Examining Student Effort on Hint through Response Time Decomposition

by

Ashish Gurung

A Thesis Proposal
Submitted to the Faculty
of
WORCESTER POLYTECHNIC INSTITUTE
In partial fulfillment of the requirements for the
Degree of Master in Science
in
Computer Science
April 2021

APPROVED:

_____
Professor Neil T. Heffernan, Master's Thesis Advisor

_____
Professor Joseph E. Beck, Thesis Reader

_____
Professor Craig E. Wills, Head of Department

ABSTRACT

Many teachers have come to rely on the affordances that computer-based learning platforms offer in regard to aiding in student assessment, supplementing instruction, and providing immediate feedback and help to students as they work through assigned content. Similarly, researchers commonly utilize the large datasets of clickstream logs describing students' interactions with the platform to study learning. For the teachers that use this information to monitor student progress, as well as for researchers, this data provides limited insights into the learning process; this is particularly the case as it pertains to observing and understanding the effort that students are applying to their work. From the perspective of teachers, it is important for them to know which students are attending to and using computer-provided aid and which are taking advantage of the system to complete work without effectively learning the material. In this paper, we conduct a series of analyses based on response time decomposition (RTD) to explore student help-seeking behavior in the context of on-demand hints within a computer-based learning platform with particular focus on examining which students appear to be exhibiting effort to learn while engaging with the system. Our findings are then leveraged to examine how our measure of student effort correlates with later student performance measures.

ACKNOWLEDGMENTS

I want to thank my advisor, Professor Neil T. Heffernan. His guidance, encouragement, and support throughout the years have been instrumental in the pursuit of my personal and career goals. I would also like to thank my reader, Professor Joseph E. Beck. Finally, I would like to thank Professor Anthony Botelho for his constant support and encouragement. Anthony has been a guru, and more importantly, a friend.

Examining Student Effort on Hint through Response Time Decomposition

1 INTRODUCTION

Computer-based learning platforms guide students' learning through the implementation of various principles of learning and cognitive sciences. Learning platforms have adopted differing approaches in supporting learners' needs through varying degrees of student- or instructor-paced approaches in determining the content presented to students. In the self-paced paradigm, the systems determine the sequence, and often the difficulty, of content that is presented to the student based on demonstrated performance and mastery of the material; conversely, instructor-paced systems rely on the instructor to determine these assignment parameters. Despite these differences, both of these learning system designs rely on the system to supplement the instruction and provide additional aid to students as they work; this can simply be done through, for example, immediate correctness feedback, but many systems incorporate more involved instructional aids in the form of hint messages[1, 24, 33], scaffolding problems[34, 40], or other forms of explanations or worked examples. Although the implementation of self-paced and instructor-paced systems often differ, there is a significant overlap in the design principles between the two approaches as both utilize principles of learning sciences and cognitive sciences to enhance learning through these offered supports. These principles have been extensively researched, and various works have explored their effectiveness [25, 32].

Regardless of the learning system's design, there is an underlying assumption that is commonly made regarding student engagement with help provided by the platform. It is presumed that students, when requesting or offered help through the system, are attending to the delivered feedback and using this to learn effectively. While this assumption is likely true for a large population of students, there is certainly evidence that many students take advantage of computer-provided help to work through assignments without effectively learning the material [12, 26]. It is important for students to use help productively, and it

is similarly important for instructors to know which students are effectively learning that assigned material.

Our goal in this work is to explore student help-seeking behavior within a computer-based learning platform with a focus on identifying and examining students who are attending to hints they receive through the system. The purpose of this work is to explore this behavior toward the development of a measure of student effort, accounting for systemic differences in the format of help provided (e.g., text-based hint messages or video-based worked examples). In self-paced systems, such a metric could help the system more accurately assess student knowledge and deliver content appropriately [4], or otherwise help instructors monitor and assess student performance more effectively. In either scenario, a measure of student effort, particularly on the help they receive, can help in better understanding the behavior and deploying learning interventions that promote more productive help-seeking strategies.

Using data collected from students interacting with a learning system in real classrooms, we conduct a series of exploratory analyses based on Response Time Decomposition (RTD; c.f., [15, 38, 41]). We further use the findings of these analyses to explore the relationship between identified student help-seeking behavior and later student performance. In this way, this paper addresses the following research questions:

(1) Are students using hints appropriately as determined by the amount of time spent on problems?
(2) What is the relationship between time spent on hints and later performance?
(3) What is the relationship between the time spent on hints and the prior knowledge of a student?
(4) Does the relationship between *time spent and hint usage* help us explore user behaviour in Randomized Control Trials(RCT) that explored hint usage on learning platforms?

The remainder of this paper is structured as follows. Section 2 describes the related works in the field of Learning Analytics with a focus on student help-seeking behavior. Section 3 explains our theoretical framework that decomposes help usage by users and our hypothesis of the user's mental model that dictates the actions a user takes after receiving help. Section 4 describes the dataset used in this work and Section 5 breaks

down the exploratory analysis conducted to test if the data supports the cogency of our theoretical framework. We use our findings from the exploratory analysis to define user behavior in terms of effort. Section 6 builds on our findings from Section 5 and explores the relationship between effort and other performance metrics. Section 6, 7, and 8 examine our findings and their relevance to research areas in learning analytics to inform future directions.

2 BACKGROUND

Most, if not all, computer-based learning platforms log the actions (clickstream data) of all users interacting with the system. The actions of the students, coupled with measures of performance, are commonly used to generate reports that help teachers monitor student progress. Although these reports provide an overview of the learners' activity on a given problem set, often in aggregate, the reports provide only limited insight into the learners' engagement and learning behavior exhibited while working. Efforts in the learning analytics community have helped develop better reports and visualizations that describe several dimensions of student performance, and activity [13, 20]. In this way, developers have attempted to leverage learning analytics research to develop measures that provide finer-grained insights into student learning. Measures of partial credit, for example, help to inform teachers about their students' knowledge and performance beyond a simple binary correctness measure [37]. Similarly, developing measures of student engagement can better direct teachers' attention to the students in most need. Researchers have found that the real-time reporting of related measures help teachers spend more time with lower-performing students [20].

The study of help within computer-based learning platforms has similarly led to questions pertaining to the effectiveness of tutor-provided aid within such systems among the learning science and learning analytics communities [3, 6, 19]. In some cases, studies conducted into the role of on-demand help within learning platforms have provided us with valuable insight into help seeking behaviors and various design approaches and principles that can lead to a more effective usage of hints by users [2]; this has been supported, in part, through the study of help-seeking behavior exhibited by learners [1, 36]. Related to this, Researchers have previously studied the use of self-explanation strategies as a method of helping students engage with content [35], while others have

explored the format of help delivery through text-based and video-based feedback [24]. Similarly, researchers have explored the effect of hints versus explanation [17] on student learning, as well as the use of erroneous examples to encourage student engagement with help and learning in general [23]. Finally, there has been other noteworthy research conducted in the field considering how the source or authorship of computer-provided help impacts student learning and engagement [31, 39].

In many cases, these studies have concluded that the effectiveness of help varies greatly and depends on many factors, with perhaps the most prominent of these being the level of student engagement. Regardless of the type of help provided, format, or authorship (e.g., expert-authored versus crowdsourced), these supports cannot help a student who does not attend to and engage with the provided aid. In this way, previous works examining student engagement, or conversely a lack of engagement, are particularly relevant to the study of student help-seeking behavior. Most notably, perhaps, is the large body of work pertaining to the study of students who "game the system" [5, 12, 26, 27]. Commonly referred to simply as "gaming," this behavior is characterized by students who take advantage of aspects of the system to complete assignments rather than effectively learn the material. In the context of help, students may exhaust available hints [28] or other aids to be given the correct answer or to be given easier questions. Many have theorized and explored aspects that may cause students to disengage, including work pertaining to the study of student affect [9, 16, 23]. Building off these and similar ideas, some researchers have tried to use affect detection to effectively adjust teaching strategies for disengaged behaviour [21], and explore how affect and engagement relate to future student performance [11, 29].

It is clearly important to promote engagement among students and to similarly promote positive help-seeking strategies, but it is also the case that engagement and persistence is not always productive. The example of "wheel spinning" behavior (c.f., [7]), for example, illustrates the negative aspects of persistence. Wheel spinning is defined as a student's struggle to master a given skill despite being given multiple practice opportunities; practically speaking, wheel spinning as been previously defined as a student being unable to demonstrate understanding of a concept by answering three consecutive questions correctly by the tenth item on a mastery learning assignment [7]. In light of wheel spinning behavior, and in consideration of the many works referenced in this section, it is important to identify students who are truly struggling and where the computer-provided

help is failing to aid them. Toward this, it is the goal of this work to develop a measure of student effort as defined by engagement and attentiveness to assigned work. We seek to distinguish students who are applying effort from those who may appear to be exhibiting wheel spinning, but are, in actuality, not "spinning their wheels" in the context of computer-provided hints. This paper focuses its attention to the sub-action level, observing variations in time between requested help actions within a learning platform to examine these aspects of student learning and help-seeking behavior.

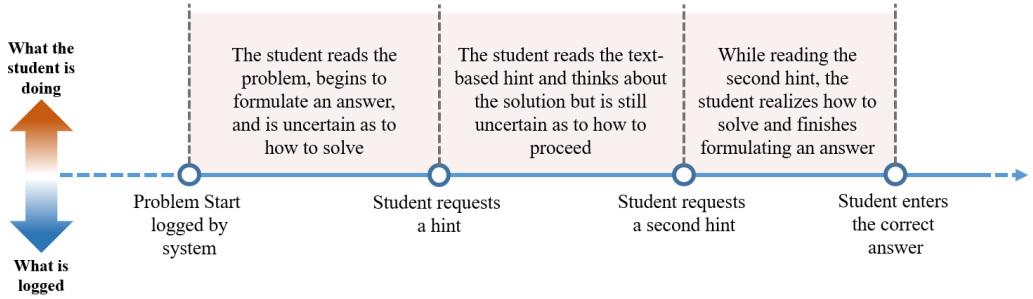

Fig. 1. Visual representation of the student behaviour for a user interacting with a Computer-based learning platform.

3 THEORETICAL FRAMEWORK BEHIND DECOMPOSITION OF HELP USAGE

While students work through assigned problems, regardless of the learning platform, there is a subtle disconnect between what is being logged and the learning processes taking place. It is certainly the case that actions logged by a learning system provide evidence to latent learning constructs (e.g., knowledge[10, 30]), as the actions were taken by a student and aspects of those actions (i.e., correctness) provide evidence of underlying cognitive and behavioral processes. However, these actions are not direct measurements of these latent attributes and must be viewed in conjunction with expectations as to what occurs between actions logged in the system to gain better insight into processes of learning.

Consider, for instance, the example illustrated in Figure 1. In this example, a student begins a problem in a learning system and is able to ultimately reach the correct solution after receiving help. From the perspective of the system, what is logged is just four actions: the start of a problem followed by two help requests, and an attempt to answer with

the correct solution. However, in that example, the actions themselves are not able to represent shifts in an activity that occurred external to the system. It is not, for example, able to capture when the student finished reading the question and began to think through how to formulate a response. We can hypothesize that the student was perhaps confused or lacked the knowledge to solve the problem in that the student requested a hint, but there is a large degree of uncertainty as to what the true reasoning for the action was in addition to the sequence of actions, behaviors, and thoughts that occurred external to the system between the start of the problem and the help request.

In order to measure these actions and behaviors, there are several approaches that can be explored. First, the use of additional sensors (such as video) or human observers can help record activity that occurs outside the learning system; such methods have previously been applied to study mind wandering [22] and student affect [8, 14], for example. These methods, however, can be potentially intrusive, expensive, and difficult to implement in classroom settings due to other ethical and privacy concerns. Another method is that of self-reports. By asking a student to reflect on their thought processes, we may be able to gain insights into aspects of the student's approach to problem-solving that was missed by the system. This method, however, can be potentially disruptive depending on when it is asked, or unreliable if the student is not able or not willing to articulate their approach with precision. The last method is the examination and analysis of data to make inferences of student activity based on the evidence provided through those actions that are logged and the time between them. While not as definitive as the other methods, as it is more difficult to externally validate many of the inferences made, this method can be applied post-hoc to large amounts of data without facing the concerns exhibited by the other two methods.

Given the actions logged by the system, coupled with the time between those actions, we hypothesize that we can gain insight into the productivity of student usage of help by decomposing the time spent after requesting help in a learning system. In the example illustrated in Figure 1, the student read through each requested hint and took the time to think through the new information as it related to formulating the correct solution; it is theorized that such students who are attending to the help would spend more time after the hint and would be more likely to answer the following, related problem correctly than a student who does not exhibit the same effort. By observing the response times in

conjunction with the following actions, we hope to gain this measure of effort, even if we are unable to specifically identify the specific latent processes exhibited beyond this valence metric.

4 DESCRIPTION OF DATASET

For our exploratory analyses, we collected a dataset[1] by randomly sampling 20,000 student-assignment interaction logs from ASSISTments [18] from the 2018-2019 and first half of the 2019-2020 school years (i.e., before the shift to remote learning in response to the COVID-19 pandemic). ASSISTments is a computer-based learning platform that allows teachers to assign content (primarily in the domain of middle-school mathematics) and monitors student progress, while supplying students with immediate correctness feedback and, on many problems, computer-provided help in the form of on-demand hints and scaffolding. Teachers are able to assign several types of assignments including a "complete all" that requires students to complete all assigned problems (similar to traditional paper-and-pencil assignments with the added benefit of computer supports), as well as "skill builder" assignments, which instead are mastery-based; skill builder assignments require students to demonstrate an understanding of the material by answering 3 consecutive questions correctly on the first attempt without the use of computer-provided aid. The data used in this paper observes both types of assignments but is primarily composed of skill builder work.

While working through assigned problems in ASSISTments, students are able to make multiple attempts to answer as well as receive aid by requesting help in the form of hints (available on many problems in the form of either text- or video-based messages and examples), or scaffolding questions that help break the problem into smaller steps. Problems may contain multiple hints which may be requested by the student, where, in all cases, the final "bottom-out" hint provides the student with the answer. Students are not able to move on to the next problem without eventually providing the correct answer.

The dataset contains the action logs from students who started work on the randomly-sampled assignments. Overall, the dataset contains 644,095 action logs from distinct 14,824 students working on problems across 6,569 problem sets that have a total of unique 36,441

[1]The data and code used in this work are made publicly available at http://tiny.cc/LAK21-28

problems. The difference between the total users and assignment logs indicates that we have records for users who did more than one assignment on the platform. The purpose of randomly sampling student-assignment interactions in this way was an attempt to create a sizeable dataset that is not based on a particular subset of content or groups of students; the selection of 20,000 such logs was an arbitrary decision, but we argue is sufficient to conduct the analyses and make impactful claims regarding the observed behaviors of students therein.

In our context, an action is logged every time a user interacts with the system. The system logs actions, for example, when the users start the assignment, start working on a problem, make an attempt, ask for help (as hint, explanation, or request for the correct answer), complete a problem and complete the assignment, among others (there are many system-level actions that can be taken describing a student ending a session and resuming, for example). Each action is accompanied by a timestamp to indicate when each action was taken by the user on the system. The dataset has a unique identifier for each individual user and each assignment as well as other descriptives incluing, for instance, the start and end time for each assignment. The dataset also has unique identifiers to represent the problem set and the problems the users are working on.

As we are interested in decomposing the amount of time a user takes between actions, we explore the data in regard to action pairs representing sequences of recorded actions; as exemplified in Figure 1, it is the goal of this work to take a step toward identifying processes that occur between actions and intend to use the observed time between actions as a means of addressing this goal. We first combined all the actions into pairs, denoted throughout this paper in the form "(first action, second action)" where these represent two consecutive actions taken within the session (i.e., we do not consider an action pair where the student logged out and resumed before continuing). Action pairs help us calculate the amount of time, in seconds, a user took after an action before taking the next action. While exploring the data, we discovered that the time a user took between first and second action ranged from close to 0 seconds to, in a small number of cases, more than an hour; as such, we applied a natural log-transform to the student response time to observe trends and relationships using the measure as an approximate-normal distribution.

4.1 Action pairs considered

As it is our goal to decompose student response time in regard to help-seeking behavior, we filtered the action pairs to include only those involving student help requests from the system. This work excludes the observance of scaffolding requests and instead focuses on hints within the system; as scaffolding problems may offer hints themselves, a deeper exploration of this type of aid is more complex and is planned as part of future work. Particularly, there are two notable types of hint requests that existed within the dataset: hints and explanations. The system defines these as separate forms of help, with hints often occurring in a series (i.e., there may be multiple hints), while explanations are singular and give the answer to the student following instruction or a worked example. We found, in our dataset, there were very few samples containing explanations, and fewer samples where the student actually requested such an explanation. As such, we further limited our analyses to explore only hint requests made within the system. We also excluded requests for the last hint in the sequence, referred to as the bottom-out hint, as this gives away the answer; we do not expect students to attend to the given answer in the same manner as a more-instructional hint, and therefore limit the scope of this work to focus specifically on non-answer-giving hints. Given this filtering to examine only hints, we will refer to help within the analyses described in this paper as "hints" to avoid conflating results with potential differences that may be examined in future works regarding other forms of help.

From this, we observe two primary types of action pairs, distinguished by the subsequent action taken after requesting a hint in the system. The intuition behind this is that students likely take additional time to formulate an answer when the subsequent action is an attempt as opposed to another hint request, or otherwise the response time is likely to incorporate different processes that lead to the different subsequent action. Namely, these action pairs are:

- (Hint Request, Attempt): The action pair (Hint Request, Attempt) represents all the instances when the user asked for a hint from the system, and the next action the user took after getting the hint was to attempt to answer the problem.
- (Hint Request, Hint Request): The action pair (Hint Request, Hint Request) represents all the instances when the user asked for a hint from the system, and the next action the user took was to ask for the next hint.

Table 1. filtered Action Pairs of students who asked for a hint

Action pairs	N
(Hint Request, Attempt)	808
(Hint Request, Hint Request)	414

In order to explore the theoretical framework behind decomposing help usage, we look at the instances when the user asked for a hint or multiple hints within the first 4 actions of working on a problem for both the action pairs. The action pair time represents the amount of time the user spent analyzing the hint before taking the second action in the action pair. We then z-scored the action pair time taken (again, represented as log-time) for each action pair and filtered the records with a value outside of the range (-3, 3); this filtering step is an attempt to remove very large outliers that may influence our results in unpredictable ways. The final resulting number of action pairs used in our analyses are shown in Table 1.

5 EXPLORATORY ANALYSES

In this section, we discuss the response time decomposition exploratory analyses conducted in examining student hint usage. As part of this, we examine not only differences in response time, but also explore potential systemic explanations for any differences observed (e.g., the format and length of hints requested). We used python for our analysis and the plots were generated using the Seaborn data visualization library. The y-axis in the charts of this section are the Kernel Density Estimation of the Gaussian distribution.

5.1 Analyzing action pairs

First, we observe student response time comparing the second action taken in regard to the first action that students take on the given problem. In other words, we hypothesize that students may use help differently depending on if they felt confident enough to attempt the problem before requesting a hint as opposed to requesting a hint as the first action on the given problem. As such, we observe first the time taken across all first actions and compare this to only the students who request a hint as the first action on the given problem.

5.1.1 Examining students across all first actions. We analyzed the two sets of action pairs by plotting the log-transformed distribution of the time taken across students exhibiting each of the action pairs. We found that the distribution of the (Hint Request, Hint Request) action pair to be distinctly bimodal in nature whereas the (Hint Request, Attempt) appeared to be closer to a unimodal distribution. Figure 2 shows the overlayed distribution of both action pairs.

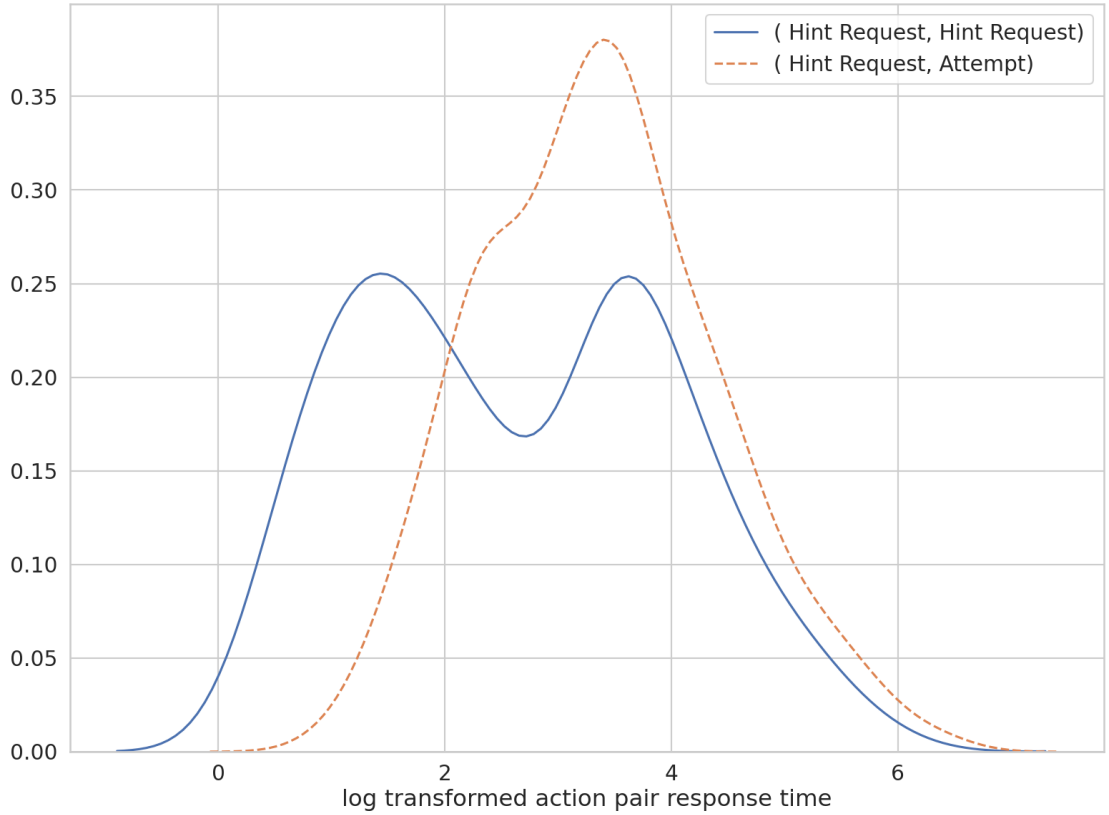

Fig. 2. distribution curve of (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs using natural log-transformed values of time taken for each action pair

The distribution illustrated in Figure 2 suggests that the users who ask for a hint and make an attempt to answer the question are similar to users spending more time on hints; we hypothesize that these students may be those who spend more time attempting to understand and appropriate the information given by a hint before taking a second action.

The alignment between the students spending more time on hints with those students who attempt an answer following a help request suggests that these students may be related in their usage of the hint; of course this claim cannot be verified from this plot alone, but does align with our theory that students who spend more time on help may be using that time productively to remedy gaps in knowledge. This also helps us intuit that users in the first half of the (Hint Request, Hint Request) action pair distribution (i.e., the left "peak" of the bimodal distribution) may not be devoting the same attention to the hint as those students spending more time; the cause of this is unclear, however, as it could suggest that these students are not reading or attending to the hint, but it could also suggest that these students are able to recognize that the hint is not helpful early and request a second hint in search of the information they need.

5.1.2 Examining students who request a hint first. In order to further refine our analysis, we also analyzed the response time for users whose first action after reading a problem was to ask for a hint. Figure 3 shows the normal distribution of both action pairs; we used the natural log-transformed values of the two pairs as that allows us to compare the two distributions. It is imporant to note that there are many similarities found between this and Figure 2, with the largest differences being seen in the shape of the (Hint Request, Attempt) distribution; we use the description of "differences" with hesitation here as there were very few meaningful differences between the two distributions.

Again, while subtle, the distributions depicted in Figure 3 show some variations. The (Hint Request, Attempt) action pairs distribution, for example, appears to be slightly smoother than was observed in Figure 2. this is rather unsurprising as we would expect observing the distribution of this subset of students would result in a smoother distribution, however, the smoothing shifts the mean of this distribution in favor of longer response times. This suggests that students who ask for a hint as the first action and make an attempt to answer as the second action, such students are spending more time on the requested hint. No such trend is observed for the students who are requesting multiple hints.

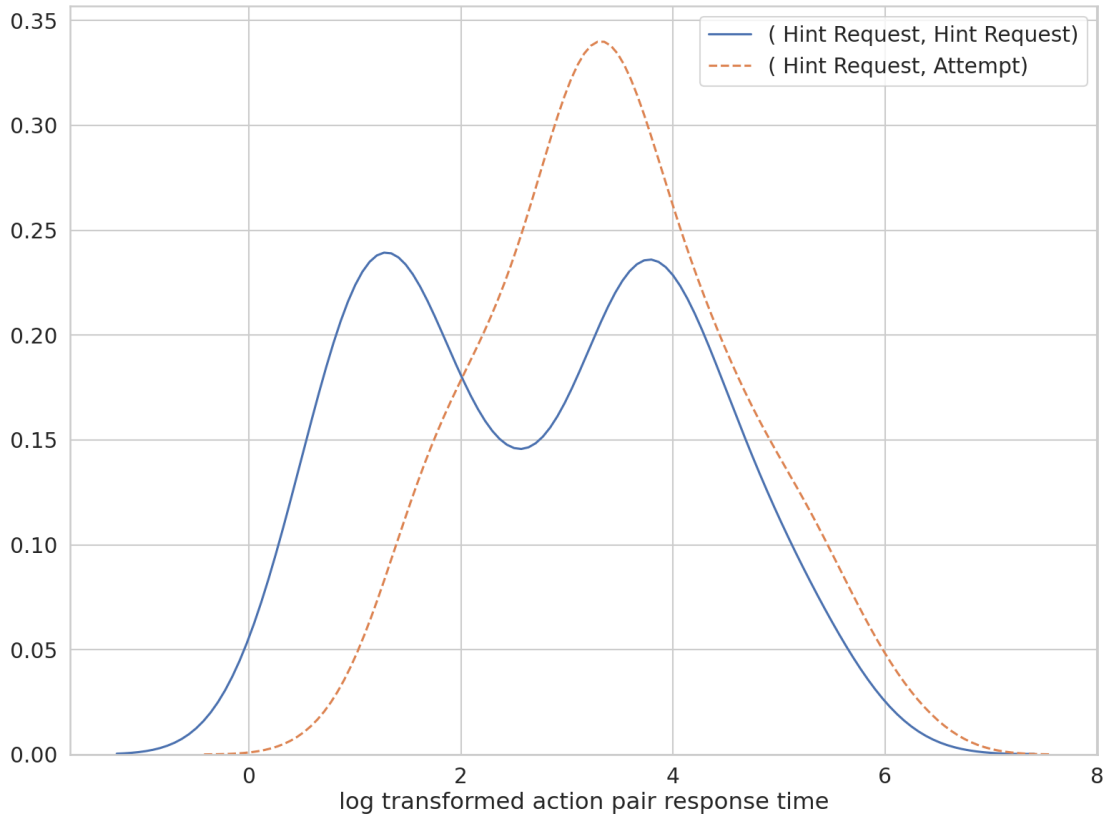

Fig. 3. distribution curve of (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs, when their first action was asking for hint after reading the problem, using natural log-transformed values of time taken for each action pair

5.2 Examining Potential Systemic Causes

In order to better understand our observations in regards to the response time during hint requests, we explore the existence of any potential systemic causes driving user behavior in both the (Hint Request, Attempt) and (Hint Request, Hint Request) action pairs. For the (Hint Request, Attempt) action pair, we also explored if the correctness/incorrectness of the user's subsequent attempt impacted the nature of the action pair's time distribution.

5.2.1 Video vs Text . The system can provide hints to a user as a text or video. We wanted to explore if the format of the hint influenced the amount of action pair time observed,

particularly examining whether this formatting could explain the bimodal distributions observed in the previous plots. Figure 4 shows another seemingly-bimodal distribution of the (Hint Request, Hint Request) action pair and the shape of the distribution when we only take text hints vs video hints; we used the log-transformed values of the two pairs as that allows us to compare the distributions as was conducted in the previous analysis.

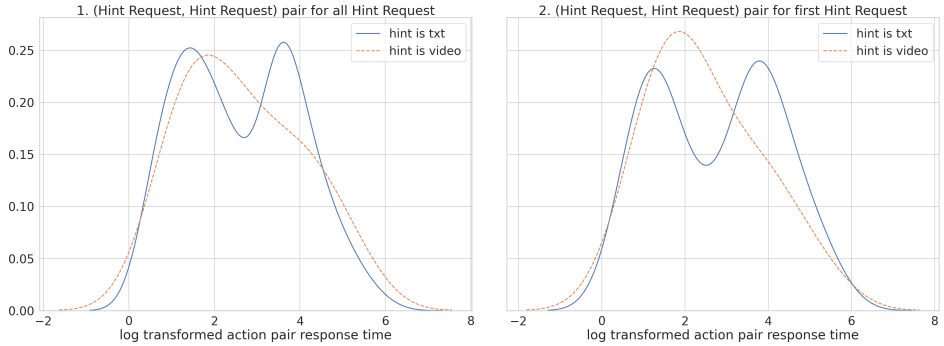

Fig. 4. There are too few instances of video hint for us to draw a conclusion but the data does seem to indicate that the type of hint does not influence the action pair response time

Figure 5 shows the normal distribution of (Hint Request, Attempt) action pair and the nature of the distribution when we only take text hints vs video hints; we used the two pairs' natural log-transformed values to compare the distributions.

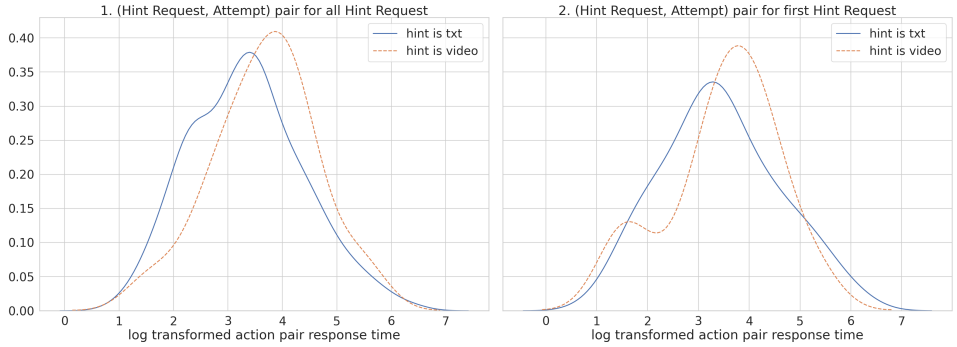

Fig. 5. There are too few instances of video hint for us to draw a conclusion but the data does seem to indicate that the type of hint does not influence the action pair response time

5.2.2 Correct Attempt vs Incorrect Attempt. In observing action pairs containing an attempt as the second action, we further examined if there were any meaningful differences in response time when the attempt was assessed to be correct as opposed to incorrect. Figure 6 shows the distribution of (Hint Request, Attempt) action pairs for these attempts. It can be seen in this figure that students tended to spend less time on incorrect attempts, but does not exhibit a large, meaningful difference; the distributions follow a nearly-unimodal shape despite the observed trend.

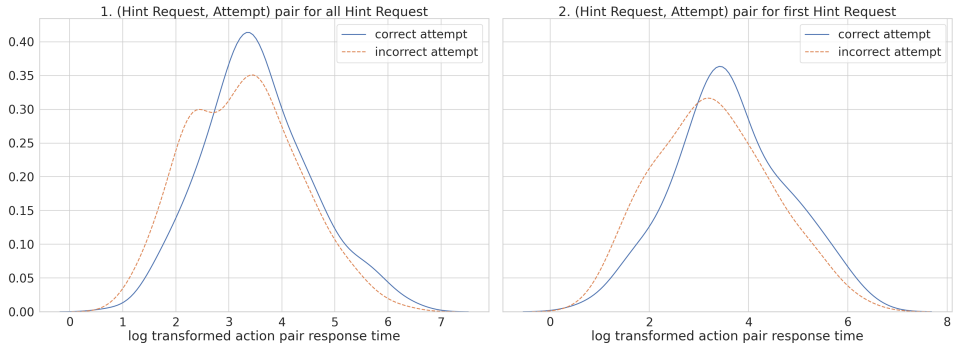

Fig. 6. The amount of time a user spends after getting a hint is the same for students who made a correct or an incorrect attempt

5.2.3 Other Explored Systemic Explanations. In addition to the systemic explanations explored above, we additionally examined the content of hints to observe whether the length and inclusion of visual components such as tables and mathematical formulae explained some of the differences in response time observed in the previous plots. These observations are summarized below; plots are not included for these due to spacing constraints.

(1) Length of Textual Hints: We analyzed the amount of time a user spent trying to understand a hint based on the length of the hint. The hints were divided into 4 quartiles based on the number of words per hint. We found users investing more time to understand hint when they were given a shorter hint i.e., hints with less than 18 words. We did not find a difference in the correctness of subsequent user attempts based on the length of the hints. While the length of hint did correlate with the amount of time spent after the request, the same bimodal distribution emerged

as before, suggesting that the length of hints did not explain away this observed difference.

(2) Tables and Formulae: We found some hints contained visual content such as tables and formulae. Performing a similar visual analysis comparing the response time for such cases, the resulting distributions did suggest that the inclusion of such content is correlated with higher observed response times, but, similar to the number of words in the hint, did not explain the previously-observed bimodal distributions. It is difficult to make stronger claims in regard to this finding, however, as the presence of tables and formulae in hints was too sparse in the data.

6 EXAMINING STUDENT EFFORT

6.1 Defining Effort

Our findings from the exploratory analysis, in the previous section, of the response time decomposition of users upon receiving help(hint) goes to support our theoretical model of user behavior. As the user response distribution for (Hint Request, Hint Request) action pair is bimodal in nature and the (Hint Request, Attempt) action pair distribution overlaps with the second peak of the bimodal distribution we use the information to formulate our definition of user exhibiting effort upon receiving help from a computer-based learning platform. In our theoretical model, we hypothesize that the amount of time a student spends on a problem trying to solve the problem is influenced by their understanding of the problem and the underlying concept the problem is trying to address. The amount of time they spend trying to understand the hint provided by the system is influenced by their understanding of the core idea behind the problem and the soundness of their mental model they formulated in order to solve the problem. A student sincerely trying to solve the problem would put in time understanding the hint, recalibrating their mental model to solve the problem, and decide if they have the answer or they need further help. Using the evidence from our analysis we hypothesize that the students in the first hump of the distribution for (Hint Request, Hint Request) action pair are not putting in the effort to understand the hint hence we define those users as exhibiting *"low-effort"* on the problem, the students in the second hump, we believe, put in the effort to understand

Table 2. the mean(μ) and standard deviation(σ) for the high and low effort clusters using Gaussian Mixture Modelling

	mean(μ)	standard deviation(σ)
Low-effort	1.7	0.757
High-effort	3.9	0.909

the hint and tried to formulate an answer using the hint hence we define those users as exhibiting *"high-effort"* on the problem.

6.2 Modeling Student Effort

The students exhibiting high-effort on both action pairs (Hint Request, Hint Request) and (Hint Request, Attempt) overlap on their time distribution for high-effort behavior; we merge our two action pairs into a single action pair (Hint Request, Action). As our primary interest is on decomposing user response to help and the amount of time a user spends unpacking the hint. As this distribution is bimodal in nature we apply Gaussian Mixture Models(GMM) to calculate the likelihood of the time spent by the student, understanding the hint, is part of the distribution of high-effort users, and the likelihood that the user is part of the distribution of low-effort users. GMM are a probabilistic model of representing a normally distributed subpopulation within an overall population. GMM is an unsupervised learning algorithm that uses Expectation Maximization to cluster the observations in a population into a subpopulation using probabilistic estimation that it is part of a subpopulation within the overall population. We clustered the bimodal distribution into two clusters using GMM; Table 2 shows the mean((μ)) and the standard deviation(σ) of the two clusters.

We now use the mean((μ)) and the standard deviation(σ) from the two clusters to calculate the area under curve for every response time if it were part of the low-effort distribution and the high-effort distribution. This provides us with insight into where the response time falls in the low effort distribution and high effort distribution if it were a user exhibiting low or high effort respectively. We realized that there were three major regions in the distribution where a user response time can fall. For the instances where the area under curve is less than 50 percent for low effort, we label them as low effort and

for the instances where the area under curve is larger than 50 percent for high effort we label them as high effort; however, for the instance that do not meet these requirements we can intuit the effort exhibited by the user but we cannot definitively say if they are exhibiting high or low effort so we did not label them.

6.3 Exploring the Relationship Between Effort and Performance Metrics

To explore the relationship between our measure of student effort and later performance metrics, we paired the action-level data used in previous analyses with both prior and later student performance measures. These additional measures include assignment completion, wheel spinning in the assignment (as defined by [7]), next problem correctness, prior percent correct (i.e. the percent of problems answered correctly by the student prior to each observed problem), and prior completion rates. We wanted to investigate if the students exhibiting effort perform better in the immediate next problem, if they are more likely to complete the assignment, and if they are more likely to exhibit wheel spinning during the assignment.

We used regression analyses to investigate the relationship between student effort and each of these outcome measures while controlling for prior completion rate, prior percent correct, and prior completion rate respectively. The observed models and results of our regression analysis are observed in Tables 3, 4, and 7, and are discussed further in the next section.

7 RESULTS

We trained a logistic regression to explore the relationship between effort and next problem correctness while controlling for prior percent correct; it is important to highlight, as this is a logistic regression, that the coefficients are reported in log-odds units and should therefore be interpreted in terms of their magnitude rather than in terms of standard deviations or percents as is commonly afforded by linear regression models. We found that the model (R^2 = 0.048) showed that low effort behaviour, B = -0.7053, p=0.4, was a significant predictor of next-problem correctness. This suggests that students exhibiting low effort are more likely to answer the next problem incorrectly. The same cannot be said for the students who are exhibiting high effort. The regression analysis is reported in Table 3. It is also important to note that the r-squared of the model is relatively low,

Table 3. Logistic Regression analysis exploring the relationship between effort and next problem correctness while controlling for prior percent correct (R^2 = 0.048)

	coefficient	std. err	conf. interval	p-value
intercept	-1.7747	0.360	[-2.481, -1.069]	0.000
High effort	-0.2652	0.271	[-0.797, 0.267]	0.328
Low effort	-0.7053	0.343	[-1.378, -0.033]	0.040
Prior percent correct	2.2975	0.615	[1.091, 3.504]	0.000

Table 4. Logistic Regression analysis exploring the relationship between effort and wheel spinning while controlling for prior completion (R^2 = 0.091)

	coefficient	std. err	conf. interval	p-value
Intercept	0.3809	0.387	[-0.378, 1.139]	0.325
High effort	-0.5815	0.301	[-1.171, 0.008]	0.053
Low effort	1.0741	0.294	[0.497, 1.651]	0.000
Prior completion	-1.8236	0.502	[-2.808, -0.840]	0.000

which, while it does not detract from our findings, suggests that there are other larger factors that we did not account for that explain the dependent variable (e.g., likely other skill- or content-based factors).

Table 5. Logistic Regression analysis exploring the relationship between effort and assignment completion while controlling for prior completion (R^2 = 0.104)

	coefficient	std. err	conf. interval	p-value
Intercept	-3.3584	0.484	[-4.307, -2.410]	0.000
High effort	0.3614	0.246	[-0.121, 0.844]	0.142
Low effort	-0.1617	0.296	[-0.741, 0.418]	0.584
Prior completion	3.6991	0.577	[2.569, 4.829]	0.000

We also examined the relationship between effort and wheel spinning behavior while controlling for prior completion. We found the model (R^2 = 0.091) found that low effort behavior, B = 1.0741, p < 0.001, was a significant predictor of wheel-spinning behavior. The analysis found that high effort behavior, B=-0.5815, p=0.053 was a strong indicator of wheel spinning behavior.This indicates that the students who are exhibiting low effort

on the problem are highly likely to wheel-spin during the assignment where as there is a strong indication that students in the students exhibiting high effort are less likely to wheel-spin. The regression analysis is reported in Table 4.

We also examined the relationship between effort and assignment completion while controlling for prior completion. We found the model (R^2 = 0.104) found neither high nor low effort to be significant predictors of assignment completion although there was an indication that high effort is a predictor for assignment completion. Here, we found that the students who exhibit high effort will likely complete the assignment however the findings were not significant. The regression analysis is reported in Table 7.

8 ANALYZING RCTS USING EFFORT

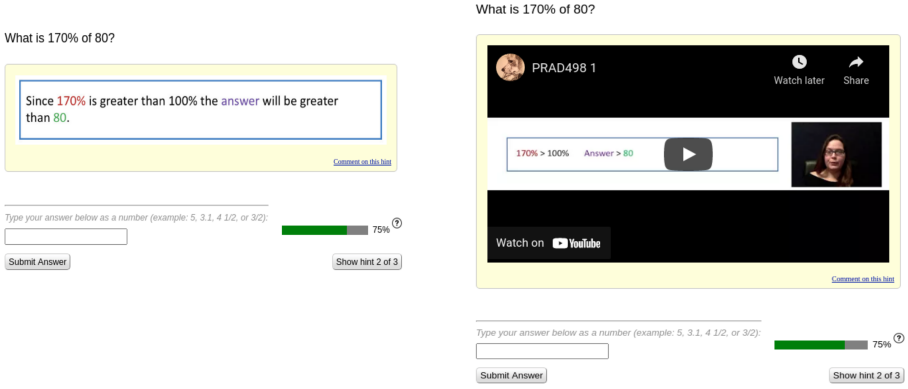

Fig. 7. Comparing the Video vs Text Hints

Prior work comparing the delivery of hints in different modalities, text vs. videos, has found video hints to be more helpful to students with lower prior knowledge [24]. Students with higher prior knowledge preferred text hints. There is an RCT currently running on ASSISTments since 2018 to replicate the research in [24]. The study explores the role of autonomy in helping students choose the modality for hints. Figure 8 shows the design of the study. The students were randomized into two conditions “Choice” and “No Choice”. The students in the “Choice” condition were allowed to choose between video and text hints. In contrast, students in the “No Choice” condition got randomized into video and text hints—certain schools using ASSISTmetns block YouTube in their

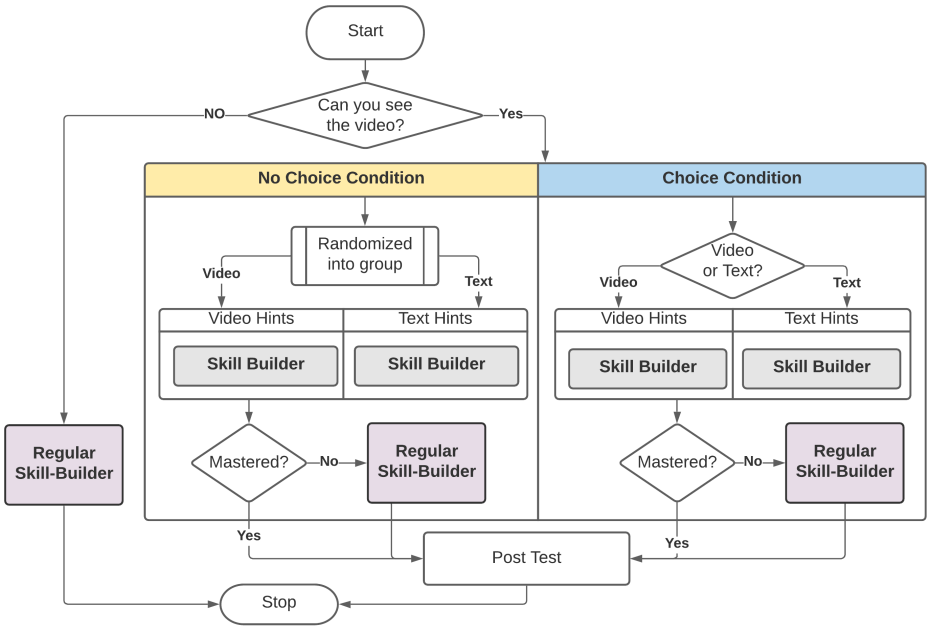

Fig. 8. Breakdown of the experimental design that explores the role of autonomy in choosing the modality for Hints i.e. Video vs Text

network to prevent misuse of school resources. The students were asked to do a video check before the randomization to filter out the students without access to YouTube from the experiment. The students who did not have access to the video at their school got a traditional Skill-Builder, whereas students who had access to the video got randomized into “Choice” vs. “No Choice”. After completing the Skill-Builder, the students had to take a Post-Test. The students in the study were assigned a custom Skill-Builder that had ten problems with either a video hint or a text hint. If the students could not exhibit mastery in the custom Skill-Builder, they were assigned the same traditional Skill-Builder as the students who were not randomized into “Choice” vs “No Choice”. Figure 7 has an example of the video and text hint that was provided to the students for the same problem.

From the students who got randomized, the dataset had 703 students who completed the post-test. There were 359 students in the “Choice” condition and 344 students in the “No Choice” condition. A detailed breakdown of the randomization per condition can be found in Table 7. It was fascinating that students were two times more likely to choose a

Table 6. Breakdown of the students per condition

Condition	N	Sub-condition	N
Choice	359	Video	112
		Text	247
No Choice	344	Video	171
		Text	173

text hint over a video hint. We explored if the students' affinity towards the hint type was motivated by their prior knowledge; however, they were still twice as likely to choose text over video hints, although the likelihood was slightly below two for low knowledge students.

Table 7. Linear Regression analysis exploring the relationship between effort and Post test percent correct while controlling for prior completion (R^2 = 0.158)

	coefficient	std. err	conf. interval	p-value
Intercept	-0.1673	0.076	[-0.316, -0.018]	0.028
C(Condition)[Choice]	0.0003	0.024	[-0.046, 0.047]	0.989
High effort	-0.0732	0.081	[-0.233, 0.086]	0.368
Low effort	-0.1328	0.118	[-0.364, 0.099]	0.260
Prior percent correct	-0.8860	0.109	[0.673, 1.099]	0.000
Total Problem Hints Used	-0.1156	0.022	[-0.158, -0.073]	0.000

We computed the response time for all the action pairs where the student asked for a hint. We used the GMM to determine the amount of effort a student put into understanding the hint. We conducted a regression analysis to compare the relationship between post-test performance, choice and effort while controlling for prior percent correct and total problem hints used. We found the model(R^2 = 0.158) found neither high nor low effort to be significant predictor for post test completion.

9 DISCUSSION AND FUTURE WORKS

Our analysis found that user behavior can be categorized into exhibiting low and high response times, which, in consideration of our exploratory analyses, we posit correspond to measures of high and low effort; we hypothesize from our findings that we are able

to identify students applying effort as evidenced by the time taken and aspects of their subsequent action. With this definition of our metric, We found low effort students to correlated strongly with wheel-spinning, even more so than the high effort students. This finding is a noteworthy contribution as it contradicts the intentional definition of wheel spinning behavior; many of the students exhibiting wheel spinning, in this way, appear to be spending little time and effort while working through their assigned work. We argue, and look to address in future work, that such students should not be considered as exhibiting wheel spinning and the definition of such behavior should be updated to consider these aspects of student work.

This work did not explore any interaction between effort and affect or other theories of behavior and engagement, but also may provide insights into student behavior across problems; the current analyses focuses at the sub-action level, and future works are planned to explore how our findings extend across an assignment. We are particularly interested in exploring the relationship between our measure of student effort and previously-developed measures of gaming behavior [25] while working on problems.

Other works have suggested that videos work better than hints in certain contexts [24], and future works intend to explore further if similar results may be better explained when accounting for student effort and attention devoted to the requsted help. Additionally, in the future, we want to investigate if the effect in such studies is mediated by indicators of effort.

Similarly, the development of student models may benefit from further insights into student effort and engagement. Cognitive models such as that of Knowledge Tracing [10], for example, rely on correctness and incorrectness of student actions for modelling knowledge state, and we intuit that using a more continuous measure of effort might improve the performance of these types of cognitive models.

We implore researchers and developers to use our findings and exploration of effort to develop better measures and reports for teachers that consider effort in the assessment of students. We strive, in future works, to develop externally-validated measures of student engagement and effort toward these goals.

10 CONCLUSION

This paper presents evidence that provides new insights into user behavior pertaining to student help-seeking behavior. User response time can be categorized into users exhibiting high-effort and low-effort in their hint usage before taking the next action. We conducted exploratory analyses that helped to eliminate obvious systemic and performance confounds and still found distinguishable groups of students by the time devoted to hint requests. The response time decomposition work is an essential step in quantifying student effort while working on a problem as teachers often rely upon the amount of effort a student exhibits in conjunction with the student's problem-level correctness scores in gauging student progress while working on their assignment.

We also explored the interaction between effort and wheel spinning as well as other student outcome measures. We found that lower effort students are highly correlated with wheel spinning behavior, contradicting the intended definition of the behavior; we argue that this is a significant finding as it attests to the fact that the definition of wheel-spinning needs further work as the current definition does not account for whether students are truly "spinning their wheels" by applying effort.

ACKNOWLEDGMENTS

We thank multiple NSF grants (e.g., 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, 1535428, 1440753, 1316736, 1252297, 1109483, & DRL-1031398), the US Department of Education IES (R305A170137, R305A170243, R305A180401, R305A120125, R305A180401, & R305C100024) and the Graduate Assistance in Areas of National Need program (e.g., P200A180088 & P200A150306), EIR, the Office of Naval Research (N00014-18-1-2768), Schmidt Futures and other anonymous philanthropy.

REFERENCES

[1] Vincent Aleven, Bruce McLaren, Ido Roll, and Kenneth Koedinger. 2004. Toward tutoring help seeking. In *International Conference on Intelligent Tutoring Systems*. Springer, 227–239.

[2] Vincent Aleven, Elmar Stahl, Silke Schworm, Frank Fischer, and Raven Wallace. 2003. Help seeking and help design in interactive learning environments. *Review of educational research* 73, 3 (2003), 277–320.

[3] Ivon Arroyo, Joseph E Beck, Carole R Beal, Rachel Wing, and Beverly P Woolf. 2001. Analyzing students' response to help provision in an elementary mathematics intelligent tutoring system. In

Papers of the AIED-2001 workshop on help provision and help seeking in interactive learning environments. Citeseer, 34–46.

[4] Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. 2014. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 387–426.

[5] Ryan SJd Baker, Albert T Corbett, and Angela Z Wagner. 2006. Human classification of low-fidelity replays of student actions. In *Proceedings of the educational data mining workshop at the 8th international conference on intelligent tutoring systems*, Vol. 2002. 29–36.

[6] Joseph E Beck, Kai-min Chang, Jack Mostow, and Albert Corbett. 2008. Does help help? Introducing the Bayesian Evaluation and Assessment methodology. In *International Conference on Intelligent Tutoring Systems*. Springer, 383–394.

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

[8] Nigel Bosch, Sidney K D'mello, Jaclyn Ocumpaugh, Ryan S Baker, and Valerie Shute. 2016. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 2 (2016), 1–26.

[9] Anthony F Botelho, Ryan S Baker, Jaclyn Ocumpaugh, and Neil T Heffernan. 2018. Studying Affect Dynamics and Chronometry Using Sensor-Free Detectors. *International Educational Data Mining Society* (2018).

[10] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[11] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (2004), 241–250.

[12] Ryan SJ d Baker, Albert T Corbett, Ido Roll, and Kenneth R Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction* 18, 3 (2008), 287–314.

[13] Dan Davis, Ioana Jivet, René F Kizilcec, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2017. Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In *Proceedings of the seventh international learning analytics & knowledge conference*. 454–463.

[14] Sidney D'Mello, Ed Dieterle, and Angela Duckworth. 2017. Advanced, analytic, automated (AAA) measurement of engagement during learning. *Educational psychologist* 52, 2 (2017), 104–123.

[15] Ehtibar N Dzhafarov and Richard Schweickert. 1995. Decompositions of response times: An almost general theory. *Journal of Mathematical Psychology* 39, 3 (1995), 285–314.

[16] Sidney D'Mello, Blair Lehman, Jeremiah Sullins, Rosaire Daigle, Rebekah Combs, Kimberly Vogt, Lydia Perkins, and Art Graesser. 2010. A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In *International conference on intelligent tutoring systems*. Springer, 245–254.

[17] Elizabeth Sarah Gould, Ruowang Li, and Steven James Southard. 2010. Effectiveness of Hints vs. Complete Explanation using ASSISTments. (2010).

[18] Neil T Heffernan and Cristina Lindquist Heffernan. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education* 24, 4 (2014), 470–497.

[19] Cecily Heiner, Joseph Beck, and Jack Mostow. 2004. Improving the help selection policy in a Reading Tutor that listens. In *InSTIL/ICALL Symposium 2004*.

[20] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. 2018. Student learning benefits of a mixed-reality teacher awareness tool in AI-enhanced classrooms. In *International conference on artificial intelligence in education*. Springer, 154–168.

[21] Blair Lehman, Melanie Matthews, Sidney D'Mello, and Natalie Person. 2008. What are you feeling? Investigating student affective states during expert human tutoring sessions. In *International conference on intelligent tutoring systems*. Springer, 50–59.

[22] Caitlin Mills, Sidney D'Mello, Nigel Bosch, and Andrew M Olney. 2015. Mind wandering during learning with an intelligent tutoring system. In *International conference on artificial intelligence in education*. Springer, 267–276.

[23] Michael Mogessie, J Elizabeth Richey, Bruce M McLaren, Juan Miguel L Andres-Bray, and Ryan S Baker. 2020. Confrustion and Gaming While Learning with Erroneous Examples in a Decimals Game. In *International Conference on Artificial Intelligence in Education*. Springer, 208–213.

[24] Korinn S Ostrow and Neil T Heffernan. 2015. The role of student choice within adaptive tutoring. In *International Conference on Artificial Intelligence in Education*. Springer, 752–755.

[25] John F Pane, Beth Ann Griffin, Daniel F McCaffrey, and Rita Karam. 2014. Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis* 36, 2 (2014), 127–144.

[26] Luc Paquette and Ryan S Baker. 2019. Comparing machine learning to knowledge engineering for student behavior modeling: a case study in gaming the system. *Interactive Learning Environments* 27, 5-6 (2019), 585–597.

[27] Luc Paquette, Ryan S Baker, Adriana de Carvalho, and Jaclyn Ocumpaugh. 2015. Cross-system transfer of machine learned and knowledge engineered models of gaming the system. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 183–194.

[28] Luc Paquette, Adriana de Carvahlo, Ryan Baker, and Jaclyn Ocumpaugh. 2014. Reengineering the feature distillation process: A case study in detection of gaming the system. In *Educational data mining 2014*. Citeseer.

[29] Zachary A Pardos, Ryan SJD Baker, Maria OCZ San Pedro, Sujith M Gowda, and Supreeth M Gowda. 2013. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the third international conference on learning analytics and knowledge*. 117–124.

[30] Zachary A Pardos and Neil T Heffernan. 2010. Modeling individualization in a bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 255–266.

[31] Thanaporn Patikorn and Neil T Heffernan. 2020. Effectiveness of Crowd-Sourcing On-Demand Assistance from Teachers in Online Learning Platforms. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*. 115–124.

[32] Jeremy Roschelle, Mingyu Feng, Robert F Murphy, and Craig A Mason. 2016. Online mathematics homework increases student achievement. *AERA open* 2, 4 (2016), 2332858416673968.

[33] Adam C Sales and John F Pane. 2018. Student Log-Data from a Randomized Evaluation of Educational Technology: A Causal Case Study. *arXiv preprint arXiv:1808.02528* (2018).

[34] James R Segedy, Gautam Biswas, Emily Feitl Blackstock, and Akailah Jenkins. 2013. Guided skill practice as an adaptive scaffolding strategy in open-ended learning environments. In *International Conference on Artificial Intelligence in Education*. Springer, 532–541.

[35] Benjamin Shih, Kenneth R Koedinger, and Richard Scheines. 2011. A response time model for bottom-out hints as worked examples. *Handbook of educational data mining* (2011), 201–212.

[36] Bram E Vaessen, Frans J Prins, and Johan Jeuring. 2014. University students' achievement goals and help-seeking strategies in an intelligent tutoring system. *Computers & Education* 72 (2014), 196–208.

[37] Yan Wang, Korinn Ostrow, and Neil Heffernan. 2016. Partial credit revisited: Enhancing the efficiency and reliability of group differentiation at scale. *Student Modeling from Different Aspects* 7, 6,314 (2016), 22.

[38] Robert Whelan. 2008. Effective analysis of reaction time data. *The Psychological Record* 58, 3 (2008), 475–482.

[39] Jacob Whitehill and Margo Seltzer. 2017. A Crowdsourcing Approach to Collecting Tutorial Videos–Toward Personalized Learning-at-Scale. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*. 157–160.

[40] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[41] Jeremy M Wolfe and W Gray. 2007. Guided search 4.0. *Integrated models of cognitive systems* (2007), 99–119.