A Response Time Model for Bottom-Out Hints as Worked Examples

Benjamin Shih, Kenneth R. Koedinger, and Richard Scheines
{shih, koedinger, scheines}@cmu.edu
Carnegie Mellon University

Abstract. Students can use an educational system's help in unexpected ways. For example, they may bypass abstract hints in search of a concrete solution. This behavior has traditionally been labeled as a form of gaming or help abuse. We propose that some examples of this behavior are not abusive and that bottom-out hints can act as worked examples. We create a model for distinguishing good student use of bottom-out hints from bad student use of bottom-out hints by means of logged response times. We show that this model not only predicts learning, but captures behaviors related to self-explanation.

1 Introduction

There are many reasons to measure a student's affective and metacognitive state. These may include adapting instruction to individual needs or designing interventions to change affective states. One technique is to build classification models with tutor interaction data, often using student response times as an independent variable[4, 6, 7, 13]. Several lines of research using this approach have targeted tutor help abuse as a behavior negatively correlated with learning[4, 13], but these classification rules for help abuse can be quite broad and thus proscribe behaviors that may be good for learning. For example, students who drill down to the most detailed hint may not be "gaming" the system, but instead searching for a worked example. The goal of this work is to show that a simple response time-based indicator can discern good bottom-out hint behaviors.

Such an indicator has several primary uses. It may be indicative of general traits, such as good metacognitive self-regulation. It might also be useful as a proxy for student affective states. The indicator, however, might be most useful in improving the design of computer-based educational scaffolding. There is significant evidence to suggest that optimal learning comes from a mix of worked examples and scaffolded problems[11]. A response time indicator for self-explanation behavior can help determine the conditions under which worked examples or scaffolded problems are better.

In this paper, we will discuss the background literature, with a focus on the potential discrepancy between existing tutor models and research in the worked example and self-explanation literature. We will then describe the data used, with an emphasis on the timing information available in the logs. We then focus on a simple student model that leads to an equally simple indicator for good bottom-out hint behaviors. We will demonstrate that the indicator has high correlation with learning gain in the forementioned data, and show

This work was supported in part by a Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education (#R305B040063)

that that study's experimental condition provides strong evidence that the indicator does capture some form of reasoning or self-explanation behaviors.

2 Background

Generally, a response time is the time required for a subject to respond to a stimulus. For educational systems, this is often detailed at the transactional level. How long does it take a student to perform a task, such as request help or enter an answer? The limitations of log data means that many details are obscure, such as the student's thought process or the presence of distractions. The student's actions may have been triggered by an event outside the systems' purview or the student may be engaging in other off-task behavior. While a limitation, it is exactly this off-task aspect of response times that make them so valuable for measuring students' affective states[7].

There is prior work on response time-based models for affective and metacognitive states. Beck et. al. modeled student disengagement using a response time-based model[7]. Their work is particulary relevant because, like this study, they use a combination of model building and elimination of irrelevant examples to construct their detector. There are also a number of help-seeking models that utilize response times in their design to better detect "gaming" behavior, particularly behaviors that involve drilling through scaffolding or repeatedly guessing[4, 13]. Still, we have found no examples of models that can detect help abuse aimed at soliciting a worked example. Indeed, one rule shared by many of these models is that bypassing traditional scaffold structure to retrieve the bottom-out hint (which is often the final answer) is considered an undesirable behavior[4]. There are models that attempt to distinguish between positive forms of help abuse and forms of help abuse that negatively impact learning[6]. They do not, however, operate under the same assumptions as this study: that good help abuse potentially involves worked examples and self-explanation.

In the literature on worked examples, particularly in the literature for self-explanation, hint abuse is not always bad for learning. If a student truly needs a worked example, drilling down through the help system may be the easiest available means of getting one. Not only is there plenty of research on worked examples to suggest that this is true[11, 12], there have been attempts to use worked examples in computerized educational systems[5, 10]. Nevertheless, for this study, the focus is on considering worked examples as improving learning through the self-explanation mechanism. Thus, the self-explanation literature provides our hypothesis as to how bottom-out hints might improve learning. In the original self-explanations studies, students were given worked examples and then asked to explain the examples to themselves[8]. Later interventions resulted in improved pre-post gain[9], but extending the self-explanation effect to computerized educational systems has not been straightforward[10]. Using self-explanation interventions in other situations, such as with students explaining their own work, has found some success in computerized systems[3]. This suggests that a more complicated mechanism relates self-explanation behavior to learning in computer tutors. Indeed, a specific connection between bottom-out hints and self-explanation has actually been suggested before[1].

3 Data

Our data is a log of student interactions with the Geometry Cognitive Tutor[2]. In the tutor, students are presented with a geometry problem and several empty text fields. A step in the problem requires filling in a text field. The fields are arranged systematically on each problem page and might, for example, ask for the values of angles in a polygon or for the intermediate values required to calculate the circumference of a circle. In the tutor, a transaction is defined as: interacting with the glossary, requesting a hint, entering an answer, or pressing "done". The done transaction is only required to start a new problem, not to start a new step.

The data itself comes from Aleven et. al.[2] They studied the addition of required explanation steps to the Geometry Cognitive Tutor. In the experimental condition, after entering a correct answer, students were asked to justify their answer by citing a theorem. This could be done either by searching the glossary or by directly inputting the theorem into a text field. The tutor labeled the theorem as correct if it was a perfect, letter-by-letter match for the tutor's stored answer. This type of reasoning or justification transaction is this study's version of self-explanation (albeit with feedback).

The hints for the tutor are arranged in levels, with each later level providing a more specific suggestion on how to proceed on that step. Whereas the early hint levels can be quite abstract, the bottom-out hint does everything shy of entering the answer. The only required work to finish a step after receiving a bottom-out hint is to input the answer itself into the text field. This type of transaction is the study's version of a simple worked example.

There were 39 students in the study that had both pre- and post-test scores. They were split 20-19 between the two conditions. All times were measured in 100th's of a second. The other details of the data will be introduced in the discussion of results.

4 Model

The core assumption underlying this work is that bottom-out hints can serve as worked examples. In our data, a bottom-out hint is defined to be the last hint available for a given step. The number of hints available differs from step to step, but the frequency of bottom-out hint requests suggests that students were comfortable drilling down to the last hint.

Assuming that bottom-out hints might sometimes serve as worked examples, there are a couple ways to detect desirable bottom-out hint behaviors. The first method is to detect when a student's goal is to retrieve a worked example. This method is sensitive to properly modeling student intentions. The second method, which is the focus of this work, is to detect when a student is *learning* from a bottom-out hint. We assume that learning derives from a mechanism similar to self-explanation and that the time spent thinking about a hint is a sufficient measure for that learning. Thus, this approach is applicable regardless of whether the student's intention was to find a worked example or to game the system.

To detect learning from bottom-out hints, our model needs estimates for the time students

Table 1: TER Model

N	Transaction	Cognition	Observed T
1	Request Hint	Think (1)	2.5
2	Try Step	Reflect (1) Think (2) Enter Answer (2)	4.3
3	Try Step	Reflect (2) Think (3) Enter Answer (3)	9.3

Table 2: Hypothetical Student Transactions
(*Observed Data*, Unobserved Data)

Step	Transaction	Observed T	Cognition	Actual T
Step 1	*Request Hint*	*0.347*	...	...
Step 1	*Enter Answer*	*15.152*	Thinking About Hint Looking at Clock Thinking About Date Typing Answer	3 2 9 1.152
Step 2	*Enter Answer*	*4.944*	Thinking About Last Step Thinking About Next Step Typing Answer	2 1 1.944

spend thinking about those hints. Call the hint time HINT_t, where HINT_t is the time spent thinking about a hint requested on transaction t. Estimating HINT_t is nontrivial. Students may spend some of their time engaged in activities unobserved in the log, such as chatting with a neighbor. However, even assuming no external activities or any off-task behavior whatsoever, HINT_t is still not directly observable in the data.

To illustrate, consider the following model for student cognition. On an answer transaction, the student first thinks about an answer, then they enter an answer, and then they reflect on their solution. Call this model TER for Think-Enter-Reflect. An illustration of how the TER model would underly observed data is shown in Table 1. Notice that the reflection time for the second transaction is part of the logged time for the third transaction. Under the TER model, the reflection time for one transaction is indistinguishable from the thinking and entry time associated with the next transaction. Nevertheless, we need an estimate for the Think and Reflect times to understand student learning from bottom-out hints.

The full problem, including external factors, is illustrated in Table 2, which shows a series of student transactions on a pair of problem steps, along with hypothetical, unobserved student cognition. Entries in italics are observed in the log while those in normal face are unobserved, and ellipses represent data not relevant to the example. The time the student spends thinking and reflecting on the bottom-out hint is about 6 seconds, but the only *observed* durations are 0.347, 15.152, and 4.944. In a case like this, the log data's observed response times includes a mixture of Think and Reflect times across multiple steps.

Table 3: TER Model With Estimators

Transaction	Cognition	Notation
Hint	...	...
Enter Answer	Think About Step	K_t
	Off-Task	
	Enter Answer	E_t
Enter Answer	Reflect on Previous	R_t
	Think About New Step	K_{t+1}
	Off-Task	
	Enter Answer	E_{t+1}

Unfortunately, while the reflection time is important for properly estimating HINT$_t$, it is categorized incorrectly. The reflection time for transaction t is actually part of the logged time for transaction $(t+1)$. Teasing those times apart requires estimating the student's time not spent on the hint.

The first piece of our model separates out two types of bottom-out hint cognition: Think and Reflect. Thinking is defined as all hint cognition before entering the answer; reflecting is all hint cognition after entering the answer. Let Think time be denoted K_t and Reflect time be denoted R_t. We define HINT$_t = K_t + R_t$.

The task then reduces to estimating K_t and R_t. As shown earlier, this can be difficult for an arbitrary transaction. However, we focus only on bottom-out hints. Table 3 provides an example of how bottom-out hints differ from other transactions. Note the absence of a Reflect time R_{t-1} in the bottom-out case. Except for time spent on answer entry and time spent off-task, the full time between receiving the hint and entering the answer is K_t. A similar, but slightly more complicated result applies to R_t. For now, assume off-task time is zero - it will be properly addressed later. Let the answer entry time be denoted E_t. Let the total time for a transaction be T_t. Then the equation for HINT$_t$ becomes

$$\begin{aligned} \text{HINT}_t &= K_t + R_t && (1)\\ &= (T_t - E_t) + R_t && (2)\\ &= (T_t - E_t) + (T_{t+1} - (K_{t+1} + E_{t+1})) && (3) \end{aligned}$$

where T_t and T_{t+1} are observed in the log data. The first term consists of replacing K_t with measured and unmeasured times from before the answer is submitted. The second term consists of times from after the answer is submitted. If we have an estimate for E_t, we can now estimate K_t. Similarly, if we have an estimate for K_{t+1} and E_{t+1}, we can estimate R_t.

Constructing reliable estimates for any of the above values is impossible on a per transaction basis. However, if we aggregate across all transactions performed by a given student, then the estimators become more reasonable. There are two other important points regarding the estimators we will use. First, response times, because of their open-ended nature,

are extremely prone to outliers. For example, the longest recorded transaction is over 25 minutes in length. Thus, we will require our estimators be robust. Second, some students have relatively few (≈ 10) bottom-out hint transactions that will fit our eventual criteria. Thus, our estimators must converge quickly.

Now we need some new notation. We will be using the s subscript, where s represents a student. We will also use the $\hat{E}_s$ notation for estimators and the $m(E_t)$ notation for medians. You can think of $m(E_t)$ as approximating the mean, but we will always be using the median because of outliers. E_s, the per student estimator, will represent some measure of the "usual" E_t for a student s. Also let A be the set of all transactions and A_s be the set of all transactions for a given student. Let A_s^1 be the set of all correct answer transactions by a student s where the transaction immediately follows a bottom-out hint. Similarly, let A_s^2 be the set of all transactions that follow a transaction $t \in A_s^1$. For convenience, we will let $T_s^1 = m_{t \in A_s^1}(T_t)$ be the median time for transactions $t \in A_s^1$ and $T_s^2 = m_{t \in A_s^1}(T_{t+1})$ be the median time for transactions $t \in A_s^2$. These two types of transactions are generalizations of the last two transactions shown in Table 3. This gives an equation for our estimator $\hat{\text{HINT}}_s$,

$$\hat{\text{HINT}}_s = (T_s^1 - E_s) + (T_s^2 - (K_s^2 + E_s)) \tag{4}$$

Here, $K_s^2 = m_{t \in A_s^2}(K_t)$ is the thinking time that takes place for transaction $t \in A_s^2$.

Consider $\hat{E}_s$, the median time for student s to enter an answer. It always takes time to type an answer, but the time required is consistently short. If we assume that the variance is small, then $\hat{E}_s \approx \min_{t \in A_s}(E_t)$. That is, because the variance is small, E_t can be treated as a constant. We use the minimum rather than a more common measure, like the mean, because we cannot directly observe E_t. Instead, note that if $K_t \approx 0$, then the total time spent on a post-hint transaction is approximately E_t. Thus, the minimum time student s spends on an answer step is a good approximation of $\min_{t \in A_s}(E_t)$. In practice, the observed $\hat{E}_s$ is about 1 second. With $\hat{E}_s$, we can now estimate K_t for $t \in A_s^1$.

To isolate the reflection time R_s, we need an approximation for K_s^2, the thinking time for transactions $t \in A_s^2$. Unfortunately, K_s^2 is difficult to estimate. Instead, we will estimate a value related to K_s^2. The key observation is, if a student has already thought through an answer on their own, without using any tutor help, they presumably engage in very little reflection after they enter their solution. To put it mathematically, let N_s be the set of transactions for student s where they do *not* use a bottom-out hint. We assume that $R_t \approx 0$, $\forall t \in N_s$. We can now use the following estimator to isolate R_s,

$$\begin{aligned} R_s &= T_s^2 - (K_s^2 + E_s) && (5)\\ &= m(T_t)_{t \in A_s^2} - (K_s^2 + E_s) && (6)\\ &\approx m(T_t)_{t \in A_s^2} - m(T_t)_{t \in N_s} && (7) \end{aligned}$$

where the change from line 6 to line 7 derives from the assumption $R_t \approx 0$, $\forall t \in N_s$. This is the last estimator we require: R_s is approximately $m(T_t - m(T_v)_{(u \in N_s, v = u+1)})_{t \in A_s}$.

Table 4: Indicator Correlations in the Control Condition
(* $p < 0.10$, ** $p < 0.05$)

	Pre	Post	Adjusted Gain
K_s	-0.11	0.37*	0.34*
R_s	0.29	0.42**	0.36*
HINT$_s$	0.04	0.53**	0.48**

That is, we use the median time for the first transaction on a step where the prior step was completed without worked examples. This approach avoids directly estimating K_s^2 and estimates the sum $(K_s^2 + E_s)$ instead.

There is still the problem of off-task time. We have so far assumed that off-task time is approximately zero. We will continue to make that assumption. While students engage in long periods of off-task behavior, we assume that for most transactions, students are on-task. That implies that transactions with off-task behaviors are rare, albeit potentially of long duration. Since we use medians, we eliminate these outliers from consideration entirely, and thus continue to assume that on any given transaction, off-task time is zero.

A subtle point is that the model will not fit well for end-of-problem transactions. At the end of a problem there is a "done" step, where the student has to decide to hit "done". Thus, the model no longer accurately represents the student's cognitive process. These transactions could be valuable to an extended version of the model, but for this study, all end-of-problem transactions will be dropped.

5 Results

We first run the model for students in the control condition. These students were not required to do any formal reasoning steps. The goal is to predict the adjusted pre-post gain, $\max(\frac{\text{(post}-\text{pre)}}{\text{(1}-\text{pre)}}, \frac{\text{(post}-\text{pre)}}{\text{(pre)}})$. We will not use the usual Z-scores because the pre-test suffered from a floor effect and thus the pre-test scores are quite non-normal (Shapiro-Wilks: $p < 0.005$). Two students were removed from the population for having fewer than 5 bottom-out hint requests, bringing the population down to 18. The results are shown in Table 4.

The first result of interest is that none of the indicators have statistically significant correlations with the pre-test. This suggests that they measure some state or trait of the students that is not well captured by the pre-test. The second result of interest is that all three indicators correlate strongly with both the post-test and learning gain. Notably, HINT$_s$, our main indicator, has a correlation of about 0.5 with both the post-test and the learning gain. To the extent that HINT$_s$ does distinguish between "good" versus "bad" bottom-out hint behaviors, this correlation suggests that the two types of behavior should indeed be distinguished.

It's possible that these indicators might only be achieving correlations comparable to time-on-task or average transaction time. As Table 5 shows, this is clearly not the case. All three

Table 5: Time-on-Task Correlations in the Control Condition

	Pre	Post	Adjusted Gain
Time-on-Task	-0.31	-0.10	0.23
Average Transaction Time	-0.03	0.27	0.20

Table 6: Correlations in the Experimental Condition
(* $p < 0.10$, ** $p < 0.05$)

	Pre	Post	Adjusted Gain
K_s	-0.02	0.23	0.25
R_s	0.02	0.31*	0.35*
HINT$_s$	0.00	0.38*	0.41**

hint time indicators out-perform the traditional time-on-task measures.

Nevertheless, these results still do not show whether the indicator HINT$_s$ is actually measuring what it purports to measure: self-explanation on worked examples. For that, we use the experimental condition of the data. In the experimental condition, students are asked to justify their correct solutions by providing the associated theorem. This changes the basic pattern of transactions we are interested in from HINT-GUESS-GUESS to HINT-GUESS-JUSTIFY-GUESS. We can now directly measure R_s using the time spent on the new JUSTIFY steps. R_s is now the median time students spend on a correct justification step after a bottom-out hint, subtracting the minimum time they ever spend on correct justifications. We use the minimum for reasons analogous to those of $\hat{E}_s$ - we only want to subtract time spent entering the reason. In this condition, there were sufficient observations for all 19 students. The resulting correlations are shown in Table 6.

There is almost no correlation between our indicators and the pre-test score, again showing that our indicators are detecting something not effectively measured by the pre-test. Also, the correlations with the post-test and learning gain are high for both R_s and HINT$_s$. While R_s by itself has a statistically significant correlation at $p < 0.10$, K_s and R_s combined demonstrate a statistically significant correlation at $p < 0.05$. This suggests that while some students think about a bottom-out hint before entering the answer and some students think about the hint only after entering the answer, for all students, regardless of style, spending time thinking about bottom-out hints is beneficial to learning. The corollary is that at least some bottom-out hints are proving beneficial to learning.

Thus far, we have shown that the indicator HINT$_s$ is robust enough for strong correlations with learning gain despite being measured in two different ways across two separate conditions. The first set of results demonstrated that HINT$_s$ can be measured without any direct observation of reasoning steps. The second set of results showed that direct observation of HINT$_s$ was similarly effective. Our data, however, allows us access to two other interesting questions. First, does prompting students to explain their reasoning change their bottom-out hint behavior? Second, do changes in this behavior correlate with learning gain?

Table 7: Changes in Behavior in the Experimental Condition
(* $p < 0.10$, ** $p < 0.05$)

	Mean	**Var**	**Pre**	**Post**	**Adjusted Gain**
ΔHINT$_s$	0.56	5.06	0.41*	0.47**	0.38*

To answer both questions, we look at the indicators trained on only the first 20% of each student's transactions. For this, we use only the experimental condition because, when 80% of the data is removed, the control condition has too few remaining students and too few observations. Even in the experimental condition, only 15 remaining students still have more than 5 bottom-out hint requests that meet our criteria. The results are shown in Table 7, with ΔHINT$_s$ representing the difference between HINT$_s$ trained on the first 20% of the data and HINT$_s$ trained on the full data.

To answer the first question, the change in HINT$_s$ is not statistically different from zero. The prompting does not seem to encourage longer response times in the presence of bottom-out hints, so this mechanism does not explain the experimental results of Aleven et. al.'s study[2]. However, some of the students did change their behaviors. As shown in Table 7, students who increased their HINT$_s$ times demonstrated higher learning gain. The evidence is substantial that HINT$_s$ measures an important aspect of student reasoning.

6 Conclusions and Future Work

In this study, we presented evidence that some bottom-out hint use can be good for learning. The correlations between our indicators and pre-post learning gain represent one form of evidence; the correlations between *changes* in our indicators and pre-post learning gain represent another. Both sets of results show that thinking about bottom-out hints predicts learning. However, extending our results to practical use requires additional work.

Our indicators provide estimates for student thinking about bottom-out hints. However, these estimates are aggregated across transactions, providing a student level indicator. While this is useful for categorizing students and offering them individualized help, it does not provide the level of granularity required to choose specific moments for tutor intervention. To achieve that level of granularity, a better distributional understanding of student response times would be helpful, as would an indicator capable of distinguishing between students seeking worked examples versus engaging in gaming. Exploring how the distribution of response times differs between high learning bottom-out hint students and low learning bottom-out hint students would go a long way to solving both problems.

That issue aside, our indicators for student self-explanation time have proven remarkably effective. They not only predict learning gain, they do so better than traditional time-on-task measures, they are uncorrelated with pre-test scores, and changes in our indicators over time also predict learning gain. These indicators achieve this without restrictive assumptions about domain or system design, allowing them to be adapted to other educational systems in other domains. Whether the results transfer outside of geometry or to other

systems remains to be seen, but they have so far been robust.

The inclusion of two conditions, one without justification steps and one with justification steps, allowed us to show that the indicators do measure something related to reasoning or self-explanation. We estimated the indicators for the two conditions in different ways, yet both times, the results were significant. This provides a substantial degree of validity. However, one direction for future work is to show that the indicators correlate with other measures of self-explanation or worked example cognition. One useful study would be to compare these indicators with human estimates of student self-explanation.

References

[1] Aleven, V., Koedinger, K. R. Limitations of student control: Do students know when they need help? *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, 2000, p. 292–304.

[2] Aleven, V., Koedinger, K. R. An effective meta-cognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, 26(2), 2002, p. 147–179.

[3] Aleven, V., Koedinger, K. R., Cross, K. Tutoring answer explanation fosters learning with understanding. *Proceedings of the 9th International Conference on Artificial Intelligence in Education*, 1999.

[4] Aleven, V., McLaren, B. M., Roll, I., Koedinger, K. R. Toward tutoring help seeking - applying cognitive modeling to meta-cognitive skills. *Proceedings of the 7th Conference on Intelligent Tutoring Systems*, 2004, p. 227–39.

[5] Atkinson, R. K., Renkl, A., Merrill, M. M. Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology*, 95(4), 2003, p. 774–783.

[6] Baker, R. S., Corbett, A. T., Koedinger, K. R. Detecting student misuse of intelligent tutoring systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 2004, p. 531–540.

[7] Beck, J. Engagement tracing: using response times to model student disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 2005, p. 88–95.

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

[9] Chi, M. T. H., de Leeuw, N., Chiu, M. H., LaVancher, C. Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 1994, p. 439–477.

[10] Conati, C., VanLehn, K. Teaching meta-cognitive skills: Implementation and evaluation of a tutoring system to guide self-explanation while learning from examples. *Proceedings of the 9th International Conference on Artificial Intelligence in Education*, 1999, p. 297–304.

[11] Paas, F. G. W. C., Van Meeriënboer, J. J. G. Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, 86(1), 1994, p. 122–133.

[12] Sweller, J., Cooper, G. A. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2, 1985, p. 251–296.

[13] Walonoski, J. A., Heffernan, N. T. Detection and analysis of off-task gaming behavior in intelligent tutoring systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 2006, p. 722–724.