Predictiveness of Prior Failures is Improved by Incorporating Trial Duration

Luke G. Eglington
University of Memphis
lgglngtn@memphis.edu

Philip I. Pavlik Jr
University of Memphis
ppavlik@memphis.edu

In recent years, there has been a proliferation of adaptive learner models that seek to predict student correctness. Improvements on earlier models have shown that separate predictors for prior successes, failures, and recent performance further improve fit while remaining interpretable. However, students who engage in "gaming" or other off-task behaviors may reduce the predictiveness of learner models that treat counts of prior performance equivalently across gaming and non-gaming student populations. The present research evaluated how sub-groups of students that varied in their potential gaming behavior were differently fit by a logistic learner model, and whether any observed differences between sub-groups could inspire the creation of new predictors that might improve model fit. Student data extracted from a college-level online learning application were clustered according to speed and accuracy using Gaussian mixture modeling. Distinct clusters were found, with similar cluster patterns detected in three separate datasets. Subsequently, each cluster was separately fit to a Performance Factors Analysis model (PFA). Significantly different parameter coefficients across clusters implied that students more likely to have been gaming benefitted less from prior failures. These differences inspired new and modified predictors that were found to improve overall model fit - an improvement that varied in magnitude across clusters. The present findings indicate that incorporating trial duration into counts of prior failures can improve the predictive power of learning models.

Keywords: learning, learner models, Performance Factors Analysis, clustering, student variability, feature engineering

1. INTRODUCTION

A major research focus in educational data mining involves developing models to estimate the probability that a student will correctly answer some future question. Several models have been developed, leveraging information about performance on prior test items to infer the students' knowledge level (e.g., Ayers and Junker, 2006; Cen et al., 2006; Corbett and Anderson, 1992; Galyardt and Goldin, 2015; Gong et al., 2011; Pavlik Jr. et al., 2009; Piech et al., 2015). The actual knowledge and motivations of students are of course unknowable, and thus models also benefit from drawing on cognitive theories of learning and behavior to infer future performance from prior behavior.

The present work demonstrates how the future performance of sub-groups of students can be predicted to different extents by common features of a logistic regression adaptive learning model, and how trial duration can be used to address these differences and improve model fit. The intuition for why this may be true is relatively straight-forward – the success of some features of these models depends on behaviors and motivations that may vary substantially

R305A190448; IES : National Center for Education Research

across students. Previous research has demonstrated how distinct clusters of students can have widely variable patterns of behavior in learning systems (Desmarais and Lemieux, 2013). Clustering students has also helped develop separate models for distinct sub-groups, which have been shown to improve predictive accuracy (Pardos et al., 2012). Student clusters with distinctly off-task behavior may be predicted particularly poorly by specific features. For instance, some measures (such as a count of prior incorrect answers) may predict better future performance because when a student finds they are incorrect, they are more likely to attend to related information (Baker et al., 2004; Kornell et al., 2009), which is not true if off-task behavior interrupts restudy. However, student learning goals may not be aligned with those of the researcher or learning system developer (especially outside of a laboratory). This misalignment may result in large differences in the predictive utility of common features of learner models. We evaluated how the behavior of distinct sub-groups could lead to reduced utility of particular predictive features (such as counts of prior failures), and we then used that information to develop an improved learner model.

We will begin by describing some well-known learner models and explanations for their efficacy. Subsequently, we will describe research on "gaming" behavior (Baker, Corbett and Koedinger, 2004), how such behavior could negatively impact performance, as well as how it may influence the predictive utility of learner models. We then describe how accounting for gaming behavior may influence learner model predictiveness, which motivated our creation of new features that incorporated trial duration information.

2. LEARNER MODELS

Many modern models of learning are built upon the framework of item response theory (IRT) models, which aim to predict performance on dichotomous outcomes (successes or failures) by estimating a skill parameter for each student (Rasch, 1961). The fit of a learner model can be further improved by grouping items based on shared underlying procedural rules or information (Fischer, 1973), which are also called knowledge components (KCs). Many modern logistic regression models utilize this approach, such as the Additive Factors Models (AFM; Cen, Koedinger and Junker, 2006), Performance Factors Analysis (PFA; Pavlik Jr., Cen, and Koedinger, 2009), and Recent-Performance Factors Analysis (R-PFA; Galyardt and Goldin, 2015). These types of models differ by how they use prior performance to predict future performance. AFM, for example, counts the number of prior practice attempts for a given KC, the assumption being that more practice attempts with a KC will lead to better future performance. However, a common theme to describe the evolution of logistic regression learner models is that not all learning events are equally meaningful (e.g., Chi et al., 2011; Pavlik Jr. et al., 2011).

The PFA model differs from AFM by how it distinguishes between successful and unsuccessful trials. Successful trials (e.g., applying the appropriate procedural rule or recalling the correct information) may benefit later performance differently than failed attempts and are more efficient (Carrier and Pashler, 1992; Izawa, 1970; Pavlik Jr. and Anderson, 2008). Furthermore, correct responses for a KC indicate more knowledge prior to the testing event (Pavlik Jr., 2007). In addition to failed attempts being inefficient (in terms of practice time), reviewing feedback may not necessarily elicit the same cognitive processes necessary for successful responding later (Morris, Bransford, and Franks, 1977). But failed attempts may also benefit learning if feedback is provided (Kornell, Hays, & Bjork, 2009). Furthermore, errors made with high confidence are better corrected from feedback than those made with low

confidence (Butler et al., 2011). In short, both successful and unsuccessful trials can improve later performance, but they are likely doing so in different ways. By partitioning counts of prior practice according to success or failure, the PFA model typically outperforms AFM. This partitioning also allows greater insight into what types of practice are most helpful for different types of material. Failures at tests may impact learning of some materials more than others (e.g., Kelly et al., 2015). The R-PFA model, taking a mechanism from related research (Gong, Beck and Heffernan, 2011) includes a parameter that weighs recent performance more heavily, based on the notion that recent correct answers about a particular KC are more informative than older responses. This idea is clearly reasonable if one considers that learning may entail sudden increases in understanding, rather than being a purely gradual process (Baker et al., 2011). One implementation of the recency predictor is the number of prior successful trials divided by the number of total trials, with an exponential decay parameter d. A $d = 1$ provides equal weight to previous trials, with lower d resulting in more weight to recent trials.

Another prominent approach to student modeling is Bayesian Knowledge Tracing (BKT, Corbett and Anderson, 1992). The original version of this model assumed knowledge had two possible states (known or unknown), that students learned at the same rate, and that each KC could be defined by its own set of parameters. Two important distinctions between BKT and logistic models are how they can be modified and the interpretability of parameter estimates from the models, since BKT uses failure to infer the knowledge state directly. Within the framework of logistic regression, the additive nature of predictors makes it easier to interpret the effect of individual predictors and also makes it simpler to add or subtract them and interpret the resulting change in fit. The relative influence of particular predictors across clusters was an important aspect of our research, and thus we chose to utilize logistic regression rather than BKT.

Although the aforementioned logistic regression models typically consider the quantity of prior practice as a linear effect, it is important to note that in many cases log-transforming the counts of prior practice improves model fit (e.g., Chi, Koedinger, Gordon, Jordan, and VanLehn, 2011). The benefit of this transformation is explained by the observed power-law relationship between performance measures and practice (Newell and Rosenbloom, 1981). In the present analyses, we applied log transformations to predictor variables derived from counts of prior successes or failures.

3. GAMING AND OFF-TASK BEHAVIOR

The underlying motivations of students is another important area of educational data mining research. Specifically, relevant to the present study is how students' motivations for completing a learning task may influence how they interact with the system, which in turn can influence the extent that model features may effectively track (and predict) their performance. Student motivations can be particularly influential in educational software because students may be able to choose whether to attempt to answer a question, how long to spend trying to answer a question, and when to solicit hints from the learning system. Students that do not behave as intended (e.g., rushing through tasks or seeking hints instead of attempting to answer questions) are frequently referred to as "gamers" (Baker, Corbett and Koedinger, 2004). There is ample evidence that student study preferences are at odds with what is suggested by learning science research. For instance, most students' metacognitive beliefs about how to best learn are miscalibrated (Kornell and Bjork, 2007; McCabe, 2011), and they typically do not recognize that attempting to answer a question can be an efficient way to improve memory (e.g., Roediger

and Butler, 2011). Some students may simply not want to engage fully with the learning system but still technically complete tasks as required by instructors or learning systems. Unfortunately, data from off-task students is often superficially similar to more diligent students. For example, diligent students may incorrectly answer a question and subsequently attend to feedback. In contrast, gamer students may not be attempting to recall answers at all; they may be skipping through to view the answers instead (Aleven and Koedinger, 2001; Wood and Wood, 1999). Gamers may even ignore corrective feedback, seeking instead to complete the task as quickly as possible. Baker et al. (2004) developed a model to predict these gamer students using data from an intelligent tutoring system (ITS). Most relevant to the present research, they found that speed of actions taken after errors could accurately classify gamers. Subsequent research has demonstrated that this approach can generalize to students learning other materials on other tutoring platforms (Baker et al., 2008). In short, trial duration appears to be able to identify gamer students whose learning is more poorly predicted by learner models, which may be because counts of prior practices for gamers are not tracking the same underlying processes.

4. Gaming Behavior and Predictiveness of Learner Models

If a student engaging in gaming chooses an answer randomly in order to skip to the corrective feedback, they are depriving themselves of known benefits of self-testing (Rowland, 2014), benefits that exist even if the student could not have known the answer (Grimaldi and Karpicke, 2014). While missing an opportunity to practice answering a question may reduce the efficacy of that particular study event, the consequences may extend to subsequent trials in an adaptive system. This is because the assumptions underlying a learner model may be violated in various ways. For instance, a count of prior failures may overestimate later performance for students who are ignoring corrective feedback and rushing to the end of a session. This issue would directly reduce the efficacy of adaptive systems that determine optimal practice schedules based on the relative benefit of possibly successful practice versus the possible cost of a failure (that would likely entail a more time-consuming and effortful restudy trial). Such maladaptive behavior may be partially captured by trial duration and has also been previously used to improve learner models of mathematics (Rihák, 2015).

In the present study, trial duration was used to improve the predictiveness of a logistic regression learner model. Distinct subgroups that varied according to potential gaming behavior were identified to a) allow evaluation of whether the subgroups of students were fit differently by typically effective models and then b) inspire new and adjusted predictor features that incorporated trial duration and improved overall prediction across the heterogeneous subgroups. Models of learning are constrained by the validity of the predictor features and their correspondence to the outcome measures (correctness in this case). Incorporating trial duration into the computation of predictor features was predicted to improve this correspondence.

5. The Data

The data used in the present analyses were extracted from an online learning platform for college students. In this system, students studied chapters of a textbook, and at the end of chapter sections, they were quizzed with questions pertaining to previously studied textbook content. When a student was presented with a question, they were asked first to give a confidence rating (coded as low or high) once they felt prepared to provide an answer. The student would then select their chosen answer among multiple alternatives. Immediately afterward, the student

would be provided with corrective feedback that included the correct answer as well as an explanation. Students could proceed at their own pace at every stage of each quiz question. Each of the datasets we analyzed contained 1035-1210 student participants, and each dataset concerned a different topic (Finance, Management, and Nutrition). Participation varied substantially; the number of trials (questions answered) per student ranged from as few as 4 trials to over 1000. Time intervals between quiz sessions also varied, some students had no intersession intervals (only using the system once) and others returned multiple times with over a month between some sessions. Of the available data for each quiz question, this paper used a deidentified student ID, a KC ID code, a timestamp, the correctness of the response (dichotomous), and trial duration. Other participant data were ignored for this analysis.

6. CLUSTERING

Participants were clustered according to their overall median trial duration time and overall proportion correct. Median trial duration was included because of evidence that trial durations correlate with gaming behavior (Baker, Corbett and Koedinger, 2004). However, faster trial durations can also indicate better mastery of learning material or of apparent gaming that benefits learning (Shih et al., 2008). These latter two points are why proportion correct was also included as a second input feature. The goal was to use these two features as dimensions of the clustering space to potentially find sub-groups of students that traded accuracy for speed. At one extreme, some participants may rush through the task simply to view answers and have low median trial duration and proportion correct. Others may be more diligent and have long trial durations and high proportion correct. Finally, other subgroups of participants may fall somewhere in between these two extremes. Our question was how these subgroups might be predicted differently by our model. There are many researcher degrees of freedom when it comes to choosing clustering features; we chose to limit our clustering analysis to the essential dimensions of interest - speed that may indicate gaming that is contextualized by accuracy[1]. We also hoped that fewer input features could make any distinct clusters more interpretable and large enough to be fit by our learner model.

Before clustering, the data were preprocessed. In some cases, the system did not detect that a student had logged out and would report outlier trial durations (e.g., a day-long trial) or a trial duration was not successfully recorded. In those instances, the trial duration was imputed to the median of the other trial durations from that student. This was done for approximately 1% of trials. Finally, only trials for which the trial duration was below the 95th percentile were used to calculate median trial durations and proportion correct. This resulted in the exclusion of another 4% of trials. Participants were clustered using a Gaussian mixture modeling-based method (Fraley et al., 2012). With this approach, the data were each assumed to be independently drawn from some unknown population density. The fitting task was to estimate the number of population groups g that are most likely to have together generated the entire dataset submitted for clustering. The modeling algorithm attempted to find the optimum number of g components that maximized a penalized likelihood. The ellipsoids that represented these g components could vary in their shape, orientation, and density. Means and variances of the components were fit via maximum likelihood. This flexible methodology gave more freedom

[1] We also explored clustering solutions with trial duration separated into durations of successes and failures. Highly similar clusters were obtained, as well as a similar pattern of results with subsequent model fitting.

for the patterns in the data to determine clusters, with fewer constraints imposed by assumptions of the clustering technique.

The results of the mixture modeling indicated that the datasets were best described with four or five components (or clusters) depending on the dataset. To evaluate the robustness of these solutions, we performed 10-fold cross-validated discriminant analysis on each dataset separately. The labels assigned by the original clustering solution were treated as the "known" labels to test against the cross-validation predictions. There was low classification error, ranging from 1.8 to 2.2% across datasets. The clusters that emerged from our analysis (Figures 1, 2 and 3, see Table 1 for descriptive statistics) seemed to loosely follow continuums from slower and more accurate (e.g., blue squares) to much faster and less accurate (e.g., red circles). However, clusters 1 and 2 in the nutrition dataset were more similar in correctness to cluster 3 than cluster 4, differing mainly in terms of median trial duration. The same clustering methods revealed similar[2] patterns in the other two datasets.

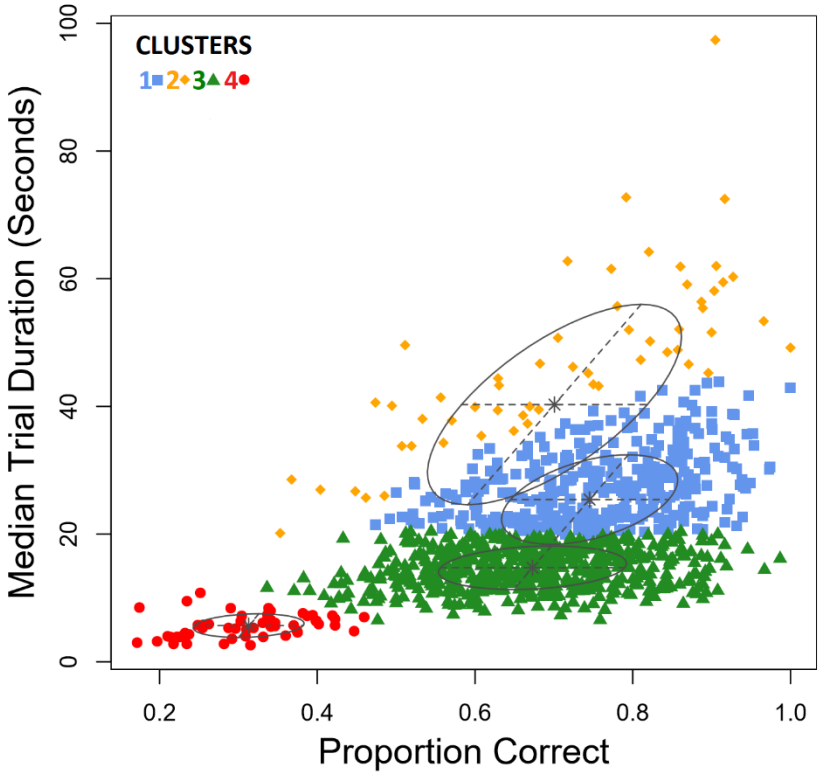

Figure 1: A plot of the nutrition dataset clusters.

[2] Although the management dataset had a five-cluster solution, the clusters still tracked the speed/accuracy tradeoff indicated in the two datasets. Subsequent analyses described later revealed a correlation between PFA model coefficient values and cluster trial duration that was also consistent across datasets.

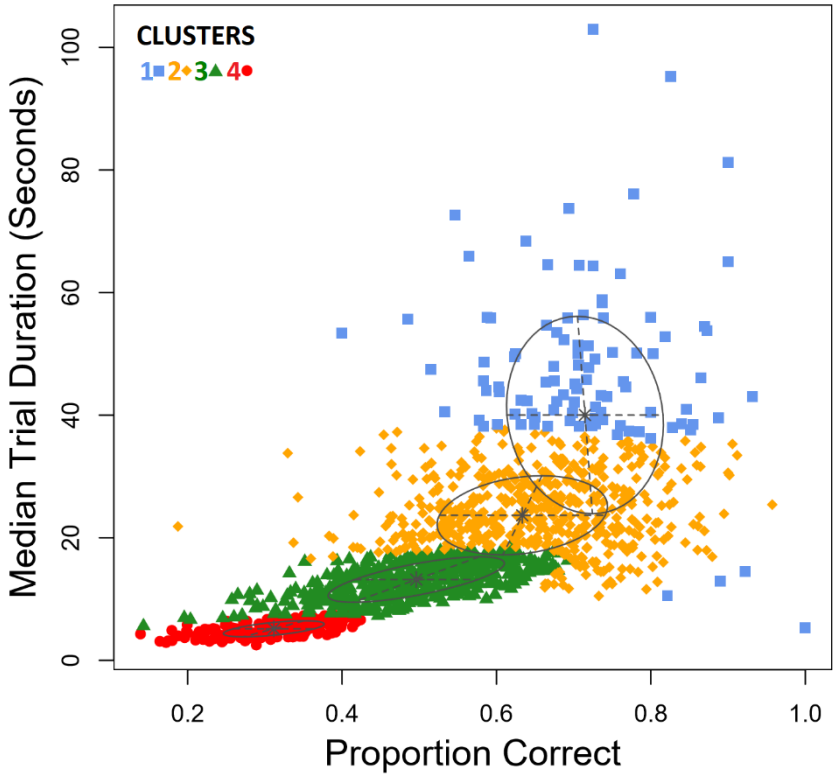

Figure 2: A plot of the finance dataset clusters.

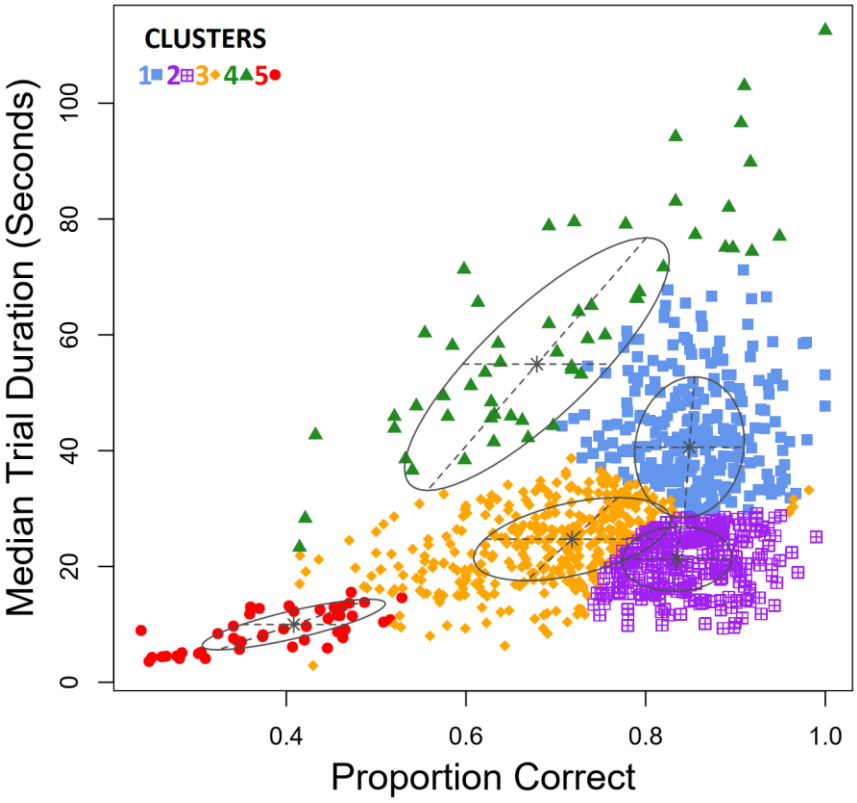

Figure 3: A plot of the management dataset clusters.

Table 1: Descriptive statistics of proportion correct and trial duration of clusters for each dataset.

Nutrition	Cluster 1	Cluster 2	Cluster 3	Cluster 4	
Mean Accuracy (SD)	.76(.11)	.70(.17)	.67(.12)	.31(.07)	
Median Duration (MAD)	26.80(6.07)	44.80(10.97)	14.55(3.63)	5.70(2.08)	
Sample Size	362	57	573	55	
Finance	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	
Mean Accuracy (SD)	.72(.10)	.65(.11)	.49(.10)	.31(.06)	
Median Duration (MAD)	44.65(9.04)	24.97(6.95)	13.20(3.56)	5.10(1.48)	
Sample Size	101	539	413	157	
Management	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Mean Accuracy (SD)	.85(.06)	.84(.05)	.72(.09)	.68(.14)	41(.08)
Median Duration (MAD)	41.65(9.45)	20.88(5.82)	24.80(7.41)	58.60(19.12)	8.82(4.52)
Sample Size	272	55	368	294	46

7. LEARNER MODELS

Our starting model was a modified PFA model (see Equation 1), hereafter referred to as LPFA. There were two predictors, log-transformations of prior success (logsuc) and failure trials (logfail) estimated for each KC level (1 was added to all values so that they were defined at 0 attempts). We chose this model because it allowed us to potentially detect different effects of successes versus failures. This model also provides reasonable fit while remaining simple enough to be fit to smaller clusters without overfitting. Due to the small number of trials from some students, we chose to treat student intercepts as random effects (Bates et al., 2014). Students sometimes only practiced a given KC a few times and typically did not practice all KCs. This variability prompted us to treat KC intercepts and slopes (of logsuc and logfail) as random effects. Our approach was modeled after that of Goldyardt and Goldin (2015) and DeBoeck et al. (2011). We also tried an alternative random effects structure in which random intercepts and slopes were correlated but did not find substantive differences in the final results. Logsuc, logfail, and modified versions of those predictors were treated as fixed effects. The model was written in R as follows: `glmer(y ~ logsuc + logfail + (1|student) + (1|kc) + (0+logsuc|kc) + (0+logfail|kc), family=binomial("logit"))`. Reported R^2 in Tables 2 and 3 pertain only to the fixed effects included in the models.

$$logit(p_{ijt}) = \underbrace{\beta_1 \log_e S_{ij} + \beta_2 \log_e F_{ij}}_{\text{Fixed}} + \underbrace{\theta_i + \theta_j + \mu_{sj} S_{ij} + \mu_{fj} F_{ij}}_{\text{Random}}$$

Equation 1: LPFA predicts future performance using the logs of prior successes *S* and failure counts *F* for each student *i* and KC *j*. θ_i and θ_j represent random intercepts for student ability

and KC difficulty, respectively. $\mu_{sj}S_{ij}$ and $\mu_{fj}F_{ij}$ represent random slopes for logsuc and logfail across KCs.

Table 2: Coefficients of the fixed effects for the LPFA model fit to each of the three datasets, as well as to each cluster within each dataset separately. Values in parentheses are standard errors. RMSE is average of subject-level root mean squared error. The R^2 column denotes a pseudo R^2 estimate of fixed effects (Johnson, 2014; Nakagawa and Schielzeth, 2013). See Nakagawa, Johnson, and Schielzeth (2017) for examples and explanations of the logic of the method. * = $p < .01$. ** = $p < .001$.

Nutrition	Logsuc	Logfail	AIC	R^2	RMSE
Cluster 1(n=362)	.774(.236)**	4.960(.184)**	12259	.455	.3063
Cluster 2(n=57)	.695(.396)	3.544(.213)**	3411	.413	.3355
Cluster 3(n=573)	.889(.167)**	3.750(.140)**	36772	.366	.3487
Cluster 4(n=55)	-.329(.203)	1.393(.069)**	12982	.138	.3933
Fit to All Data	.569(.130)**	2.901(.109)**	68846	.313	.3632

Finance	Logsuc	Logfail	AIC	R^2	RMSE
Cluster 1(n=101)	2.420(.778)**	4.280(.249)**	2612	.384	.3350
Cluster 2(n=539)	1.170(.168)**	3.554(.078)**	32587	.354	.3608
Cluster 3(n=413)	1.091(.142)**	2.272(.062)**	60467	.257	.3904
Cluster 4(n=157)	-.663(.085)**	1.042(.039)**	53403	.096	.3804
Fit to All Data	.090(.085)	1.800(.048)**	156207	.273	.4000

Management	Logsuc	Logfail	AIC	R^2	RMSE
Cluster 1(n=272)	1.219(.188)**	4.177(.157)**	9643	.314	.3153
Cluster 2(n=55)	.744(.259)**	3.187(.179)**	3516	.293	.3648
Cluster 3(n=368)	1.303(.113)**	2.381(.069)**	36712	.251	.3954
Cluster 4(n=294)	.886(.098)**	2.328(.108)**	21712	.170	.3671
Cluster 5(n=46)	.184(.216)	1.530(.071)**	11091	.160	.4092
Fit to All Data	1.088(.090)**	2.177(.056)**	83571	.204	.3834

Data were fit to the LPFA model via maximum likelihood. If a student only had a single trial regarding a particular KC, that trial was excluded to facilitate fitting the model. This resulted in excluding 42.9% of the trials for the model fitting portion of the analysis (across all three datasets). Table 2 displays parameter coefficients from fitting each cluster in the nutrition dataset with the LPFA model separately, as well as coefficients for when all students in that dataset were fit with one LPFA model. The separate fits for each cluster allowed us to evaluate how the importance of logsuc and logfail may have varied. The model fits and coefficients did indeed vary considerably across clusters (see Table 2). Clusters that tended to have longer trial durations and higher accuracy were better fit (e.g., cluster 1 in each dataset). Very fast and

inaccurate participants were poorly fit by the model (i.e., the red clusters in each dataset). Inspection of the coefficients for the overall fit and clusters revealed some interesting patterns. The log of prior failures predictor (logfail) coefficients varied widely across clusters, which was especially notable due to how much stronger a predictor it was than logsuc. The coefficients for the log of prior successes predictor (logsuc) also varied but were more similar across clusters. In general, faster trial durations seemed to correlate with smaller coefficients for logsuc and logfail. One consequence of this variability is that overall coefficient estimates for logsuc and logfail were suboptimal for large sub-groups of students when the model was fit to the entire dataset.

How can this variability among students be accounted for to improve overall fit? Some students who spent considerably less time on trials (red clusters) did not appear to benefit from failures to the same extent as those who spent more time per trial. The material students were studying could be fairly complex, and reviewing corrective feedback may have been time-consuming. It has been previously demonstrated that failed tests can improve learning, but they do require attention to meaningful feedback (Kornell, Hays and Bjork, 2009). In the present case, speeding through questions may have attenuated potential benefits from failed quiz questions. In contrast, speeding through and consistently answering *correctly* may indicate skill mastery, and so it is less reliable predictor. This apparent relationship between trial duration and logfail predictiveness motivated modifications to that predictor. We first demonstrate how separate logfail predictors for each cluster can improve fits. Although logsuc was also a significant predictor, we demonstrate in the following analyses that separating the logsuc parameter by cluster improved model fit much less than with logfail. Subsequently, we show how weighing counts of prior failures by the students' cumulative median trial duration accounted for a substantial proportion of the benefit conferred by separate logfail predictors for each cluster.

Each dataset was fit to a modified version of the model that included a single logsuc predictor as before, but separate logfail predictors for each cluster of students (LFperCluster in Table 3). This is in contrast to the analysis described in Table 2, in which each cluster was fit separately to the original LPFA model. Thus, LFperCluster had five or six fixed effect parameters instead of two (an extra logfail predictor for each cluster in the dataset, two datasets had four clusters, one had five). Partitioning logfail by cluster improved model fit for all three datasets (see Table 3). Overall, there was a clear benefit of allowing the influence of logfail to vary by cluster. The benefit of separating logfail by cluster did not appear to be simply due to adding extra predictors; randomly assigning cluster membership almost entirely removed the benefit. In other words, the benefit of additional logfail predictors per cluster depended on accurate cluster membership, not simply having additional logfail predictors. If cluster membership was randomly assigned, model fit was not substantially improved beyond the modified PFA model (and sometimes had higher AIC due to extra predictors). One hundred simulation runs indicated random cluster assignment was significantly worse than with correct cluster membership, $p < .000001$.

In a separate analysis, we attempted to fit separate logsuc predictors to each cluster instead of separate logfail predictors (LSperCluster in Table 2). LSperCluster only reduced AIC 12% as much as LFperCluster. In the present datasets, trial duration appears to be more relevant to modulating the influence of counts of prior failures than of prior successes. Trial duration may be less relevant for logsuc predictiveness because both fast and slow successful trials can indicate prior learning and successful retrieval. In contrast, fast failed trials may be insufficient to review the corrective feedback at all, with longer trials being much more likely to cause more learning. In sum, we think the result is due to how separate logfail predictors for each cluster

allowed more diligent students (who were typically clustered together) to have larger logfail coefficients that may have better reflected their increased attention to corrective feedback.

Table 3: Descriptive statistics of five model fits to the three datasets. ΔAIC denotes the relative reduction in AIC from the LPFA model (Equation 1). LFperCluster is a model with separate logfail predictors for each cluster within a given dataset. LFxLD is a model with a single modified logfail predictor - logfail times prior median trial duration on previous failed attempts (measured in seconds). LSperCluster and LSxLD are models with the same principles applied to counts of successes. The R^2 column denotes a pseudo R^2 estimate of fixed effects (Nakagawa, Johnson and Schielzeth, 2017).

Dataset	Model	AIC	ΔAIC	Specificity	Sensitivity	R^2	RMSE
Nutrition	LPFA	68846		.826	.709	.313	.3632
Nutrition	LFperCluster	65607	-3239	.845	.710	.365	.3381
Nutrition	LSperCluster	68467	-409	.828	.714	.319	.3606
Nutrition	LFxLD	66558	-2288	.832	.713	.335	.3453
Nutrition	LSxLD	68868	+22	.827	.708	.312	.3634
Finance	LPFA	156207		.909	.394	.212	.4000
Finance	LFperCluster	150535	-5672	.904	.444	.239	.3747
Finance	LSperCluster	155358	-849	.909	.400	.198	.3986
Finance	LFxLD	151226	-4981	.901	.428	.219	.3832
Finance	LSxLD	156277	+70	.909	.390	.210	.4003
Management	LPFA	83571		.713	.829	.204	.3834
Management	LFperCluster	82618	-953	.726	.828	.227	.3656
Management	LSperCluster	83617	+46	.713	.830	.203	.3826
Management	LFxLD	82455	-1116	.721	.830	.219	.3700
Management	LSxLD	84604	+33	.713	.829	.203	.3836

One important limitation of the above model is that the improvement was based on a clustering solution derived from the entire dataset. This post hoc method would, therefore, be difficult to utilize in a learning system. The clusters depicted in Figure 1-3 seemed to be tracking the diminishing benefits of logfail as a function of trial duration. Some participants were clearly spending far less time reviewing the material on a given trial, but the counts are treated similarly by the original model. In the following analysis, we hypothesized that the same benefit of separate logfail predictors per cluster could be captured by weighting the count of prior failures by the median trial duration of previous attempts for a given student.

7.1. A Modification to the LogFail Predictor

For each student, for each trial *t*, a median trial duration was calculated from the previous trials 1:*t*-1 for which the student was incorrect. For the first trial for a particular student, the value was set to zero. The logarithm of this value plus one (to prevent undefined values) was multiplied by the original logfail predictor to generate a new composite measure, referred to hereafter as

LFxLD (LogFail times LogDuration). The logarithm of the median value was used because of evidence that study duration provides steeply diminishing benefits to learning (Metcalfe & Kornell, 2003; Pavlik, 2007). The logic underlying the composite measure is that the predictive utility of the count of prior successes or failures is likely influenced by how long the student spent on those attempts. If a student failed many times previously but had very fast trial durations on those attempts, they are unlikely to have successfully encoded the feedback information. Conversely, a student with longer trial durations is more likely to have attended to the feedback. We expected that the effect of trial duration on prior attempts to be substantially more important for unsuccessful trials. We chose to calculate median trial duration at the student level because we believed the tendency to rush through trials may be a more trait-based behavior, and not especially sensitive to specific KCs. Furthermore, the datasets tended to have few trials per KC for a given student which resulted in lower reliability if medians were computed per KC per student. Regardless, calculating median trial duration at KC level for each student provided very similar results (albeit slightly worse) to those reported below. Replacing logfail with LFxLD improved model fit for all three datasets (see Table 3). Importantly, the LFxLD model provided approximately 85% of the reduction in AIC conferred by including separate logfail predictors for each cluster. Although separate logsuc predictors were not very helpful in the previous analysis, for completeness we created an analogous predictor from successful trials (LSxLD). LSxLD did not reliably improve fits in the present datasets (see Table 3). The clear contrast between LFxLD (clearly better than LPFA) and LSxLD (slightly worse than LPFA) highlights how the meaningfulness of trial duration can depend on student performance on that trial. A straightforward explanation for these results is that LFxLD is accounting for how time-on-task variability influences learning benefits. This variability may be especially relevant when the study materials are complex, and the feedback is an explanation of a concept (as opposed to simpler feedback consisting of the correct vocabulary word) as it often was for our questions.

How uniform was the benefit of this new model across different students? Separate fits per cluster indicated that the LPFA model (Equation 1) was typically better fit to clusters made up of slower and more accurate students. Faster and less accurate students were not well predicted. The speed/accuracy tradeoff appears to be a continuum, but some of the fastest clusters were spending too little time per trial to really gain to any degree. Thus, one might expect that LFxLD not improve fit very much for very fast students – longer trial durations may only predict more learning from failures once beyond a minimum threshold needed to read feedback. Figure 4 depicts the relative benefit of the LFxLD model over LPFA, as a function of student median trial duration. Students with longer median trial durations tended to have larger reductions in RMSE from LFxLD model versus LPFA in all three datasets, ts > 17.56, ps < .00001. Correlations averaged -.49(.016).

If trial duration modulates the predictiveness of logfail, then the slope of that predictor should vary across students according to their respective median trial duration. Such a finding would provide converging evidence for the influence of trial duration on predictiveness of prior counts of performance. We tested this hypothesis by refitting the LPFA model and included random slopes for logsuc and logfail predictors nested within student. Our hypothesis was that the student-level slopes would be positively correlated with their respective median trial durations. The per-student slopes for logsuc and logfail were both significantly correlated with median trial duration in all three datasets, rs > .3, ps < .0001. It might be surprising at this point that logsuc slopes were correlated with median trial duration. However, in these datasets, the actual magnitude of the slopes for logsuc was typically 4 times smaller than that for logfail. Thus, the actual impact on model performance was small despite the significant correlation. In short,

analysis of the individual slopes provided converging evidence that failure trial duration modulated the predictiveness of counts of prior performance.

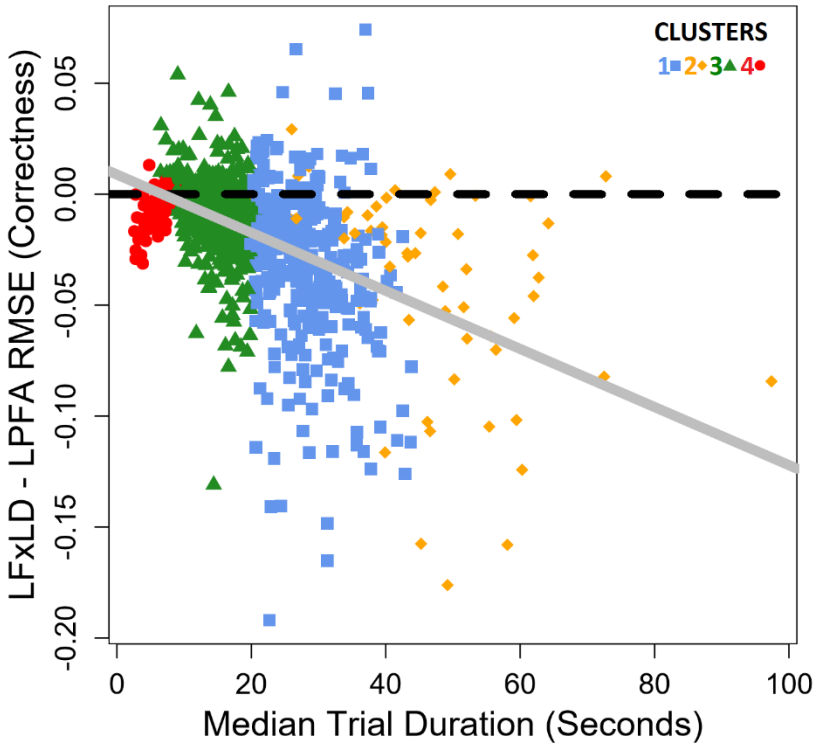

Figure 4: LFxLD model correctness RMSE for each student in nutrition dataset subtracted from their respective RMSE from LPFA model fit, as a function of their median trial duration in seconds. The gray line indicates a linear regression slope. The dashed line marks where the two models are equivalent.

8. Discussion

In the present study, the predictive utility of counts of prior failures was found to vary substantially across students. Clustering by trial duration and accuracy revealed sub-groups of slower students that were well predicted by the LPFA model, with prior successes and especially failures predicting performance. Other faster students were more poorly fit by the LPFA model and had notably smaller coefficients for predictors based on counts of prior failures (logfail). Partitioning the logfail predictor by cluster substantially improved model fit in all three datasets. However, the clusters seemed to capture a more general trend of a tradeoff between trial duration and learning from errors. Replacing the original logfail predictor with one that was multiplied by the logarithm of prior median trial duration (LFxLD) improved the model fit almost as much as having separate logfail predictors per cluster. The regression slope in Figure 4 indicates that these altered predictors were especially beneficial for students with longer median trial durations (who were typically higher-performing overall). Longer median trial durations were correlated with larger relative benefits of LFxLD over LPFA. Many learner models are biased towards predicting success (Gong, Beck and Heffernan, 2011), and have significant difficulty with poor performers. It may be that the faster but poorer performing students are in essence completing a different task; they may be trying to finish as many trials as possible, rather than learn from

the materials. This distinction implies that improving fits to these students may require different features (or interventions to change their behavior; Baker et al., 2006).

Student motivations and understanding of how to learn are closely related to the efficacy of self-regulated study. Learning can be difficult to change with external motivators (Kang and Pashler, 2014) or with verbal instruction about best study practices (Yan et al., 2016). The model fit improvement conferred by LFxLD may be related to student motivations and preconceptions about how to learn. The apparent clusters seemed to be mostly explained by a continuous relationship between logfail predictiveness and trial duration, although the overall model fit (and relative benefit of new models) was generally lower for poorer performing and faster students. Fast and inaccurate students may have had particularly misaligned motivations or preconceptions about studying. They were basically not doing the task in a way that standard learner models could predict. This limitation indicates that learner models either need to account for their unexpected behavior or that the behavior of those students needs to be changed to suit the model. The latter may be easier; a separate classifier could be employed to detect these students and modify their study behavior with targeted interventions. Early classification would also allow data from these off-task students to be excluded or down-weighted when estimating parameters for other on-task students. Some learning systems may benefit from accounting for trial duration more than others. Weighing counts of prior trials could be especially relevant in learning environments when students can choose whether or for how long that they attend to feedback (as in the present datasets). Alternative implementations of feedback may indirectly reduce the influence of time-on-task. For instance, if students are required to provide overt responses in response to the feedback (e.g., clicking on the correct answer) they may be less likely to rush through the feedback portion of the task. Similarly, imposing minimum trial durations may also reduce the influence of time-on-task, but of course this would not guarantee student attention and would reduce efficiency for students who learn more quickly.

8.1. Limitations & Future Directions

In the present study, we analyzed datasets in which feedback consisted of text explaining the target concept. In contrast, many learning systems provide hints that may be specific to the particular type of error that the student committed. Students may also solicit hints themselves. For a given problem, multiple hints may be provided in a sequence concerning the same problem. These structural differences between the learning systems raises the question of how trial duration could be incorporated across systems. Trial durations are likely to still be informative, but their dependence within a sequence of hints may be a relevant factor to consider when using them to weigh counts of prior attempts. Additionally, in a sequence of hints concerning one problem, the hints themselves may not be equally informative for the student or predictive for a learner model. For instance, the first hint may offer a small amount of additional information to the student, with subsequent cues offering more information or even an entire worked solution (Koedinger and Aleven, 2007). The predictiveness of trial duration on an initial hint may be contingent on time spent on subsequent hints. A motivated (non-gamer) student may rush through initial hints in order to view a final worked example, where they may spend more time (Shih, Koedinger and Scheines, 2008). This pattern of shorter followed by longer durations within one sequence of hints may predict learning when performed by a diligent student, and thus a cumulative median (as was used in the present study) could be suboptimal relative to utilizing a model of trial duration (Shih, Koedinger and Scheines, 2008).

The above example illustrates how student strategies could influence the predictiveness of across-trial dependence and trial duration. These patterns of behavior may result in more distinct

clusters of students, but the clustering strategy in the present study would not be helpful – the clustering solution was based on the entirety of the data, and thus the cluster labels wouldn't be available *during* learning. However, if versions of new features similar to LFxLD would be even more effective given accurate cluster labeling, it may be worth inferring cluster membership given incomplete data during learning. For instance, in the present datasets, cluster membership (given the full dataset) could be predicted with approximately 75% accuracy given the mean accuracy and median trial duration from the initial 30% of trials from each student. Early cluster prediction could also be improved by tailoring the selection of early practice items to optimize estimation of the cluster input features (e.g., an appropriate mixture of harder and easier trials could provide more accurate estimates of trial duration). Given a high confidence cluster prediction, the relation between trial duration and counts of practice (including hints) may be more effectively customized to students with various learning strategies.

Mean accuracy and median trial duration were chosen as input features for clustering to facilitate grouping students who varied in how they traded speed for accuracy. Two features were used to increase the likelihood of obtaining interpretable results. However, in some contexts, other features could be valuable inputs for clustering analysis. For example, a students' tendency to ask for hints, and how early in practice they do so, may be useful features that could group students according to their comfort with attempting more difficult problems. If the text of specific problems is available (not in the present datasets) trial duration could also be measured relative to what would be expected given an average attentive reader. Adults can read approximately 300 words per minute (Rayner et al., 2016), and can type approximately 52 words per minute (Dhakal et al., 2018). Those measures together could help weight a previous attempt based on likely completion times (e.g., a question containing 50 words is extremely unlikely to be read, processed, and correctly answered in 3 seconds).

Finally, the nonlinear relationship between trial duration and correctness may vary in ways not explored in the present study. We characterized the interaction with a log-transform of trial duration times the practice count, but another transformation may be more optimal. The particular transformation may also depend on what is being learned (e.g., learning facts versus procedural skills), since the meaning of fast or slow practice may differ based on the cognitive processes needed for the task.

8.2. Conclusion

Overall, the present study illustrates how incorrect trial duration can improve the predictive power of counts of prior incorrect performance. Accounting for this relationship could significantly improve model fit depending on the context of the learning system. Such improvements may therefore improve pedagogical selection by providing a more accurate input to infer when items have been learned. Because the benefit to fit was greatest for students with longer durations, who were given more credit for their additional time spent, the LFxLD feature may improve the pedagogy of adaptive learning systems by reducing excessively repetitive item selection for these careful students.

Acknowledgments

The present research was funded by the Schmidt Futures foundation and NSF grant #1443068. We thank McGraw Hill Education for providing data for analysis. The views expressed by the authors do not necessarily represent those of the National Science Foundation, the Schmidt Futures foundation, or McGraw Hill Education.

Editorial Statement

Luke Eglington had no involvement with the journal's handling of this article in order to avoid a conflict with his Special Track Editor role. The entire review process was managed by Associate Editor Ryan Baker.

References

Aleven, V. and Koedinger, K. R. 2001. Investigations into help seeking and learning with a cognitive tutor. In *Proceedings of the Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments*, 47-58.

Ayers, E. and Junker, B. 2006. Do skills combine additively to predict task difficulty in eighth grade mathematics? In *Educational Data Mining: Papers from the AAAI Workshop*, J. Beck, E. Aimeur and T. Barnes Eds. AAAI Press, Menlo Park, CA, 14-20.

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*, 185-224.

Baker, R. S., Corbett, A. T. and Koedinger, K. R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Proceedings of the Intelligent Tutoring Systems*, Berlin, Heidelberg, J. C. Lester, R. M. Vicari and F. Paraguaçu Eds. Springer Berlin Heidelberg, 531-540.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., Naim, M., Raspat, J., Baker, D. J. and Beck, J. E. 2006. Adapting to When Students Game an Intelligent Tutoring System. In *Proceedings of the Intelligent Tutoring Systems*, Berlin, Heidelberg, M. Ikeda, K. D. Ashley and T.-W. Chan Eds. Springer Berlin Heidelberg, 392-401.

Baker, R. S. J. D., Goldstein, A. B. and Heffernan, N. T. 2011. Detecting learning moment-by-moment. *International Journal of Artificial Intelligence and Education 21*, 5-25.

Bates, D., Mächler, M., Bolker, B. and Walker, S. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1506.05908*.

Butler, A. C., Fazio, L. K. and Marsh, E. J. 2011. The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review 18*, 1238-1244.

Carrier, M. and Pashler, H. 1992. The influence of retrieval on retention. *Memory & Cognition 20*, 633-642.

Cen, H., Koedinger, K. R. and Junker, B. 2006. Learning Factors Analysis - A general method for cognitive model evaluation and improvement. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* Springer Berlin / Heidelberg, 164-175.

Chi, M., Koedinger, K. R., Gordon, G., Jordan, P. and VanLehn, K. 2011. Instructional Factors Analysis: A cognitive model for multiple instructional interventions. In *Proceedings of the 4th International Conference on Educational Data Mining*, Eindhoven, The Netherlands, M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero and J. Stamper Eds., 61-70.

CORBETT, A. T. AND ANDERSON, J. R. 1992. Student modeling and mastery learning in a computer-based programming tutor. In *Intelligent Tutoring Systems: Second International Conference on Intelligent Tutoring Systems*, C. Frasson, G. Gauthier and G. Mccalla Eds. Springer-Verlag, New York, 413-420.

DE BOECK, P., BAKKER, M., ZWITSER, R., NIVARD, M., HOFMAN, A., TUERLINCKX, F. AND PARTCHEV, I. 2011. The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software 39*, 1-28.

DESMARAIS, M. C. AND LEMIEUX, F. 2013. Clustering and Visualizing Study State Sequences. In *Proceedings of the 6th International Conference of Educational Data Mining*, S. K. D'mello, R. A. Calvo and A. Olney Eds., Memphis. TN, 224-227.

DHAKAL, V., FEIT, A. M., KRISTENSSON, P. O. AND OULASVIRTA, A. 2018. Observations on Typing from 136 Million Keystrokes. In *Proceedings of the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada ACM, 3174220, 1-12.

FISCHER, G. H. 1973. The linear logistic test model as an instrument in educational research. *Acta Psychologica 37*, 359-374.

FRALEY, C., RAFTERY, A. E., MURPHY, T. B. AND SCRUCCA, L. 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.

GALYARDT, A. AND GOLDIN, I. 2015. Move your lamp post: Recent data reflects learner knowledge better than older data. *Journal of Educational Data Mining 7*, 83-108.

GONG, Y., BECK, J. E. AND HEFFERNAN, N. T. 2011. How to Construct More Accurate Student Models: Comparing and Optimizing Knowledge Tracing and Performance Factor Analysis. *International Journal of Artificial Intelligence in Education 21*, 27-46.

GRIMALDI, P. J. AND KARPICKE, J. D. 2014. Guided retrieval practice of educational materials using automated scoring. *Journal of Educational Psychology 106*, 58-68.

IZAWA, C. 1970. Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology 83*, 340-344.

JOHNSON, P. C. D. 2014. Extension of Nakagawa & Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution 5*, 944-946.

KANG, S. H. K. AND PASHLER, H. 2014. Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition 3*, 183-188.

KELLY, J. W., CARPENTER, S. K. AND SJOLUND, L. A. 2015. Retrieval enhances route knowledge acquisition, but only when movement errors are prevented. *Journal of Experimental Psychology: Learning, Memory, and Cognition 41*, 1540-1547.

KOEDINGER, K. R. AND ALEVEN, V. 2007. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review 19*, 239-264.

KORNELL, N. AND BJORK, R. A. 2007. The promise and perils of self-regulated study. *Psychonomic Bulletin & Review 14*, 219-224.

KORNELL, N., HAYS, M. J. AND BJORK, R. A. 2009. Unsuccessful retrieval attempts enhance

subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition 35*, 989-998.

McCabe, J. 2011. Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition 39*, 462-476.

Morris, C. D., Bransford, J. D. and Franks, J. J. 1977. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior 16*, 519-533.

Nakagawa, S., Johnson, P. C. D. and Schielzeth, H. 2017. The coefficient of determination R(2) and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface 14*(134): 20170213.

Nakagawa, S. and Schielzeth, H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4*, 133-142.

Newell, A. and Rosenbloom, P. S. 1981. Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition 1*, 1-55.

Pardos, Z. A., Trivedi, S., Heffernan, N. T. and Sárközy, G. N. 2012. Clustered Knowledge Tracing. In *Proceedings of the Intelligent Tutoring Systems*, Berlin, Heidelberg, S. A. Cerri, W. J. Clancey, G. Papadourakis and K. Panourgia Eds. Springer Berlin Heidelberg, 405-410.

Pavlik Jr., P. I. 2007. Understanding and applying the dynamics of test practice and study practice. *Instructional Science 35*, 407–441.

Pavlik Jr., P. I. and Anderson, J. R. 2008. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied 14*, 101–117.

Pavlik Jr., P. I., Cen, H. and Koedinger, K. R. 2009. Performance factors analysis -- A new alternative to knowledge tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, V. Dimitrova, R. Mizoguchi, B. D. Boulay and A. Graesser Eds., Brighton, England, 531–538.

Pavlik Jr., P. I., Yudelson, M. and Koedinger, K. R. 2011. Using contextual factors analysis to explain transfer of least common multiple skills. In *Artificial Intelligence in Education*, G. Biswas, S. Bull, J. Kay and A. Mitrovic Eds. Springer, Berlin, Germany, 256–263.

Piech, C., Spencer, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. and Sohl-Dickstein, J. 2015. Deep Knowledge Tracing. *arXiv preprint arXiv:1506.05908*.

Rasch, G. 1961. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* University of California Press Berkeley, CA, 321-333.

Rayner, K., Schotter, E. R., Masson, M. E. J., Potter, M. C. and Treiman, R. 2016. So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? *Psychological Science in the Public Interest 17*, 4-34.

Rihák, J. 2015. Use of Time Information in Models behind Adaptive System for Building Fluency in Mathematics. In *Proceedings of the 8th International Conference on Educational Data Mining*, O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, . . . M. Desmarais Eds., Madrid, Spain, 642-644.

Roediger, H. L. and Butler, A. C. 2011. The critical role of retrieval practice in long-term

retention. *Trends in Cognitive Sciences 15*, 20-27.

ROWLAND, C. A. 2014. The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin 140*, 1432-1463.

SHIH, B., KOEDINGER, K. R. AND SCHEINES, R. 2008. A response time model for bottom-out hints as worked examples. In *Proceedings of the 1st International Conference on Educational Data Mining*, R. S. Baker and J. E. Beck Eds., Montreal, Canada, 117–126.

WOOD, H. AND WOOD, D. 1999. Help seeking, learning and contingent tutoring. *Computers & Education 33*, 153-169.

YAN, V. X., BJORK, E. L. AND BJORK, R. A. 2016. On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General 145*, 918-933.