Modelling and using response times in MOOCs

Ilia Rushkin
Harvard
Cambridge, USA

Ike Chuang
MIT
Cambridge, USA

Dustin Tingley
Harvard
Cambridge, USA

ABSTRACT

DRAFT, UPDATED—August 22, 2017. Each time a learner in a self-paced online course is trying to answer an assessment question, it takes some time to submit the answer, and if multiple attempts are allowed and the first answer was incorrect, it takes some time to submit the second attempt, and so on. Here we study the distribution of such "response times". We find that the log-normal statistical model for such times, previously suggested in the literature, holds for online courses qualitatively. Users who, according to this model, tend to take longer on submits are more likely to complete the course, have a higher level of engagement and achieve a higher grade.

INTRODUCTION

When users interact with assessment questions in an online course, the data that usually receives the most attention is the answers they submit, and sometimes only the correctness of these answers or the received score. But the time spent by the user on the question is also important: it is arguably the most readily acquired data that reveals something about the process by which a user arrived at an answer. Analyzing these "response times" allows one to quantify some properties of the questions (how long does a question typically take and how much it varies) as well as some properties of the users (how slow they tend to be in answering questions). The question properties have implications for course design, and the user slowness may be related to the user's ability and preferred mode of interaction with the course. Extracting such parameters necessitates a parametric statistical model for the response times. This is similar to how in IRT (item response theory) an item response function is needed for extracting question parameters and users' abilities from the response correctness data [2]. The appropriate choice of a parametric model for the response times is important.

We evaluate the appropriateness of using a log-normal distribution to model the time it takes an online course user to submit an answer to an assessment question.[1] On the basic level, we may choose it as one might choose the normal distribution to model any histogram with a relatively unskewed bell shape. The log-normal distribution is a model of convenience: it is familiar, easy to work with, and has qualitatively correct features: no negative values in the domain, a peak, and a long tail on the right. But there is a likely and deeper reason for log-normality. The central limit theorem is the reason for the ubiquity of normal distributions in nature, because observed quantities are often the result of addition of many independent random contributions. Should such contributions be multiplicative, rather than additive, they would give rise to a log-normal distribution, and this seems a reasonable idea when dealing with response times.

To see the multiplicative nature of the process of responding to assessment questions, suppose there is a certain basic response time t_B for a user-question interaction (longer for harder questions and for slower users). The actual response time is affected by a large number of diverse factors, such as having to think about different aspects of the question, calculations, looking up information, fatigue, distractions, etc. The extra time taken up by any factor should scale with the difficulty of the question and with the overall slowness tendency of the user, i.e. with the basic time t_B. Therefore, it is natural to assume that the effect of each such factor is multiplicative: the factor i multiplies the basic response time by $(1+r_i)$, where r_i is a random variable ("rate"), resulting in the response time $t = t_B \prod_i (1+r_i)$. In such a paradigm, the central limit theorem predicts that the distribution of t will approach the log-normal distribution when the number of contributing factors is large.[2]

User slowness, extracted from the response times with the help of the log-normal model, is an interesting and little-used parameter. It can be interpreted in two fundamentally different ways. In the first, taking a longer time is viewed as a sign of user's lower mastery, and in the second it is viewed as a sign of diligence and thoroughness. In the first interpretation, higher slowness should be associated with lower achievement, while in the second the opposite is expected. Which interpretation applies, depends on the context. Thus, if users solve problems in a timed exam, the first interpretation could be more likely, but in a self-paced MOOC environment – the second. Indeed, we find in our study of MOOCS that higher slowness is positively correlated with some measures of engagement and achievement. Moreover, the interpretation of slowness as thoroughness suggests causality rather than just a

[1] All code for analysis in this paper will be made open source.

[2] We assume that the conditions of the theorem are fulfilled. In practice, the most vulnerable condition of the theorem is that the variables $x_i = \ln(1+r_i)$ should be independent, or at least not universally non-independent (they could form distinct independent groups with high internal correlation, but then the number of such groups needs to be large), which is the mathematical expression of the assumption that the nature of the variables is diverse.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

correlation. If so, user slowness in a self-paced online course is a desirable quality, somewhat like the latent ability in the item response theory. While it may be difficult to intervene on low latent ability, in the user-slowness analysis the situation seems more straightforward: we can imagine an intervention, in which a user who is going through a course too fast receives a recommendation to slow down.

MODEL DESCRIPTION

Following [8] and [1], we model the response time logarithms as independent normally distributed variables with probability density

$$P(\ln t_{qu}) = \frac{\alpha_q}{\sqrt{2\pi}} \exp\left(-\frac{\alpha_q^2}{2}(\beta_q + \zeta_u - \ln t_{qu})^2\right), \qquad (1)$$

where $q = 1, 2, ...N_q$ is a question, $u = 1, 2, ...N_u$ is a user, and t_{qu} is the response time of user u on question q. Conceptually, this model is somewhat analogous to item response theory. There too, the user-question interaction is modeled by criss-crossing a set of user parameters (latent ability) and a set of question parameters (discrimination, difficulty, and possibly guess and slip probabilities). In Eq. 1, the question parameters α_q and β_q can be interpreted as a type of discrimination and "time intensity" (a type of difficulty measure), whereas the user parameter ζ_u is the user slowness. There is a freedom of shifting all β's and all ζ's by opposite constants without affecting the probability distribution. We fix this freedom by imposing the condition $\sum_u \zeta_u = 0$. Thus, if the response times are measured in seconds, $\exp(\beta_q)$ is the question q's characteristic response time in seconds, and $\exp(\zeta_u)$ is the multiplicative factor by which user u's response times tend to differ from those characteristic response times.[3] The question discriminations α_q measure the size of random fluctuations of the response times around the expectation values: high discrimination means small fluctuations and vice versa.

In this way, the model is defined by $2N_q + N_u - 1$ free parameters, and the number of observed values t_{qu} scales as $N_u \times N_q$ (more or less, since not all users respond to all questions). For substantial numbers of users and/or questions the number of observations will be much greater than the number of parameters, making it possible to fit the parameters by maximizing likelihood. Namely, given the observed response times t_{qu} in the set of observations $(q, u) \in \mathscr{O}$, we find the parameters of the questions and the slowness of the users via minimization of negative logarithmic likelihood:

$$\{\alpha, \beta, \zeta\} = \operatorname{argmin} \sum_{(q,u)\in\mathscr{O}} \left(\frac{\alpha_q^2}{2}(\beta_q + \zeta_u - \ln t_{qu})^2 - \ln \alpha_q\right) \quad (2)$$

METHODS

We performed the minimization from Eq. 2 using a non-linear conjugate-gradient routine for 47 diverse HarvardX courses from 2015-2017. Among these, there were 16 STEM courses (natural and health sciences, computing and programming) and 31 non-STEM courses (humanities, law, social sciences). To reduce the number of responses from non-committed users, we restricted the data to those users who visited at least half of the chapters in a the course (a standard measure in HarvardX data analysis, where such users are said to have "explored" the course [3]). Further, we considered only the questions for which we see no more than 5 attempts submitted by a user: questions with a large or even unlimited number of attempts might provoke a different, guess-driven behavior. We wanted to avoid questions responded by only a few users, and so we considered only questions with at least 10 users attempting them. Similarly, we wanted to avoid users who responded to only a few questions, and to this end we imposed 10 question minimum here as well, except that in some courses there were not enough questions, and we find no users with 10 questions. In this case, we lowered the cutoff to the maximum encountered value. We can call the questions and users who remain in the data after this procedure "qualified".

Since questions in HarvardX courses often allow multiple submit attempts, we attempted to fit the model in each course on 1st and 2nd submits separately, taking care to include 2nd submits only if the response on the 1st submit was incorrect (2nd responses after the correct 1st are understandably rare, but they do occur, and we remove them from the data). For both 1st and 2nd submits, we fit the model to three different groups of submit events: only correct, only incorrect, and of any incorrectness, thus producing up to 6 model fits in a course.

Our definition of the 2nd response time is simply the difference between the timestamps of the 1st and the 2nd submit clicks, assuming that the user starts thinking about the second answer immediately after seeing that that first one was incorrect.

Calculation of the 1st response time requires more care. In principle, it is the difference between the time the question was served and the timestamp of the 1st submit click, and the role of the serve time can be played by the timestamp of the user loading the question page. The challenge is that sometimes multiple questions are served on the same page, and since a typical user works through them in a sequence, the common page-loading timestamp of these questions of the course will artificially lengthen the 1st response time for all the questions except the one on which the user worked first. Our strategy of resolving this problem can be described as follows: in case of multiple questions on a page, assume that the user starts working on a question after the chronologically last submit click on a different question from the same page. Namely, suppose we observe in the data that, for a given user, a group of questions have the same page-loading timestamp t_0, and some submit timestamps arranged and indexed chronologically: $t_0 < t_1 < t_2 < ... < t_p$. These submit events belong to different questions, possibly with multiple submits on a question, and it is not assumed that the user works on questions completely sequentially (e.g. it can be that t_1 is 1st submit on question A, t_2 1st submit on question B, t_3 is 2nd submit on question A again.). If the timestamp of the 1st submit for one of the

[3] The mean time intensity across questions equals the mean expected logarithm of response times across all questions and users: $N_q^{-1} \sum_q \beta_q = N_q^{-1} N_u^{-1} \sum_{q,u} (\beta_q + \zeta_u)$.

questions is t_i ($i > 0$), then the 1st response time for this question is calculated as $t_i - t_{i-1}$.[4]

After preparation, the data for each of the 6 model fits (correct/incorrect/any responses on 1st/2nd attempts) in a course is in the form of an $N_u \times N_q$ matrix, where each row is a qualified user, each column is a qualified question, and the entries are the natural logarithms of times in seconds, or missing. Since measurement of 2nd response times uses fewer assumptions, it may seem more reliable. However, 2nd responses occur only when the question allows more than one attempt and the 1st response was incorrect (which in case of a partially-correct answer involves an extra dichotomizing step), meaning a smaller and possibly skewed data sample. For these reasons, we regard 1st response times of any correctness as the most valuable subset of data. Its data matrix is guaranteed to have the biggest dimensions and the most data. Other matrices contain fewer observations. Convergence on the data from 1st responses of any correctness was achieved in 45 out of 47 courses, but only in 21 of them on the data from 2nd responses (also of any correctness). When aggregating the data across courses, we include only the converged fits.

Table 1 lists some parameters related to the amounts of data available.

Table 1. Dataset parameters for 1st and 2nd responses of any correctness, across courses. N_u is the number of users (rows) in the data, N_q is the number of questions (columns), m is missingness (the fraction of missing matrix entries) and $r = (2N_q + N_u - 1)/(N_u N_q (1-m))$ is the ratio of the number of fit parameters to the number of observations. Only the cases where convergence was reached are included.

	min	median	max
N_u, 1st/2nd	13/13	567/674	3,055/7,628
N_q, 1st/2nd	7/17	61/63	447/316
m, 1st/2nd	0.03/0.62	0.25/0.76	0.72/0.89
r, 1st/2nd	0.009/0.035	0.035/0.085	0.171/0.674

ASSESSING MODEL QUALITY

After the fit, we check how close to log-normal the distribution of response times is by forming the variables $x_{qu} = \alpha_q(\ln t_{qu} - \beta_q - \zeta_u)$. The model assumes that these should be standard normal variables, and so we can plot the observed cumulative distribution (percentile curve) CDF(x) vs. the cumulative distribution of the standard normal variable $\Phi(x)$. The result is in Figure 1, where we list the first four moments with respect to the origin: $m_k = (1/|\mathcal{O}|)\sum_{(q,u)\in\mathcal{O}}(x_{qu})^k$ (the standard normal distribution has $m_1 = 0$, $m_2 = 1$, $m_3 = 0$, $m_4 = 3$). We see the model as qualitatively correct, and although deviations are sizeable, it should be kept in mind how big the range of times in it is: the interquartile range is from 30 to 670 seconds for 1st response times and from 6 to 44 seconds for 2nd response times. At the very least, it is clear that time logarithms are suitable variables for analysis: the skewness

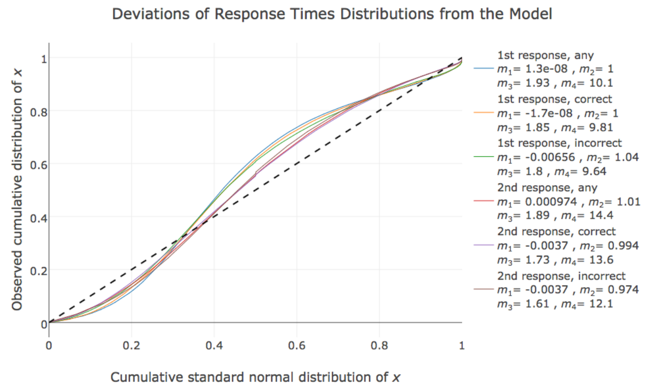

Figure 1. Comparison of the observed cumulative distribution of the values x_{qu} to the standard normal distribution $\Phi(x)$, predicted by Eq. 1. The identity line (shown in dashed black) represents the ideal agreement with the model. The listed distribution moments are calculated with respect to the ideal mean 0.

of distribution of response times themselves is extreme, but logarithmic transformation accounts for virtually all of it.

The curves in Figure 1 appear to form two groups based on the submit number, whereas the submit correctness has a lesser effect. In essence, we can focus on the data coming from 1st and 2nd submits of any correctness, and use the correctness-specific data get the idea for the uncertainty size. The distribution of the 1st response times has a much smaller excess kurtosis and skewness than the distribution of the 2nd response times, and almost perfect first and second moments. In all cases the skewness and the excess kurtosis are positive (meaning that the sample distribution has heavier tails than the model predicts).

Selecting one row or column in the data matrices gives the distributions of x_{qu} by question or by user. To quantify how frequently large deviations from Eq. (1) occur, we calculate the deviations of the first four moments of the distributions by question from the standard normal distribution, namely the quantities $d_k = (m_k)^{1/k} - (m_k^{(0)})^{1/k}$ for $k = 1,2,3,4$, where $m_k^{(0)} = (0,1,0,3)$ are the central moments of the standard normal distribution. Thus, d_1 is the mean, d_2 – the excess of standard deviation, d_3 – skewness, and d_4 – a measure similar to excess kurtosis. In Figure 2 we plot the percentile curves for these quantities.

In a course, there is some overlap in users and questions in the data from 1st and 2nd responses of different correctness, which allows comparing the parameters α_q, β_q for the same question (or the slowness ζ_u for the same user) but obtained from different subsets of data. Using a STEM course with high degree of overlap as a representative example, we plot the model parameters found in it from 1st responses of different correctness, and from 1st and 2nd responses (Figures 3, 4). Here too, submit correctness is not a major factor: the points in Figure 3 cluster around the $y = x$ line and show substantial correlation, although the correlation of α_q always proves to be the lowest of the three (hence, whatever effect the correctness has, it is primarily on the degree of variability in the response

[4]We do not impose any timeout cutoff on the response times. Only about 7% of 1st response times and 0.9% of 2nd response times in our entire dataset exceed 24 hours. The median response times in the dataset are 112 seconds for 1st reponses and 17 seconds for second responses.

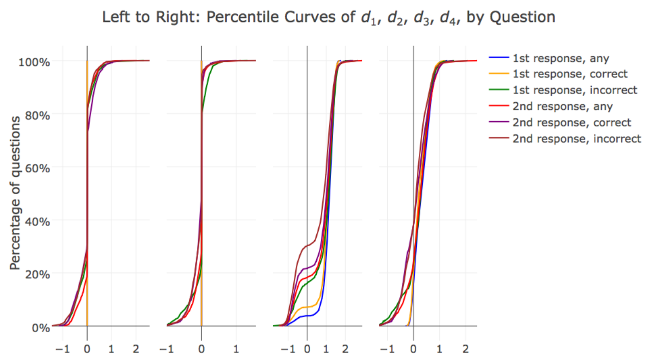

Figure 2. Parameters of distributions observed for response time logarithms for each question. Ideal agreement with the model would mean all curves collapsing onto unit-step functions. Note that this virtually happens for the two of the 1st submit curves in the first two plots.

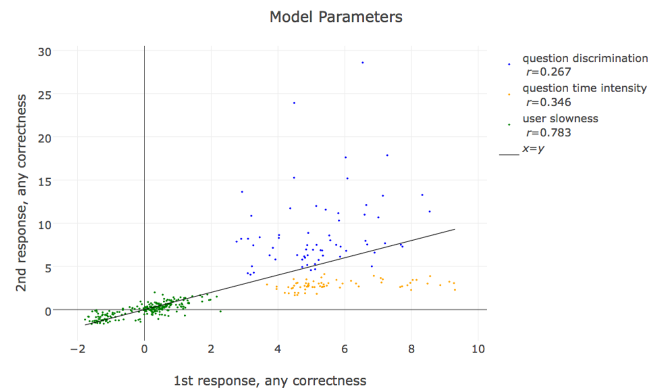

Figure 4. A STEM course example. Model parameters obtained from 1st submits and 2nd submits, of any correctness. Blue points are question discriminations α_q, multiplied by 10 for better visibility. Yellow points are question time intensities β_q. Green points are user slownesses ζ_u, multiplied by 0.5. The r values are the correlations of values on the x and y axes.

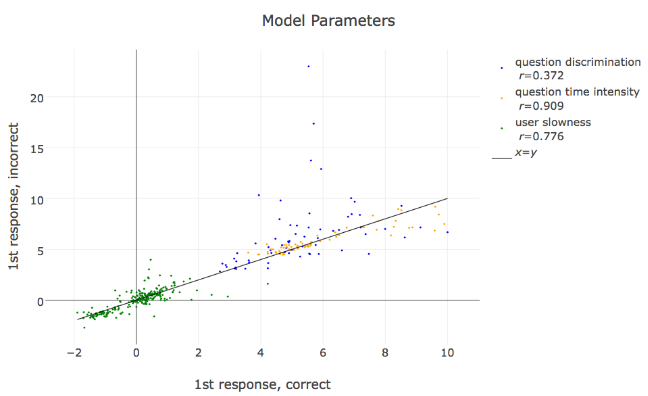

Figure 3. A STEM course example. Model parameters obtained from 1st submits, correct vs. incorrect submits. Blue points are question discriminations α_q, multiplied by 10 for better visibility. Yellow points are question time intensities β_q. Green points are user slownesses ζ_u, multiplied by 0.5. The r values are the correlations of values on the x and y axes.

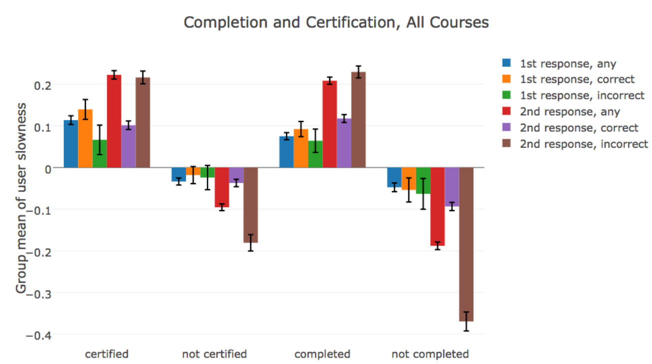

Figure 5. Group means of slowness for users depending on completion and certification in the course. Data from all courses is treated together. The error bars represent one standard error.

times). Correlations between 2nd correct and incorrect submits are lower, but otherwise the picture is similar. On the other hand, the difference between 1st and 2nd submits is big (Figure 4). We expect that the typical time spent on the second submit is much shorter than on the first, so it is no surprise that the points for time intensity cluster well below the $x = y$ line, and that their trend has slope much less than 1. Less predictably, the discrimination tends to increase on the 2nd responses (less variability in the 2nd response times).

While the differences between 1st and 2nd submits may seem moderate in the plots, it is worth reminding that these relate to time logarithms. Exponentiated, they translate into very sizeable time differences. Thus, in this course data the median response time was 123 seconds on the 1st submit, and 4 seconds on the 2nd submit.

USER SLOWNESS

We find that user slowness is correlated with measures of user engagement and success in the course: course completion, certification and final grade, number of explored course chapters, viewing videos in the course, posting on the course forum, and viewing correct answers for assessment questions. Figure 5 shows the mean[5] user slowness in different success-level groups (completion is defined as getting a grade greater or equal than the passing grade, set by the course instructor). We have already excluded users who did not explore courses or interacted with few questions, so the results are not dominated by users not committed to learning. We see that achieving certification or completion is related to the tendency to take longer on questions.

The same result holds qualitatively on the course-by-course level, although the exact slowness values may differ. Figures 6-7 show the same data as Figure 5, but splitting the courses by subject matter.

[5]Thanks to the use of time-logarithms in our model, the distribution of user slownesses is rather symmetric and resembles the normal distribution, which justifies the use of mean slowness for cross-group comparisons.

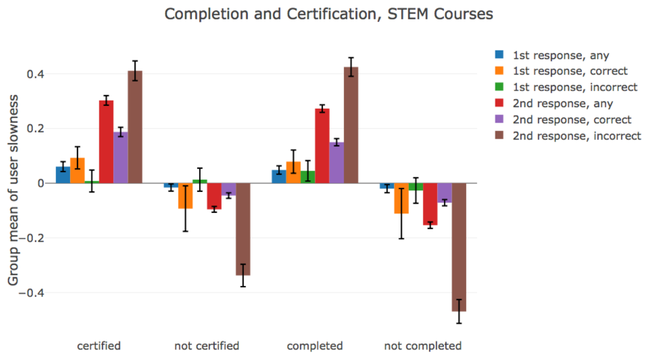

Figure 6. Group averages of slowness for users depending on completion and certification in the course. STEM courses only. The error bars represent one standard error.

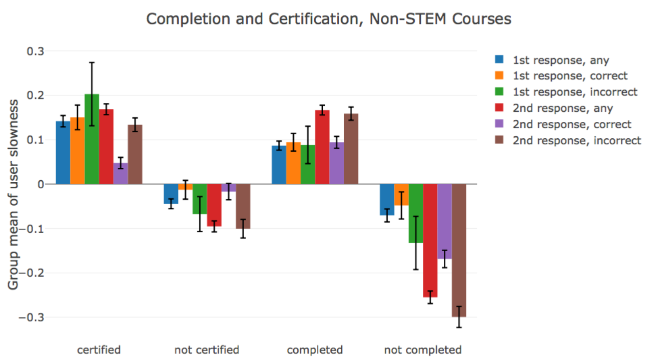

Figure 7. Group averages of slowness for users depending on completion and certification in the course. Non-STEM courses only. The error bars represent one standard error.

An alternative to Figures 5-7 way of presenting the relationship between slowness and engagement is a logistic regression with certification or completion as the dependent variable, i.e. modeling $p/(1-p) = a \cdot \exp(b\zeta)$, where p is the completion (or certification) rate and a, b are constants, so that $\exp(b)$ represents the percentage change in the odds $p/(1-p)$, associated with the increase of slowness by 1 (Table 2).

Table 2. Percentage increases in the odds of getting certification (O_{cert}) and the odds of completing the course (O_{comp}), associated with an increase by 1 in user slowness on 1st/2nd responses. The values are obtained by one-variable logistic regression.

	increase in O_{cert}	increase in O_{comp}
STEM, 1st/2nd	8%/41%	7%/49%
Non-STEM, 1st/2nd	17%/36%	16%/76%
All, 1st/2nd	14%/36%	12%/53%

Figure 8 shows the relation between slowness and final course grade (courses were graded on 0-to-1 scale[6]). The relationship is generally neutral on first submits and positive on second

[6]In a few instances grades above 1 were assigned in two courses. Removing these instances from the dataset does not affect our results.

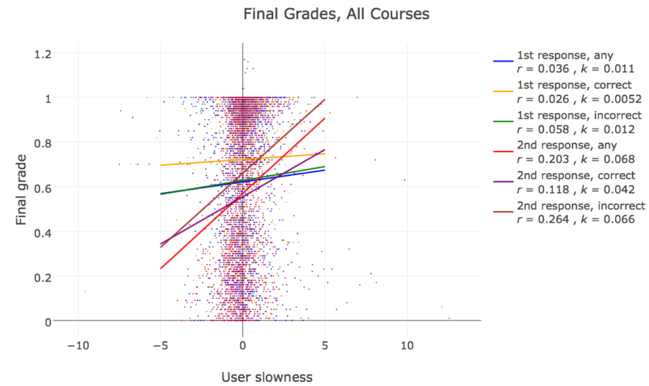

Figure 8. User slowness vs. final grade. About 5% of data points are shown. Linear regression lines are calculated using all data points. The r values are the correlation coefficients between user slowness and final grade. The k values are the regression slopes, interpreted as a gain in grade associated with taking $e \approx 2.7$ times longer to respond.

submits, which could imply that users with higher grades are those who pause and think after an incorrect answer, rather than hurrying to make a change and resubmit. Inspecting the STEM and non-STEM courses separately, we find very similar results. For instance, the slopes of the regression lines, listed in the same order as in Figure 8, are $k = [0.0075, 0.013, 0.0096, 0.069, 0.043, 0.072]$ in STEM courses and $k = [0.011, 0.0045, 0.017, 0.064, 0.040, 0.061]$ in non-STEM courses.

We found no significant correlation between slowness and the level of prior competency, as measured by the user's self-reported number of online courses taken previously, or by the self-reported level of fluency in English (all courses in our dataset are in English). On the other hand, slowness is consistently correlated with the user's age (Pearson correlation of ≈ 0.17 for both STEM and non-STEM courses, and based on 1st and 2nd responses of any correctness). In STEM courses, the correlation is noticeably higher when restricted to only the course-completing users: ≈ 0.3 for 1st responses and ≈ 0.22 for 2nd responses of any correctness. Furthermore, the course-completing users in STEM courses show some correlation of slowness with the self-reported highest level of education: Spearman correlation coefficient of 0.2/0.125 for 1st/2nd responses of any correctness (inclusion of non-completing users causes these correlation values to drop to ≈ 0.12; and in non-STEM courses these values are low both for all users and for course-completing users, in the 0.06-0.09 range).

The link between slowness and the main measures of user engagement – course completion and certification – was shown above. We also examined the correlation of slowness with several other measures of user engagement, such as the numbers of unique videos viewed by a user in the course, of video "play" events, of forum posts, of visited course chapters, of submitted question responses, of "Show answer" clicks. To make these comparable across courses, we normalized these measures in each course to mean value 1 prior to computing the correlations. To reduce the confounding dependence on user's mastery and drop-out, in Table 3 we compute correla-

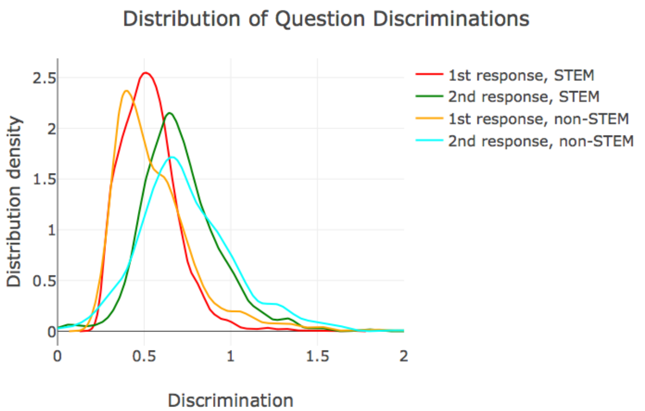

Figure 9. Distribution density of question discriminations α_q. Responses of any correctness.

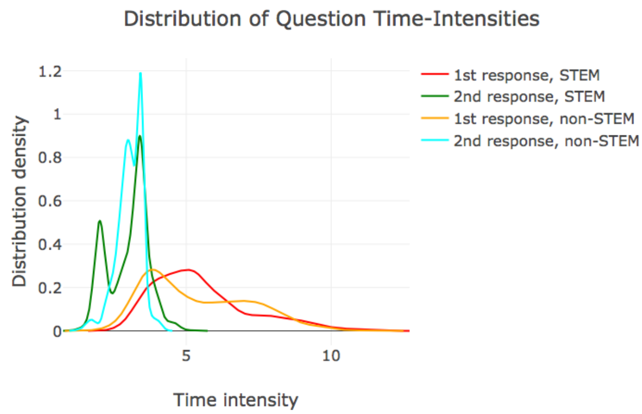

Figure 10. Distribution density of question time intensity β_q. Responses of any correctness.

tion only for the users who completed the course, and only for 1st and 2nd responses of any correctness. Nevertheless, correlations with these variables are ambiguous in interpretation, since a response time might be long precisely because the user spent it re-watching a video or posting on the course forum. Apart from the number of question submits (which appears uncorrelated with slowness), we observe higher correlations with slowness on 1st responses.

Table 3. Correlations of several measures of user engagement with slowness. Only the course-completing users are included. The values separated by "/" are computed with slowness in 1st/2nd responses of any correctness.

	STEM	non-STEM
unique videos	0.25/0.10	0.21/-0.07
video "play" clicks	0.21/0.13	0.13/0.04
forum posts	0.11/0.07	0.21/0.08
chapters visited	0.13/0.06	0.11/0.07
question submits	0.03/0.06	0.02/0.01
"Show answer" clicks	0.25/0.12	0.20/0.08

The overall conclusion is that greater slowness is associated with higher achievement (measured by grades) and engagement (measured by rates of completion and certification, watching videos).

QUESTION DISCRIMINATION AND TIME INTENSITY

Figures 9, 10 show the distribution densities (obtained by gaussian-kernel smoothing) for α_q and β_q, as obtained from all converged data subsets from all courses. These reiterate the conclusions made from Figures 3-4, but now on all courses globally: on the second submit the time intensity tends to be smaller (i.e. the second responses tend to be much quicker), but the discrimination is higher (although, also has a broader distribution across questions). The distributions of time intensities on 2nd submits are bimodal, with smaller peaks at $\beta = 2$ for STEM and $\beta = 3$ for non-STEM courses. Exponentiated, these correspond to typical response times of 7 and 20 seconds. These peaks are due to the questions, where user tends to make a small change in the answer and quickly resubmit (e.g. the sign change in the numeric answer in a STEM course).

The main distribution peaks lie near $\beta = 3.5$ ($\exp(\beta) \approx 33$ seconds) for all courses.[7]

The median time intensity of a question across all courses is 5.098 and 3.155 on the 1st and 2nd responses (of any correctness). Exponentiated, these become the user-averaged typical response times: 164 seconds and 23 seconds, respectively. The median question discriminations on 1st and 2nd responses (of any correctness) are 0.511 and 0.691. To remind, discrimination is the inverse standard deviation of the distribution of time logarithms on a given question. Exponentiating the inverses of these values, we find 7.08 on the 1st submits and 4.25 on the 2nd. For instance, we can describe in rough terms (made precise by Eq. 1 and the medians of distributions in Figures 9-10), the situation in 2nd submits as follows. After an unsuccessful 1st submit on a typical question, a typical user is expected to spend 23 seconds before submitting it again. The user variability is such that for most users the actual time lies in the range between $23/4.25 \approx 5$ seconds and $23 \cdot 4.25 \approx 98$ seconds. Obviously this is a broad range, which after all is the reason for us to choose time logarithm, rather than time, as the model variable.

It is an interesting question what relationship can be expected between α_q and β_q in general: what part of the observed increase in α_q is due specifically to the 2nd attempt on the question, and what part is a simple corollary of the lower β_q, observable for quicker questions even on the same submit attempt. Indeed, we find in our data that on 1st responses β_q and $1/\alpha_q$ are positively correlated ($\rho = 0.63$). In other words, if a question's response times are shorter on average, so is the multiplicative spread in them. A likely explanation is that quicker questions are not just scaled-down versions of slower ones. They are of a simpler nature, less open-ended

[7] We also repeated the data analysis using the page-load timestamp to calculate 1st response times, i.e. ignoring the fact that this extends the response times when multiple questions are served on the same page. In this case, the 1st response time intensity distribution also becomes bimodal, but for a different reason. The main peak is around $\beta = 8$ and the secondary peak was around $\beta = \ln(24 \cdot 3600) \approx 11.4$, due to users loading a page with multiple questions and working on some of them the next day.

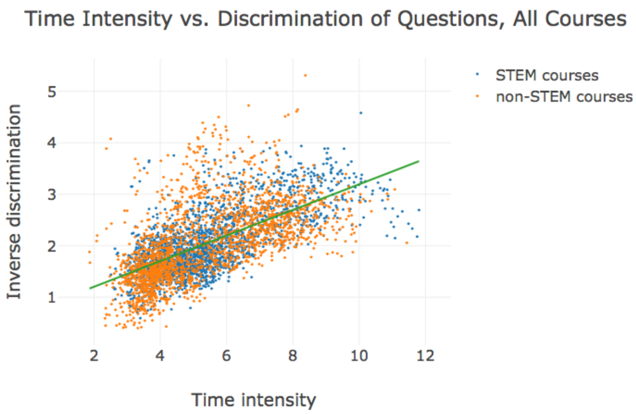

Figure 11. A plot of β_q vs. $1/\alpha_q$, for all courses. Data from 1st responses of any correctness.

or with fewer alternative solution paths, which decreases the variability in response times.

RELATED LITERATURE

The use of logarithms of response times (rather than response times themselves), and of somehow fitting the time data with a log-normal distribution, is at least as old as 1983 ([7]), in which study response time logarithms are combined with the parameters of the item response theory, in order to model the trade-off between speed and accuracy, as well as the relation between time intensity of question and its difficulty. Other studies ([6],[9]) modeling response times without incorporating the response variables, such as IRT parameters. The specific form of the model used here as Eq. 1 was first investigated in [8].

It should be noted that time distributions other than log-normal were also tried. Notably, [4] and [5], where the distribution is taken to be exponential, i.e. $P(t_{qu}) = \lambda_{qu} \exp(-\lambda_{qu} t_{qu})$, where the distribution parameter is taken to be a sum of a question-specific parameter and a user-specific parameter: $\lambda_{qu} = \theta_u + \varepsilon_q$. The implied assumption is that the problem-solving process is modeled as waiting for an epiphany, which can occur equally likely at any time (probability $\lambda_{qu} dt$ for any infinitesimal time interval dt), and the problem is submitted as soon as it happens. This may be appropriate for some types of mental activity, but clearly not for submitting questions in an online course: in these the observed distributions of response times invariably have a qualitatively log-normal, not exponential, shape.

Originally, the log-normal model of response times was developed for test items. To our knowledge, it was first applied to assessment items in MOOCs in [1], where the main direction of the investigation is in linking the user slowness and IRT latent ability.

CONCLUSION

The described log-normal model of response times can be used for estimating the characteristics for both the learners (slowness) and the assessment questions (discrimination and time intensity). We find that higher user slowness is linked to higher achievement and engagement levels. This poses an experimental question: is this relationship causal, and if so, what is the direction of the causality? This can be resolved by interventions, encouraging users to slow down when necessary, and if such interventions have a measurable positive effect, they can be made a recurring MOOC feature.

The question characteristics, extracted from the model, may be useful in course design. Questions are commonly transferred, with no or minimal alterations, from one version of a course to the next. Examining the time intensity and discrimination of each question in the previous course version can inform these decisions: special attention may be paid to questions that are outliers either in time intensity or in discrimination.

It was mentioned in the above that convergence is not always achieved in Eq. 2. In the data analysis we omitted all non-convergent cases, but they are, in fact, also informative for the course design purposes. Inspection shows that if the minimization did not converge after a substantial number of iterations (we tried from 500 to 2,500), this was due to having relatively few clear outliers – the questions prompting unusual learner behavior. Typically, the outliers were more pronounced in the discrimination than in time intensity. In other words, in case of non-convergence, a plot such as Figure 3 would contain a few discrimination-points very high up (and if enough of these outlier questions are removed from data, convergence will be achieved on the remaining data). Hence, non-convergence is not an obstacle for determining outlier questions, which deserve a second look from the course developer, but rather the most transparent way to see them.

REFERENCES

1. Maria Bertling and Isaac Chuang. 2015. Response time as a measure of motivation and applied effort in MOOCs. (2015). In preparation.

2. Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*. Vol. 2. Sage.

3. Andrew Dean Ho, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. 2014. HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. (2014).

4. Hartmann Scheiblechner. 1979. Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology* 19, 1 (1979), 18–38.

5. Hartmann Scheiblechner. 1985. Psychometric models for speed-test construction: The linear exponential model. *Test design: Developments in psychology and psychometrics* (1985), 219–244.

6. Deborah L Schnipke and David J Scrams. 1999. *Representing response-time information in item banks*. Vol. 97. Law School Admission Council.

7. David Thissen. 1983. Timed testing: An approach using item response theory. *New horizons in testing: Latent trait test theory and computerized adaptive testing* (1983), 179–203.

8. Wim J van der Linden. 2006. A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics* 31, 2 (2006), 181–204.

9. Wim J Van Der Linden, David J Scrams, and Deborah L Schnipke. 1999. Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement* 23, 3 (1999), 195–210.