Effect of student ability and question difficulty on duration

Yijun Ma
Teachers College,
Columbia University
New York, NY 10027
ym2476@tc.columbia.edu

Lalitha Agnihotri
McGraw Hill Education
2 Penn Plaza
New York, NY 10121
lalitha.agnihotri@mheducation.com

Ryan Baker
Teachers College,
Columbia University
New York, NY 10027
ryanshaunbaker@gmail.com

Shirin Mojarad
McGraw Hill Education
281 Summer Street
Boston, MA 02210
shirin.mojarad@mheducation.com

ABSTRACT

Time has become a standard feature used in EDM models, and is used in models of meta-cognitive strategies to models of disengagement. Most of these models consider whether a student action is "too fast" or "too slow". However, an open question remains on how we define and select these cut-offs. Moreover, it is not clear that the same cut-offs are appropriate across different situations. Some students may generally respond faster than others; more difficult items may take different amounts of time. In this paper, we consider whether absolute or relative indicators of time are more appropriate as cut-offs, and whether simple transformations (such as log time) are useful when representing time. We do so through visualizing student performance in relation to general student ability, item difficulty, and different ways of representing time. We find that student knowledge and item difficulty should be taken into account when choosing cut-offs, and that there are advantages to representing duration in terms of standardized log-time.

Keywords

Time taken, Duration, Visualization, Student Ability, Rasch Model, 1PL, Item Difficulty

1. INTRODUCTION

Over the decade since the Educational Data Mining community began to coalesce, one of the most common ways to interpret student behavior has been to look at the amount of time taken to respond to questions. Early work by Aleven, Baker, and Beck tried to determine whether a response was "too fast", indicating gaming the system, help abuse, try-step abuse, or disengaged behavior [1, 2, 3]. Soon, work began to consider whether a response was "too slow" as well [4]. Researchers noted that performance seemed to degrade when behavior reached either of these two extremes. This theme of trying to identify behavior as "too fast" or "too slow" continues to this day [5, 6]. Actions that are "too fast" or "too slow" are seen as components in a range of EDM models, including contemporary models of gaming the system [7], off-task behavior [8, 9], carelessness [10, 11], and self-explanation [12].

However, one of the interesting aspects of this body of literature is how remarkably inconsistent it is, as noted by [5]. Despite their conceptual simplicity, researchers do not agree what "too fast" or "too slow" means. This inconsistency may not be a major concern when these parameters are empirically fit using training labels, but is somewhat more concerning when cut-offs are rationally defined.

Part of the reason for inconsistency, of course, is that "too fast" and "too slow" are inherently contextual. Interfaces matter. A student completing division problems by typing in answers is likely to respond faster than a student chasing down a skeleton and hitting the right divisor key [13]. Ability matters. A 7-year old solving arithmetic problems is likely to perform more slowly than a 38-year old. Difficulty matters. Even for the same user interface and an experienced adult, "49 / 7" will be solved more quickly than "602 / 7".

For this reason, it is unlikely there is a universal answer to how fast is too fast, and how slow is too slow. Nor will it be easy to find a simple formula or set of formulas that can predict this. Mathematical models based on memory [14] can make predictions about speed in some situations, but are incomplete for many of the complex types of problem-solving and the activities surrounding problem-solving in modern learning environments. At the same time, there exist simple psychometric models that can predict a considerable amount of variance in performance, which may be useful in investigations of this nature.

One solution, as discussed above, is to empirically select a single cut-off, but part of the challenge is that even within a learning environment, cut-offs both vary contextually, and exist on a continuum. In this paper, we will examine this continuum in a visual fashion, across different situations within a single online learning environment. Specifically, we will analyze how the relationship between time and performance varies when students vary in knowledge, and for items of different overall difficulty.

We will also investigate whether the most commonly used way to represent time (number of seconds) is the best representation for understanding these issues, or whether standardizing or transforming time makes it easier to understand the relationship between time and performance.

By better understanding these relationships, we will be able to select more appropriate cut-offs, and develop more precise models for discovery with analysis and interventions.

2. DATA SET

We investigate these issues in the context of one of the world's most widely used digital learning environments, McGraw-Hill Education's Connect system [16, 17]. Connect is currently actively used by approximately two million students and 25,000 instructors. Within Connect, instructors select questions from question banks and the system then administers them to the student as homework, quiz, exam, or practice assignments. Most items are auto-graded by the system, and immediate feedback is provided when relevant (e.g. not during exams). Within homework and practice assignments, students can make multiple attempts to answer each question, based on the policies set up by the instructor. In this paper, we use item and questions interchangeably.

Connect is organized into courses; each course is tied to a McGraw-Hill book title, and question banks are organized in relation to book chapters. In this paper, we focus on a single textbook in order to avoid including radically different material together in the same analysis (for example, one might expect calculus problems to take longer to solve than questions about the factual aspects of history). We analyze a data set from 173 courses that utilize the title *McGraw-Hill's Taxation of Individuals and Business Entities, 6th Edition,* by Brian Spilker, a medium-sized data set with relatively consistent item design, involving a course text with items selected as a focus for enhancement within McGraw-Hill at the time this research was being conducted. Within this textbook, there were multiple types of items: multiple choice items where single responses were correct, multiple choice items where multiple responses were correct, fill-in-the-blank items, matching questions, and ungraded essays (removed prior to analysis).

Within this textbook, within the period between August 2014 and November 2014, 3,882 students (working with 86 instructors) answered 2,947 distinct questions. In total, this set of students attempted to answer questions 536,520 times, an average of 138.21 attempts per student.

Prior to analysis, we removed all ungraded questions from the data set (as assessing correctness is outside the scope of this paper). We also removed attempts where the student timed-out due to inactivity within the system for 60 minutes, and where the student's response time was not collected or had impossible values (due to logging errors). For this specific analysis, we removed students' second and subsequent attempts to answer questions, focusing on their performance and time taken on their first attempt. Although second and subsequent attempts are relevant to issues of modeling student behaviors such as off-task behavior and gaming the system, these times are strongly influenced by the time taken on the first attempt, and are relatively more complex to consider. As such, we leave analysis of second and subsequent attempts to future work. The resultant cleaned data set involved 3,632 students answering 2,689 distinct questions, attempting to answer items 365,302 times, an average of 100.58 attempts per student.

Within these items, scores were distributed between 0 and 1, with 76% of items receiving a fully correct score of 1. However, partial credit was assigned by instructors and, as a result, is somewhat non-uniform; different items had different partial credit assigned for different responses. As such, the partial credit information was less useful for analysis than in other systems where it is assigned in a consistent fashion [15, 18]. To avoid having our results impacted by this inconsistency, we assigned a value of 0 (incorrect) to any student response that was not fully correct. Only 7.9% of the problem attempts were affected by this modification.

2.1 Tagging with Question Difficulty and Student Ability

In order to understand how student knowledge and item difficulty influence the relationship between time taken and performance, we annotated the data with a well-known psychometric model: the Rasch Model [19, 20].

The Rasch Model is one of the most widely used models in the history of psychometrics. It relates performance to student ability (treated here as overall knowledge of the domain) and item difficulty. More recent and advanced models from the psychometrics and student modeling literature consider change in knowledge over time, group items into latent skills, explicitly model the probability of guess and slip, and use different uncertainty functions for students and items [21, 22, 23, 24, 25]. However, the Rasch model is appropriate for the analysis here, as assesses student knowledge and item difficulty (which is what we focus on in the analyses below), it is known to function well when different students answer different items [19], and has high stability and reliability [20].

The equation for the Rasch model is given as follows [19]:

$$P(\theta) = \frac{1}{1 + e^{-1(\theta - b)}}$$

where b is the question difficulty parameter, θ is the student ability (knowledge) level, and $P(\theta)$ is the probability that the student will answer the current item correctly. Within this model, if a student's ability is equal to the item's difficulty ($\theta = b$), the probability that the student will answer the question correctly is 50%. As the student's ability becomes higher or the item's difficulty becomes lower, the probability of correctness increases and finally is approximately equal to 1; correspondingly, as ability becomes lower or difficulty becomes higher, the probability of correctness approaches 0.

As is standard [19], we use Maximum Likelihood Estimation, in this case converging after seven iterations, to estimate the values of θ and b for each student and item based on actual data. After fitting and applying the model, all student attempts are tagged with a difficulty parameter and an ability parameter.

This model achieves an R-squared value of 0.322, and an A' (mathematically equivalent to AUC but easier to calculate) of 0.852, calculated using the A' calculator available at http://www.columbia.edu/~rsb2162/computeAPrime.zip .

3. Analysis

We analyze the research questions discussed above through a set of visualizations, created in Python's matplotlib library. Each of the visualizations will place some variant of the time taken by the student to give a response on the X axis, and place the percentage of times when the student response was correct (percent correct) on the Y axis. In the visualizations, item responses are binned to one-second grain-size. For that

bin, we find the percent correct and plot a dot there; if there are more items in the bin, the dot is made larger.

3.1 Baseline Graph

In the first visualization, Figure 1, we consider the baseline relationship between time taken and percent correct. Item difficulty according to the Rasch model is also included in the visualization as color, with darker colored dots representing easier items and lighter dots representing harder items (e.g. if a dot is dark, the items composing that dot were on average easier).[12]

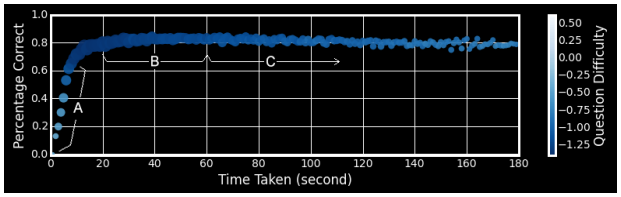

Figure 1: The relationship between the time taken to respond to an item, and correctness. Color is used to denote item difficulty.

As Figure 1 shows, students who spend very little time on an item typically achieve low percentage correct. As the time taken increases, performance improves, curving up from 0 seconds to about 12 seconds; this range of the graph is denoted "A". Percent correct remains stable from 20 seconds to 60 seconds; this range of the graph is denoted "B". As students spend over 60 seconds, their performance somewhat declines again; this range of the graph is denoted "C". This graph shows a similar qualitative pattern to the pattern seen in other systems, but with the shifts occurring at different points. For example, Beck [3] finds that performance improves up until the student has spent 4 seconds, remains stable under 7 seconds, and drops gradually after that.

It is worth noting that despite these shifts, it is non-trivial to find cut-offs. 12 seconds is approximately the inflection point where performance shifts to being stable, but it probably contains more positive behavior than would be desired. It might still be desirable to pick a lower cut-off point for "too fast". Similarly, the difference between 60 seconds and 100 seconds for "too long" is relatively minimal.

One limitation to Figure 1 is that fewer and fewer data points are seen as the times get longer, making it difficult to show all the data in a relatively limited horizontal space. This limitation can be addressed by switching from absolute time in seconds, to a logarithmic scale for time, shown in Figure 2. By switching to a logarithmic scale, the long tail of long response times is compressed to a small section of the plot and we can show more data while maintaining the essence of the graph. The log scale thus makes it easier to present our full data.

The log scale also makes it easier to see that there are more inflection points than Figure 1 showed. The same ranges (0-12 seconds, 20-60 seconds and 60+ seconds) are marked in Figure 2 as in Figure 1, to enable comparison. Note that between 0-12 seconds (range A), there is a secondary inflection point around 3.5 log time taken where performance shifts from improving slowly to improving quickly. This might be a better cut-off for "too fast" than 12 seconds. Similarly, the decline in performance can be seen to begin around 4.75 log time taken but to accelerate after 5.5 log time taken, suggesting a potentially better "too slow" cut-off. While these cut-offs are somewhat harder for a reader to interpret directly from the numbers, they allow us to make more sophisticated distinctions than were possible just from absolute time.

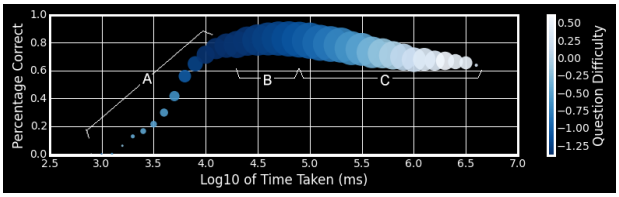

Figure 2: The relationship between the time taken (log scale) to respond to an item, and correctness. Color is used to denote item difficulty.

3.2 Standardization

One common decision seen in many models that measure student time [26, 27] is to represent student time in terms of standard deviations faster or slower than the average time, calculated as a Z-score, and referred to as standardized time or unitized time. This transformation, which assumes that time is normally distributed, uses the formula

$$Z = \frac{Time - Mean(Time)}{SD\ (Time)}$$

The logic is that this approach accounts for the fact that different items need different amounts of time to answer them, allowing fairer comparison of student time on different items.

Figure 3 shows the results of applying this transformation to our data.

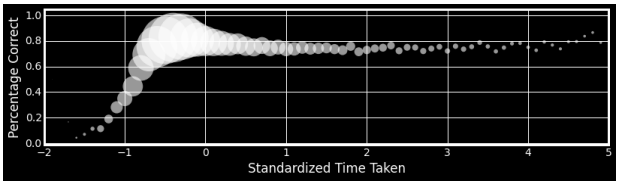

Figure 3: The relationship between the standardized time taken to respond to an item, and correctness.

As this graph shows, most of the data is now clumped together. Notably, the center of the data is not at 0 SD; instead the median is somewhere around -0.5 SD. Though 0 SD is by definition the average value, it is clearly not the median value. This is a common limitation to using standardization, and one that the authors have observed in previous data sets as well. As such, using standardization is vulnerable to skewness and outliers in the original data, making it broadly unsuitable for use across data sets – or indeed, for cases where the magnitude of the long time outliers may vary over time. This can occur, for example, when the original data set has a small number of students with extremely high outlier times, or when the system time-out may change over time. This suggests that standardized time is undesirable for use in cut-offs, since the cut-off points may vary depending on the exact outliers in the data set. This could be addressed by ignoring the outliers when computing

the SD value (i.e. truncating the values of extreme outliers [28];) but doing so will only incompletely address a second problem; the data is highly compressed relative to the previous visualizations we have examined. Most of the data points occur in a fairly small range. In this case, 64.4% of the data is clumped between Z= -1 and Z = 0. If the data were distributed according to assumptions, 68% of data would be clumped between Z= -1 and Z= 1, double the range. This clumping makes it difficult to see the inflections in performance for rapid student responses; although the graph's clumping does allow us to see that there is some rise in performance for very high response times (a set of outliers outside of bounds for the earlier representations).

One alternative, shown in Figure 4, is to use [29] modified Z-score, which is computed as:

$$M_i = \frac{0.6745\ (Time - Median(Time))}{MAD\ (Time)}$$

where MAD stands for Median Absolute Deviation.

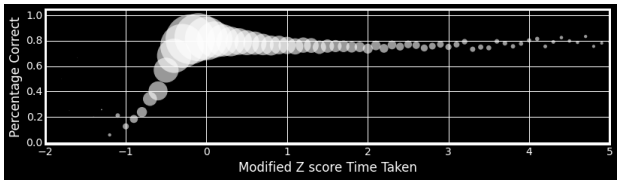

Figure 4: The relationship between the modified Z-score standardized time taken to respond to an item, and correctness.

This approach centers the data better, but does not solve the problem of the data being compressed.

Another alternative is to conduct standardization on time transformed to a logarithmic scale, shown in Figure 5. As we saw in the previous section, using a logarithmic scale spread out the data better and allowed us to see inflection points more clearly.

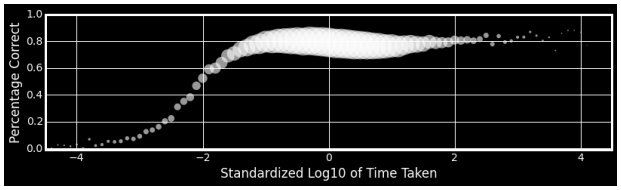

Figure 5: The relationship between the standardized log-transformed time taken to respond to an item, and correctness.

As Figure 5 shows, standardizing using a logarithmic scale centers the data as well as using modified Z-score, but spreads the data out better. The data is broadly centered on Z = 0, with most of the data (68.82%) between Z = -1 and Z = 1 (almost exactly the amount that one would expect for normally distributed data). The same inflection points visible at the left side of Figure 2 are visible at the left side of Figure 5. At the same time, while the logarithmic nature of the transformation does compress the right tail somewhat, we nonetheless can see the same rise in performance at very high time taken that we saw in Figure 3. As such, this representation helps us in understanding the data and choosing cut-offs, while gaining the benefit of comparability that standardizing variables gives us.

3.3 Studying Item Difficulty

One factor that is worth considering is that the time taken appears to be associated with how difficult the items are. Figures 1 and 2 each show difficulty in terms of color, with blue representing easier items (according to the Rasch model discussed above) and white representing harder items.

In Figure 1, we can see that the hardest items are found at the two ends of the spectrum; the briefest times taken, and the longest times taken. It is unsurprising that students take longer on hard items. The connection between difficulty and brief responses is also reasonable; students are more likely to become disengaged and engage in behaviors such as gaming the system and carelessness when encountering hard items [30]. The same pattern is seen in Figure 2, although whether the lowest difficulty is seen for higher or lower times varies between graphs. This is simply a result of the fact that Figure 2 shows more of the data set than Figure 1, due to the use of a logarithmic scale.

This leads to the question of how we should expect the relationship between the student's time taken and their performance to change based on item difficulty. In particular, does the same amount of time taken mean different things for easy items versus difficult items? It is plausible to hypothesize – for example – that rapid responses on easy items may imply fluent knowledge [31] but rapid responses on difficult items may imply disengagement [3].

We examine this by grouping items, based on their difficulty according to the Rasch model *b* parameters, into 5 bands, shown in Table 1, and displayed in Figures 6 and 7.

Table 1: The difficulty groups shown in Figures 6 and 7, based on *b* in the Rasch model. Items with *b* below -3 look very similar to items with *b* from -1 to -3, so they are included in the same group.

Difficulty < -1	Dark Blue
Difficulty 0 to -1	Light Blue
Difficulty 0 to 1	Light Yellow
Difficulty 1 to 3	Yellow
Difficulty > 3	Red

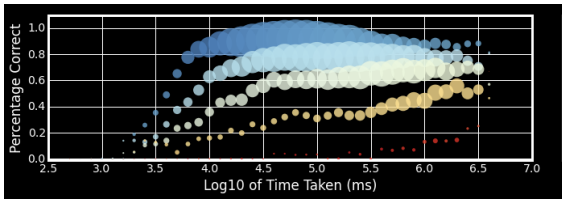

Figure 6: The relationship between the log-transformed time taken to respond to an item, and correctness, for each of the difficulty bands shown in Table 1.

As Figure 6 shows, the pattern for dark blue and light blue (the lower-difficulty items) is largely the same as in Figure 2. Correctness increases fairly rapidly when students spend more time, leveling off and then slowly declining for high amounts of time spent. However, the amount of time needed for high

levels of correctness is higher for the light blue items (*b* between 0 and -1) than for the dark blue items (*b* below -1). This suggests that the same cut-off for "too fast" is not appropriate for items with different difficulty.

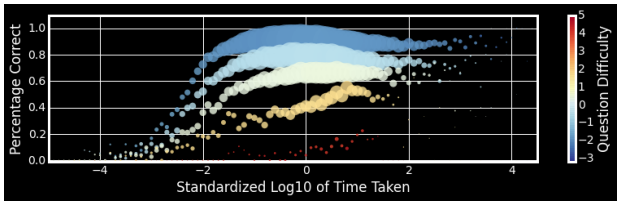

Figure 7: The relationship between the standardized log-transformed time taken to respond to an item, and correctness, for each of the difficulty bands shown in Table 1.

As Figure 7 indicates, this difference between the time needed for the lowest-difficulty items (dark blue) and the moderately low-difficulty items (light blue) cannot be controlled for, simply by switching to standardized log time. Even after we switch to standardized log time, more time is needed for the moderately low-difficulty items than for the lowest-difficulty items, to reach high levels of correctness.

The decline in performance for students who spend too much time (possibly going off-task, or asking for help) is seen for both of these two item difficulty groups, in both the log-time graph and the standardized log-time graph.

Interestingly, the patterns seen are different for the higher-difficulty items. Focusing on yellow and red, we can see that there is no clear inflection point where spending more time is associated with worse performance, or even a clear leveling off in performance. For yellow (*b* between 1 and 3), there is a range between -1 and -1.5 standardized log time where performance may be leveling off or mildly dropping, but it is at best a minor and brief shift, compared to the lower-difficulty bands. For yellow, "too fast" cut-offs could be placed within the -1 to -1.5 SD range, somewhat higher than for lower difficulty (it is hard to identify any good place for a cut-off in the non-standardized graph). For red (*b* above 3), there is essentially no range where increasing time does not improve performance. For neither of these bands is there a clear "too slow" range, where performance worsens once too high a time spent is reached.

These graphs show that time cut-offs should not be considered independently of item difficulty. We are not aware of any models of gaming the system, carelessness, off-task behavior, or related constructs that explicitly consider item difficulty. Our results suggest that this omission is lowering the quality of these models.

3.4 Studying Student Knowledge

Finally, we consider how the student's knowledge of the domain impacts their time spent. Figures 8 and 9 each show knowledge in terms of color, with green representing more knowledgeable students (according to the Rasch model discussed above) and white representing less knowledgeable students. Note that this color scheme corresponds to the color scheme used for difficulty – students are less likely to produce correct answers for white dots.

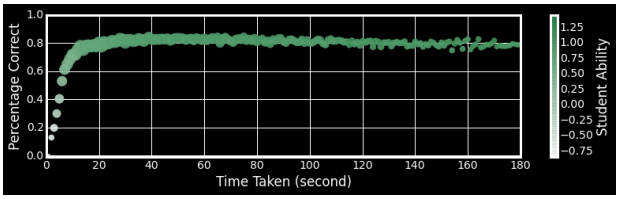

Figure 8: The relationship between the time taken to respond to an item, and correctness. Color is used to denote student overall domain knowledge, assessed using the ability parameter in the Rasch model.

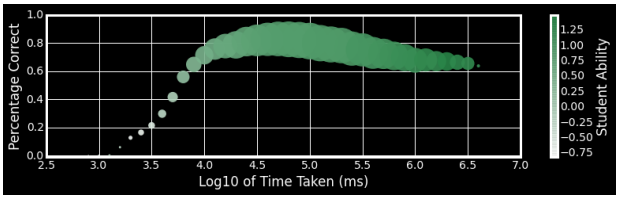

Figure 9: The relationship between the log transformed time taken to respond to an item, and correctness. Color is used to denote student overall domain knowledge, assessed using the ability parameter in the Rasch model.

Figures 8 and 9 show a different pattern than Figures 1 and 2. Whereas those earlier figures indicated that short and long times were seen for hard items, Figures 8 and 9 indicate that brief times are seen for the least able students while long times are generally seen for knowledgeable students. This result suggests that less knowledgeable students appear to be more likely to engage in behaviors such as gaming the system and carelessness, but there does not seem to be a similar pattern for off-task behavior.

Figure 10 shows the same item difficulty bands as were seen in Figure 7, but colored in terms of student ability rather than item difficulty. We can see that regardless of question difficulty, if the response time is too fast relative to the average for the item, the student is likely to be of low ability. However, we can also see from box T1 that this low ability is also seen for longer response times for harder items. For the easiest items, lower ability is seen below -2 SD for time; for the hardest items, lower ability is seen below -1.2 SD for time. As such, this figure indicates that the behavior of answering too fast is seen across questions with different difficulties, though the cut-off should differ.

For higher difficulty items, longer time taken is associated with better students, as shown in T2. But this effect only manifests for the higher difficulty items; these items are more discriminative in terms of the relationship between student ability and longer time taken. Finally, most of the examples of responses that are relatively much longer than other responses occur on the easier items – it is harder to distinguish responses that are genuinely too long for harder items.

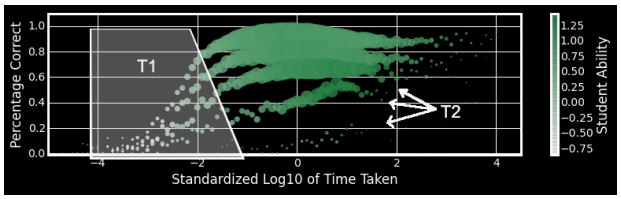

Figure 10: The relationship between the log-transformed time taken to respond to an item, and correctness, for each of the difficulty bands shown in Table 1, but colorized in terms of student ability.

Given these results, we can reasonably ask: how should we expect the relationship between the student's time taken and their performance to change based on the student's general knowledge of the item? In particular, does the same amount of time taken mean different things for knowledgeable students versus not knowledgeable students? Correspondingly, with the above, it is plausible to hypothesize – for example – that rapid responses by knowledgeable students may imply fluent knowledge but rapid responses by struggling students may imply disengagement [14].

We examine this by grouping students, based on their knowledge level according to the Rasch model θ parameters, into 5 bands, shown in Table 2, and displayed in Figure 11.

Table 2: The difficulty groups shown in Figure 11, based on b in the Rasch model. Items with θ below -3 look very similar to items with θ from -1 to -3, so they are included in the same group.

Knowledge < -1	Dark Red
Knowledge 0 to -1	Brick Red
Knowledge 0 to 1	Pink
Knowledge 1 to 3	Light Green
Knowledge > 3	Green

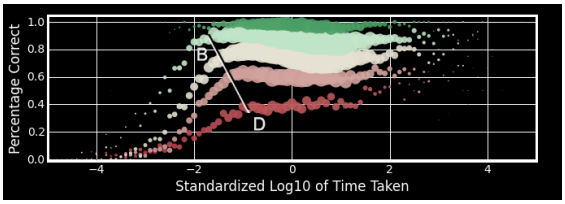

Figure 11: The relationship between the standardized log-transformed time taken to respond to an item, and correctness, for each of the student knowledge bands shown in Table 2.

As Figure 11 shows, the pattern for brick red, pink, and light green (the medium-knowledge students) is largely the same as in Figure 9. Correctness increases fairly rapidly when students spend more time, leveling off, declining, and then coming back up a little for the highest amounts of time spent. The pattern is different for the highest-knowledge students.

The highest-knowledge students (green) essentially do not have any very rapid responses and show similarly high performance across the spectrum of time taken. This can be interpreted in at least three ways. Perhaps the highest-knowledge students do not become disengaged; alternatively, perhaps the students who never become disengaged perform better, and appear to have the highest knowledge. Or perhaps being classified by the Rasch model as having the highest knowledge requires both having the highest knowledge and never becoming disengaged.

The lowest-knowledge students (dark red) have very poor performance for low amounts of time spent. However, their performance never flattens out, although the rate of improvement slows. The more time these students spend, the better they do. Despite that, these students' performance never reaches a very high level.

One other thing that is visible in the graph is that the amount of time needed for asymptotic levels of correctness is lower for the higher knowledge students (θ above 1) than for the lower knowledge students (θ below 0). See the line B-D in the Figure, which links the asymptotic point for high-knowledge students to the near-asymptotic point for low-knowledge students. This suggests that the same cut-off for "too fast" is not appropriate for students with different ability.

4. DISCUSSION AND CONCLUSIONS

In this paper, we have investigated how the relationship between the time taken by students and their performance is mediated by student general knowledge and item difficulty. We also investigate whether different ways of representing time (standardized or non-standardized; log-transformed or non-transformed) impact our ability to recognize cut-offs and inflections in student performance. We analyze these questions by visualizing the relationship between time taken and performance under each of these different conditions.

We find that using a logarithmic scale allows for showing more data while making it easy to present the full data range while standardization allows for a fairer comparison of student time on different items. We find that the combination of these approaches facilitates identifying cut-offs and infection points in student performance.

We find that students who spend very little time on an item typically achieve low percent correct and as the time taken increases, performance improves. However, as students spend over a certain time, their performance somewhat declines again. The amount of time needed for very successful performance is different for easier and harder items and is higher for the easy items compared to very easy items. Hence, we suggest that the same cut-off for "too fast" is not appropriate for items with different difficulty levels.

Student performance declines when students spend too much time on easy and very easy items. The patterns seen are different for the higher-difficulty items. For the difficult and very difficult items, we do not observe any clear inflection point where spending more time is associated with worse performance.

As such, we can conclude that time cut-offs should not be considered independently of item difficulty. We are not aware of any models of gaming the system, carelessness, off-task behavior, or related constructs that explicitly consider item difficulty. Our results suggest that this omission is lowering the quality of these models.

In terms of student overall domain knowledge, we find that the most successful students seldom respond in very short amounts of time. As discussed above, this may reflect in part the fact

that very quick responses make the student appear generally less successful within the Rasch model. However, we also see that the generally knowledgeable students show consistently high performance for most the span of time taken, whereas the less generally knowledgeable students' performance does not level off to the same degree.

For higher difficulty items, longer time taken is associated with better students. However, this effect only manifests for the higher difficulty items; these items are more discriminative in terms of the relationship between student ability and longer times taken. In future work, we will try to correlate these longer times with students' usage of other online materials during. At present we do not have access to this level of detailed data.

These results suggest overall that models that consider student time taken during online learning, and select time cut-offs, should take student general knowledge and item difficulty into account. However, the exact cut-offs will probably differ between systems and also possibly differ with content.

It would be useful to investigate whether the findings seen here are general across other contexts. In our future work, we will investigate their generality to other textbooks, and whether the findings also generalize to other online learning platforms. It would also be useful to examine existing models depending on time cutoffs, and see whether measures of general student knowledge (perhaps average correctness so far across skills) and item difficulty can produce more accurate models of constructs like gaming the system and off-task behavior. Ultimately, this type of model may enhance the effectiveness of behavior detection, leading to more effective interventions to struggling and disengaged students. One of our upcoming steps will be to use these analyses to develop behavior detectors for our platform, that can be used to help to students who are answering too fast or who are struggling and responding slowly. We will then measure the impact of these changes on learning outcomes, to see the degree to which these approaches can enhance student learning.

5. ACKNOWLEDGMENTS

Our most sincere thanks to Mark Riedesel and Alfred Essa for supporting this research. We would also like to thank Malcolm Duncan and his team at the EZTest who helped us navigate through the database and helped create queries to pull the data required for this research.

6. REFERENCES

[1] Aleven, V., McLaren, B., Roll, I. and Koedinger, K., 2004, August. Toward tutoring help seeking. In Intelligent Tutoring Systems (pp. 227-239). Springer Berlin Heidelberg.

[2] Baker, R.S., Corbett, A.T. and Koedinger, K.R., 2004, August. Detecting student misuse of intelligent tutoring systems. In Intelligent tutoring systems (pp. 531-540). Springer Berlin Heidelberg.

[3] Beck, J. E., 2005. Engagement tracing: using response times to model student disengagement. In Artificial intelligence in education: Supporting learning through intelligent and socially informed technology, 125, p.88-95.

[4] Baker, R.S., 2007, April. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 1059-1068). ACM.

[5] Kovanović, V., Gašević, D., Dawson, S., Joksimović, S., Baker, R.S. and Hatala, M., 2015, March. Penetrating the black box of time-on-task estimation. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 184-193). ACM.

[6] Muldner, K., Burleson, W., Van de Sande, B. and VanLehn, K., 2010, June. An analysis of gaming behaviors in an intelligent tutoring system. In Intelligent Tutoring Systems (pp. 184-193). Springer Berlin Heidelberg.

[7] Paquette, L., de Carvalho, A.M.J.A. and Ryan, S.B., 2014, July. Towards understanding export coding of student disengagement in online learning. In Proc. of the 36th Annual Cognitive Science Conference (pp. 1126-1131).

[8] Cetintas, S., Si, L., Xin, Y.P. and Hord, C., 2010. Automatic detection of off-task behaviors in intelligent tutoring systems with machine learning techniques. Learning Technologies, IEEE Transactions on, 3(3), pp.228-236.

[9] Pardos, Z.A., Baker, R.S., San Pedro, M.O., Gowda, S.M. and Gowda, S.M., 2013, April. 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 (pp. 117-124). ACM.

[10] San Pedro, M.O.C.Z., d Baker, R.S. and Rodrigo, M.M.T., 2011, June. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In Artificial Intelligence in Education (pp. 304-311). Springer Berlin Heidelberg.

[11] Hershkovitz, A., de Baker, R.S.J., Gobert, J., Wixon, M. and Sao Pedro, M., 2013. Discovery With Models A Case Study on Carelessness in Computer-Based Science Inquiry. American Behavioral Scientist, 57(10), pp.1480-1499.

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

[13] Habgood, M.J. and Ainsworth, S.E., 2011. Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. The Journal of the Learning Sciences, 20(2), pp.169-206.

[14] Pavlik, P.I. and Anderson, J.R., 2005. Practice and Forgetting Effects on Vocabulary Memory: An Activation - Based Model of the Spacing Effect. Cognitive Science, 29(4), pp.559-586.

[15] Wang, Y. and Heffernan, N., 2013, July. Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In Artificial Intelligence in Education (pp. 181-188). Springer Berlin Heidelberg.

[16] Feild, J., 2015. Improving student performance using nudge analytics.Educational Data Mining.

[17] Agnihotri, L., Aghababyan, A., Mojarad, S., Riedesel, M. and Essa, A., 2015. Mining Login Data For Actionable Student Insight. In Proc. 8th International Conference on Educational Data Mining.

[18] Bridgeman, S., Goodrich, M.T., Kobourov, S.G. and Tamassia, R., 2000. PILOT: An interactive tool for learning and grading. ACM SIGCSE Bulletin, 32(1), pp.139-143.

[19] Baker, F.B., 2001. The basics of item response theory. For full text: http://ericae.net/irt/baker

[20] Bond, T. and Fox, C.M., 2015. Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.

[21] Corbett, A.T. and Anderson, J.R., 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4(4), pp.253-278.

[22] Pavlik Jr, P.I., Cen, H. and Koedinger, K.R., 2009. Performance Factors Analysis--A New Alternative to Knowledge Tracing. Online Submission.

[23] Junker, B.W. and Sijtsma, K., 2001. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. Applied Psychological Measurement, 25(3), pp.258-272.

[24] Haberman, S.J., 2006. An elementary test of the normal 2PL model against the normal 3PL alternative. ETS Research Report Series, 2006(1), pp.i-8.

[25] Pelánek, R., 2014. Application of Time Decay Functions and the Elo System in Student Modeling. Proc. of Educational Data Mining, pp.21-27.

[26] Baker, R. S., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18(3), 287-314.

[27] San Pedro, M.O.Z., d Baker, R.S. and Rodrigo, M.M.T., 2014. Carelessness and affect in an intelligent tutoring system for mathematics. International Journal of Artificial Intelligence in Education, 24(2), pp.189-210.

[28] BUlrich, R. and Miller, J., 1994. Effects of truncation on reaction time analysis. Journal of Experimental Psychology: General, 123(1), p.34.

[29] Iglewicz, B. and Hoaglin, D.C., 1993. How to detect and handle outliers (Vol. 16). Asq Press.

[30] Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A. and Koedinger, K., 2008. Why students engage in" gaming the system" behavior in interactive learning environments. Journal of Interactive Learning Research, 19(2), p.185.

[31] Mettler, E., Massey, C.M. and Kellman, P.J., 2011. Improving Adaptive Learning Technology through the Use of Response Times. Grantee Submission.