a learning task) relates to correctness for STEM-related courses that were facilitated by ITSs.

Research Hypothesis

This work builds on the hypothesis that a student who takes either too little time or too long to respond to a tutor's question or carry out a step of a learning task, will most likely be unsuccessful in that step – thus, there is no linear relationship between step duration and correctness. The rationale is that, on the one hand, a student needs a minimum amount of time in order to process the problem, retrieve appropriate information, and to construct a correct response. On the other hand, taking too long to carry out a step could indicate lack of background knowledge, failure to retrieve critical information, and inability to address the step. Our research hypothesis is that there is a time frame defined by a minimum and a maximum step duration (dt_{min} and dt_{max} respectively) in which a student will likely provide the correct answer (and thus the error rate will be low in the same interval). Consequently, every step that lies outside this time frame will most likely be solved incorrectly or not solved, and the error rate will be high. From now on, we refer to this time frame [dt_{min}, dt_{max}] as the Zone of Interest (ZOI). The concept of the Zone of Interest and the research hypothesis is depicted in Figure 1. Based on the research hypothesis, the error rate for the steps answered in less time than dt_{min} and more than dt_{max} will be higher than the error rate for the steps that were answered within these time thresholds. We propose a mechanism to identify this zone from students' response data, and to extrapolate their predicted performance based on where future response times fall in the distribution (inside the zone of interest or not).

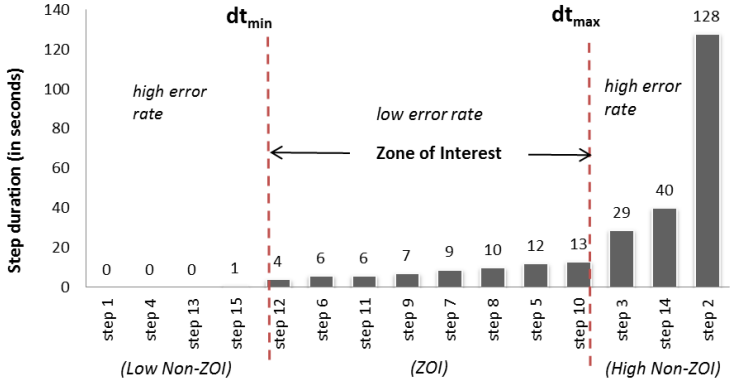

Figure 1. The Zone of Interest (ZOI) and the research hypothesis. The horizontal axis depicts the steps of a learning task and the vertical axis depicts the step duration. The steps on the horizontal axis are ordered with respect to their duration, from shorter to longer duration.

Related Work

Previous work has identified different characteristics of the relationship between response times and performance. Response time has been studied as a proxy of engagement. For example, Shih et al. (Shih, Koedinger, & Scheines, 2011) used response times as a proxy to distinguish when students use bottom-out hints to their benefit, and Beck (Beck, 2004) proposed the use of response times along with correctness to model students engagement in a task. Beck argues that hard-setting a time threshold and advising students to go faster or slower is counterproductive since it doesn't take into account personal characteristics. Instead, one should ask whether the student is engaging in the task rather than whether the student spends time on the task.

Response times have also been directly related to performance. For example, Miller et al. (Miller, Lasry, Lukoff, Schell, & Mazur, 2014) studied the relation between response times and performance for a classroom response system for Conceptual Physics. The authors pinpoint three main findings: a) response time for correct answers was faster than for incorrect answers b) background knowledge and self-efficacy affect response times (good background knowledge and high self-efficacy relates to fast response times and c) gender does not relate to response rates.

However, other work (Xiong & Pardos, 2011), was not successful at predicting future performance using a model that took into account past response times, even though, using response time and related features led to small improvement for student performance prediction when response time of the predicted action was known. Xiong noted that they did not identify a clear trend between response time and correctness. Similarly, Lin et. al. (Lin, Shen, & Chi, 2016) incorporated response times in BKT models and explored whether this addition would lead to better performance in terms of prediction accuracy in next-step's performance. The results suggested that response time can potentially be a good predictor for post-test scores but does not always support predicting performance per step. Thus, it is still not clear what would be a "good" response time - that is a response time that indicates the student is knowledgeable about the task - or a "bad" response time - that is a time that indicates that the student either is not interested in the activity or does not have the required background knowledge to address it - and how to model it.

Method of the study

Methodology

In this work, we operationalized the Zone of Interest (ZOI) using the duration of steps that students had to carry out while engaging in various learning activities. In particular, each learning activity consists of multiple steps. In order to complete a learning activity, a student has to go through all the steps of the activity. Each step is characterized by a duration - the time elapsed while the student reads the task, contemplates and provides an answer - and an outcome, that is whether the student performed this step correctly or not. For each activity and for every student we collected a set of duration times per step (step duration) and outcomes. Then, we computed the standard score (z-score) of the steps duration per student per activity. Here, the ZOI is defined as the zone between one standard deviation below and one standard deviation above the mean step duration of each student. From the definition of standard score it follows that the ZOI will make up for the 50% of the area under study while the area outside the ZOI will make up for the rest 50%. Next, we analyzed the performance of students with respect to the Zone of Interest. We computed and compared the error rates per student and per activity inside and outside the Zone of Interest.

Based on our research hypothesis, students will most likely carry out a step correctly when the time spent on this step (step duration) falls within the Zone of Interest. On the contrary, students will most likely carry out a step unsuccessfully when the time spent on this step falls outside the Zone of Interest. In other words, the error rate outside the Zone of Interest should be smaller than the error rate inside the Zone of Interest. We operationalize the ZOI in this way for two reasons: a) to address potential imbalance between correct and incorrect steps that may occur in the dataset and b) to assist the choice of thresholds (dt_{min}, dt_{max}) by using ±1 SD.

Dataset

We use five different datasets to test the research hypothesis. These datasets were collected from science courses that were supported by intelligent tutoring systems. An overview of the courses and a short description of the learning activities in these courses are given in Table X. All courses were STEM-related and they mostly aimed at problem-solving activities. All datasets are shared via the online repository *Datashop* (Koedinger et al., 2010). From these datasets we only used steps that were identified as correct or incorrect (excluded hints and unidentified steps).

Course Name	Domain	Description of the dataset
Fractions (2012)	Math	Tasks Description: Identifying and constructing fractions using graphical representations / 77 students / 25130 transactions: 73% correct, 27% incorrect
Geometry Area (1996/1997)	Geometry	Tasks Description: Geometry problem solving activities / 59 students / 6778 transactions: 80% correct, 20% incorrect
Stoichiometry (2005-2007)	Chemistry	Tasks Description: Problem solving tasks on Stoichiometry / 505 students/ 161257 transactions: 76% correct, 24% incorrect
Real Genetics (2015)	Genetics	Tasks Description: complex problem solving activities across a wide range of genetics topics / 125 students / 159714 transactions: 82% correct, 18% incorrect
Introductory Physics (2010)	Physics	Tasks Description: problem-solving tasks on Introductory Physics for USNA (Electricity, Magnetic Fields, Optics, etc) / 77 students / 97241 transactions: 70% correct, 30% incorrect

Analysis and results

To study the research hypothesis, we have computed the error rates for student steps inside and outside the ZOI, as described in the Methodology section. The error rate in the ZOI was computed as the number of incorrect steps in the ZOI over the total number of correct and incorrect steps in the ZOI. Similarly, the error rate outside the ZOI was computed as the number of incorrect steps outside the ZOI over the total number of correct and incorrect steps outside the ZOI. Furthermore, we studied the error rates in each one of the areas that lie on the left and right side of the ZOI. To compare the error rates we used the Wilcoxon signed-rank test. For all courses, the differences in error rates were statistically significant (see Table 1).

Table 1. Error rates (%) for student steps inside and outside the ZOI and W, p values for the Wilcoxon signed-rank test

Course name	Error Rate (%) ZOI	Error Rate (%) Non-ZOI	W	p
Physics	26	38	18385838	7.46E-16
Fractions	18	34	183826	6.44E-28
Geometry	18	25	328049	0.000188
Real Genetics	22	34	1952478	1.04E-23
Stoichiometry	21	25	22743389	5.72E-28

Table 2 Error rates (%) for the areas outside the ZOI: the Low Non-ZOI extending on ZOI's left and the High Non-ZOI extending on the ZOI's right.

Course name	Low Non-ZOI Error Rate (%)	High Non-ZOI Error Rate (%)
Physics	21	41
Fractions	27	34
Geometry	6	29
Real Genetics	20	34
Stoichiometry	12	26

As it can be seen from Table 2, students have more errors when taking long to respond than when taking short periods. Importantly, the error rate for short durations is non-zero, suggesting that these errors play an important role in the relationship between step duration and correctness. Moreover, when considering the error rate in the ZOI, we see that there is not a consistent linear relationship between increasing durations and error rate.

Table 1. Mean step duration (in seconds) for correct and incorrect steps inside and outside the ZOI

Course name	Mean duration for correct steps in ZOI	Mean duration for incorrect steps in ZOI	Mean duration for correct steps in Non-ZOI	Mean duration for incorrect steps in Non-ZOI
Physics	12.6	12.7	40.6	39.3
Fractions	4.3	3.3	33.7	7.3
Geometry	4.9	3.7	15	14
Real Genetics	2.5	3.4	15	14
Stoichiometry	6	6.2	61.4	10.5

Error rate inside the ZOI is statistically significantly lower (roughly 10% on average) than the error rate outside the ZOI. This suggests that when the step duration is either too fast or too long then the student is more likely to

fail this step. Moreover, the error rate in the Low Non-ZOI area was lower than the error rate in the High Non-ZOI area (roughly 15% on average). This indicates that students were more likely to fail a step when they were taking too long instead of being too fast.

Discussion

Our results contribute to the description of the relationship between time spent completing an activity and performance in that activity. Steps with durations that fall inside what we called the Zone of Interest are more often solved correctly than steps with durations that fall outside of that zone. Determining this zone takes into account the distribution of step durations for each student and, therefore, accounts for differences in reading time and other individual characteristics. Defining the zone of interest for each student is fundamental to characterize when a step might be solved incorrectly because the relationship between correctness and step duration is not linear. When students are too fast or take too long to complete a step, they are less likely to complete it correctly (and vice-versa). These results suggest that steps that are below the average step duration for similar students and content might not have been processed successfully. For example, students might not have read the whole problem or tried to retrieve critical previous information (Heckler, Scaife, & Sayre, 2010). Similarly, step duration above the average step duration for similar students and content might indicate that students were not able to access or construct the necessary information to successfully solve the problem (Miller et al., 2014). This has clear implications to ITS development: when a student is too quick she should be encouraged to further work on the step, and when a student takes too long she should be provided with a hint or another scaffold that allows her to succeed. In future work we envision creating a model to predict student performance based on step duration using the ZOI approach (instead of modeling time as continuous predictor). It is also important to test the generalizability of the ZOI approach and the model to new areas outside STEM and ITS, for example, language and social sciences in online courses. Finally, it is fundamental to test a successful model in a novel ITS that provides students with scaffold and feedback based not only on their historical accuracy, but also their response time.

References

Beck, J. E. (2004). Using response times to model student disengagement. In *Proceedings of the ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments* (pp. 13–20).

Heckler, A. F., Scaife, T. M., & Sayre, E. C. (2010). Response times and misconception-like responses to science questions. In *Proceedings of the Cognitive Science Society* (Vol. 32).

Koedinger, K. R., Baker, R. Sj., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of Educational Data Mining*, *43*.

Lin, C., Shen, S., & Chi, M. (2016). Incorporating Student Response Time and Tutor Instructional Interventions into Student Modeling. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 157–161). ACM.

Miller, K., Lasry, N., Lukoff, B., Schell, J., & Mazur, E. (2014). Conceptual question response times in peer instruction classrooms. *Physical Review Special Topics-Physics Education Research*, *10*(2), 020113.

Shih, B., Koedinger, K. R., & Scheines, R. (2011). A response time model for bottom-out hints as worked examples. *Handbook of Educational Data Mining*, 201–212.

Xiong, X., & Pardos, Z. A. (2011). An analysis of response time data for improving student performance prediction. In *Proceedings of the KDD 2011 workshop: Knowledge discovery in educational data* (pp. 103–108)

Acknowledgments (use Heading 1)

For this research we used the following datasets accessed via DataShop (pslcdatashop.org): the *'Geometry Area (1996-97)*' dataset, the *'REAL Genetics Study Data 2015*' dataset, the *'Fractions Lab Experiment 2012 - Classroom study 2013*' dataset, the *'USNA Introductory Physics Spring 2010'* dataset, the *'Pittsburgh Science of Learning Center Stoichiometry Study 1*' dataset.