Using Prior Data to Inform Initial Performance Predictions of Individual Students

Michael G. Collins (michael.collins.74@ctr@us.af.mil)
Kevin A. Gluck (kevin.gluck@us.af.mil)
Matthew Walsh (mmw118@gmail.com)
Michael Krusmark (michael.krusmark@us.af.mil)

Air Force Research Laboratory, 2620 Q Street, Building 852
Wright-Patterson Air Force Base, OH 45433

Abstract
The predictive performance equation (PPE) is a mathematical model of learning and retention that uses regularities seen in human learning to predict future performance. Previous research (Collins, Gluck, Walsh Krusmark & Gunzelmann,, 2016) found that prior data could be used to inform PPE's free parameters when generating predictions of a group's aggregate performance, allowing for more accurate initial performance predictions. Here we investigate an extension of this methodology to predict performance of individuals, rather than aggregate samples. This paper documents the results of that investigation, which is on the critical path to the use of this cognitive technology in education and training.

Keywords: Mathematical model; Performance predictions; Skill learning; Parameter generalization; Educational data mining, Individual predictions

Introduction

It is typical in training and education for instructors to have little to no information about the people who are about to begin the curriculum. Rather, individuals must complete some portion of the curriculum before for their knowledge can be assessed. This assessment period can lead to an increase in the overall amount of time that training and education takes, and can lead to individuals practicing skills that have already mastered (Beck & Chang, 2007). Ideally, instructors could be able to estimate the future performance of both the incoming cohort of students as a whole in addition to the specific individuals based on the past performance of those who learned the same curriculum. This would allow instructors to better adjust a given curriculum to fit the needs of the cohort and of specific students.

In cognitive science, models of learning and retention have been developed to account for particular regularities in human learning such as the power law of learning (Newell & Rosenbloom, 1981) and power law of decay (Rubin & Wenzel, 1996), and the spacing effect (Bahrick, Bahrick, Bahrick, & Bahrick, 1993 Although many of these models were created based on basic laboratory phenomena, they can also be used to generate predictions of future human performance (Anderson & Schunn, 2000; Jastrzembski, Gluck, & Gunzelmann, 2006; Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009; Pavlik & Anderson, 2008; Raaijmakers, 2003). These models hold promise in training and education to increase mastery and /or decrease instruction time.

The Predictive Performance Equation

The model discussed in this paper is the Predictive Performance Equation (PPE; Walsh et al, submitted). PPE is a mathematical model of human learning and retention that can generate performance predictions on declarative (know-what) and procedural (know-how) tasks. Prior research has validated PPE across a variety of different laboratory tasks (Walsh et al., submitted) as well as complex human performance data from F-16 fighter pilot training research (Jastrzembski et al., 2006) and education and training data (Collins, Gluck, & Jastrzembski, 2015).

PPE represents the effects of three factors on knowledge acquisition and retention: recency of practice, frequency of practice, and the distribution of practice over time (i.e., spacing). The first factor, recency (T_n), captures the amount of elapsed time since training began. T_n is calculated as a weighted sum of the elapsed time since each of each previous training opportunities (t_i) (Equation 1). The weight (w_i) applied to the amount of time that has passed since a particular event decreases exponentially with time (Equation 2). Although in principle a free parameter, prior model exploration has found that the exponent, x, can be set to 0.6, which we do in the analyses presented here.

$$T = \sum_{i=1}^{n} w_i * t_i \quad (1)$$

$$w_i = -t_i^{-x} \sum_{j=1}^{n} \frac{1}{t_j^{-x}} \quad (2)$$

The second factor, frequency (N_n), represents the number of times that a particular knowledge or skill has been rehearsed. These two factors, elapsed time and frequency of practice, have a multiplicative effect on activation (M_n), which is the strength of a particular memory or skill (Equation 3). Amount of practice is scaled by the learning rate c, which is fixed to 0.1. As the amount of practice increases, activation rises at a decreasing rate, producing the power law of learning. The second term, comprised of T and d, captures the effects of the power law of decay. The decay rate, d, captures spacing effects (Equation 4).

$$M_n = N^c * T^{-d} \quad (3)$$

$$d_n = b + m * \left(\frac{1}{n-1} * \sum_{j=1}^{n-1} \frac{1}{\log\left(lag_j + e\right)}\right) \quad (4)$$

The precise effect of spacing on performance is determined by the summation term within the decay parameter. When lags between successive training opportunities (lag_j) are

short, the summation term in Eq. 4 approaches 1 and decay increases, leading to a greater amount of forgetting. When the lags between training opportunities are long, and the summation term approaches 0, the decay term decreases, leading to less forgetting over time. The decay rate equation includes a decay intercept parameter (b) and a decay slope parameter (m). The activation value from Eq. 3 is scaled to performance through a logistic function (Equation 5). The function contains two additional free parameters controlling its slope (s) and the intercept (τ).

$$P_n = \frac{1}{1+\exp\left(\frac{\tau - M_n}{s}\right)} \quad (5)$$

In summary, PPE has four free parameters (i.e., b, m, s, τ). These parameters can be calibrated based on existing performance data. Once a set of best fitting parameters have been found, PPE can use these parameters to predict *future* performance.

Motivation

Reliable and valid parameter estimates for PPE cannot be found with PPE when calibrating to fewer than three training opportunities. There are two reasons for this. First, when fewer than three data points are available, multiple combinations of free parameter values that can fit the available training data equally well. This makes it difficult to determine which set of parameter values should be used to generate out-of-sample predictions (Beck & Cheng 2007). Second, when calibrating PPE to so little data, PPE will likely fit to both the performance of the individual as well as to noise in the data (Geman, Bienenstock, & Doursat, 1992). This overfitting, in turn, will reduce the accuracy of out-of-sample predictions. The combination of these two factors are likely to lead to inaccurate and uncertain out-of-sample performance predictions. To overcome this limitation, Collins et al. (2016) developed a method for using prior data (i.e., records of performance data collected from previous classes) to inform a subset of PPE's free parameters (prior predictions), under cases where there were not enough data points for accurate calibration. By using prior data to inform a subset of PPE's free parameters, PPE fits the available training data with a constrained parameter set. In circumstances where there is little training data, this increases PPE's prediction accuracy for early performance events.

This prior-informed prediction method was based on work from the Educational Data Mining (EDM) literature. EDM research applies data mining and statistical learning methodologies to educational data to improve student learning outcomes (Romero, Ventura, & Baker, 2010). EDM methods are primarily data driven, meaning they require large amounts of data to develop predictions within a specific domain (Webb, Pazzani, & Billsus, 2001). In contrast PPE is primarily theory driven, meaning that its predictions are based on mechanisms that account for general characteristics of human learning and retention. The development of PPE's prior-informed prediction method balances the data-driven and theory driven approach of these two methods.

Although Collins et al. (2016) found that prior data could be used to generate predictions of the aggregate performance of multiple students attempting a single skill, their results did not indicate how accurate the predictions are at an individual student level of analysis. Using prior data to predict the initial performance at a finer level of aggregation is more difficult for two reasons. First, the performance of a single individual is characterized by greater variability, as compared to learning curves aggregating across the performance of multiple students, making performance of a single student more difficult to predict. Second, students are likely to learn skills at different rates, meaning that best fitting parameters for an aggregate learning curve may not generalize to account for the performance of a specific student attempting a particular skill.

In spite of these additional complexities when predicting the performance of individual students, educational data mining research has shown that prior data can be used to inform valid model parameter estimates for models used to account for the performance of individual students on single skills (Cen, Koedinger & Junker 2007; Beck & Chang 2007; Ritter et al., 2009). These findings suggest that prior data can serve as a useful tool that can be used to inform predictions of individual students and not just aggregate samples. In summary, we sought to expand our previous research by examining the extent to which our method for predicting early performance of groups of students generalizes to the individual student level of analysis. To evaluate the prior-informed method, we compare it against predictions to PPE's standard non-prior predictions during an individual student's first 4 attempts on a new skill.

Method

The data used in this report were obtained from Learnlab.com's DataShop (Koedinger, Baker, Cunningham, Skogsholm, Leber, & Stamper 2010), which is an online educational data repository for student log data. DataShop contains a collection of publicly available datasets from different math, science, and English classroom and tutoring studies. The data used in our analyses, consisted of log files of performance metrics of students completing their homework for an introductory physics class during six different semesters. Students used the ANDES tutoring system to complete their homework (VanLehn et al, 2005) at the United States Naval Academy (USNA). We chose these datasets because they contain the largest collection of data from multiple semesters collected from the same domain currently available on DataShop, allowing us to better investigate the utility of using prior data to inform PPE's performance predictions.

A single semester's worth of data on DataShop is called a dataset, which is composed of a record of the

performance of individuals who attempted to solve problems in a specific domain within a specific period of time. Each dataset contains the record of all of the students' actions across the curriculum's content. A curriculum is made up of problems, defined as "a task [attempted by] a student usually involving several steps." An example of a problem would be calculating the difference in velocity between trains A and B. Successfully solving a problem involves completing a series of steps, which are "an observable part of a solution to a problem", such as finding the velocity of train A. We choose to examine the performance of students while completing particular steps for two reasons. First, steps were the smallest level of resolution of data available on Datashop. Second, each step isolates a particular knowledge component. Because learning occurs at the level of individual knowledge components (Anderson & Schunn, 2000), comparing analogous steps across problems is the proper way to observe the change in performance over time.

Prediction Procedure

We systematically selected one of the six datasets as the prediction sample, and used the remaining five datasets as prior data to inform predictions for an individual on a particular step. Then the performance data of a single student on a particular skill was selctecd, from the prediction sample. All of the students from the prior data who also attempted the same skill were selected (relevant sample) and used to inform PPE's predictions. Due to the fact that the data collected from the ANDES tutoring system are data from homework assignments, the students' first exposure to the curriculum was during class and was not their first attempt on a particular step within the tutoring system. For this reason, we assumed a six-hour lag between class and when a student began to complete their homework. This assumption of a lag between class and home time allowed for a better estimation of PPE's model time as calculated from PPE's time variables (Eq. 1and 4). For the relevant sample to be able to inform a prediction, the average performance and model time variables across each participant during each event was calculated. Based on aggregate performance and model time computed from the relevant samples, PPE model parameters were estimated, and then used to make individualized predictions of a student's performance on a particular skill on the 2nd, 3rd, and 4th event.

For the analysis in this paper, we used PPE to generate predictions for two metrics of the students' performance: time to complete a particular step (seconds) and the number of incorrect attempts made by a student during a particular event. To generate a prior prediction, PPE first calibrated to the performance (i.e., completion time in seconds or number of incorrect attempts) of the first two events from the aggregate performance of the relevant sample. This yields a set of best fitting parameters values. The best fitting *b* (b_{prior}) and *m* (m_{prior}) parameters are then generalized to inform PPE's prior informed prediction of an individual student's performance on the 2nd event given their performance on the 1st event. This is done by setting PPE's *b* and *m* free parameters to the b_{prior} and m_{prior} values and fitting PPE's remaining two free parameters *s* and τ to the student's performance during the first event. After PPE is fitted to the student's performance on the 1st event, the model is used to generate a prediction of the student's performance on the second event. This procedure was then repeated to generate predictions of the 3rd and 4th event, by increasing the number of events that PPE is calibrated to with the prior sample and the predicted individual before generating a performance prediction of the next event.

In addition to generating prior predictions, we used PPE to generate predictions of each student's performance on the 2nd, 3rd, and 4th events without using data from past participants. This involved fitting the model with the sparse, individual-specific data, and using the model to predict performance for the following event.

Across all of the six datasets collected from Datashop, a total of 10,499 predictions were made across 430 students and 161 individual steps across the 2nd, 3rd, and 4th performance event.

Results

To examine the accuracy of PPE's prior and non-prior predictions the average model predictions from the 2nd, 3rd, and 4th events were compared to the average observations from students whose performance was predicted (Figure 1).

In addition to the looking at the average performance, the students' performance and PPE's predictions were separated in to two groups (i.e., canonical and non-canonical learning). The students in the canonical learning groups were students whose performance either improved or remained the same over the four observed learning events (Figure 2-A, 2-C). Students in the non-canonical learning group were students whose performance decreased during at least one of the four learning events (Figure 2-B, 2-D). The students' performance was separated into canonical and non-canonical learning groups, due to the fact the variability in the students' performance effects PPE's performance predictions. Additionally, we wanted to observe to test if PPE's could account for the two types of learning profiles.

Completion Time

As seen in Figure 1, when predicting a student's performance on the 2nd event, given their performance on the 1st event, there is a significant difference between the mean completion time between PPE's prior (*M = 45.50, SD = 85.50*) and non-prior (*M = 192.199, SD = 196.78*; $t(10497) = 90.932$, $p < .01$) predictions compared to the students' average completion time (*M* = 37.85, *SD* = 70.43). Examining the root mean squared deviation (*RMSD*) between PPE's prior (*RMSD* = 98.49) and non-prior predictions (*RMSD* = 250.18), we see that PPE's prior-informed predictions were more accurate than non-prior predictions. These results show that informing PPE's

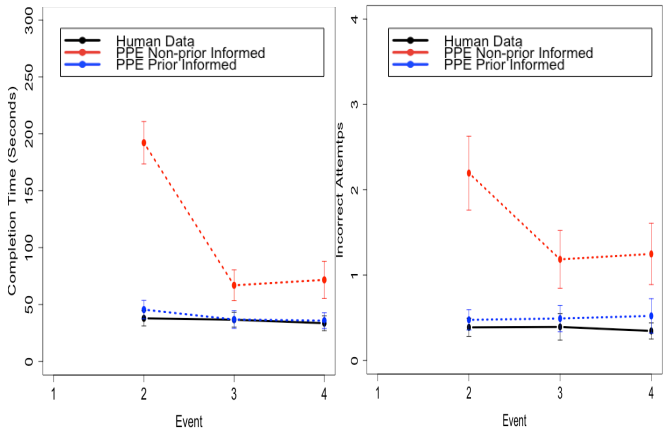

Figure 1. The average performance metric, completion time (seconds) (left plot) and number of incorrect attempts (right plot) on the 2nd, 3rd, and 4th event, across human data (solid black line), prior informed predictions (dashed blue line), and non-prior informed predictions (dashed red line).

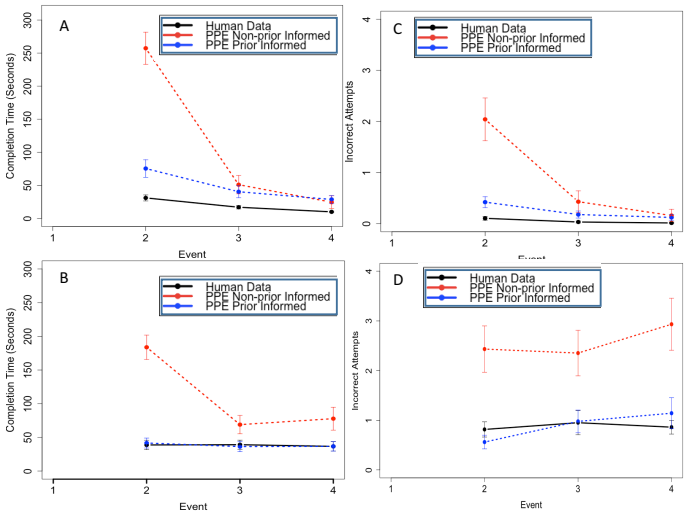

Figure 2. The average performance metric, completion time (A, B) and number of incorrect attempts (C, D) for both students who fit the canonical (A, C) and non-canonical (B D) learning profile, for both the human data (solid black line), non-prior predictions (dashed red line) and prior predictions (dashed blue line) on the 2nd, 3rd, and 4th event.

predictions using prior data can improve prediction accuracy when prediction performance of the 2nd event

When predicting the students' performance on the 3rd event, given their performance on the first 2 events, again a significant difference between PPE's prior (M = 36.79, SD = 81.76) and non-prior (M = 66.85, SD = 143.21; $t(10497)$ = 22.78, p < .01) predictions is observed, compared to the students' average performance (M = 73.96, SD = 179.11). As was seen when predicting the students' average performance on the 2nd event, a similar pattern is seen when predicting the 3rd event. A lower RMSD was found between the students' average performance and PPE's prior ($RMSD$ = 99.66) compared to non-prior predictions ($RMSD$ = 151.80).

Finally, when predicting the students' performance on the 4th event, given their performance on the previous 3 events, again a difference between the PPE's prior (M = 35.76, SD = 73.11) and non-prior (M = 71.63, SD = 172.86; t(10497) = 22.63, p < .01) predictions are observed, compared to the students' average performance (M = 33.56, SD = 68.26). Again PPE's prior informed predictions had a lower RMSD ($RMSD$ = 92.13) compared to the non-prior predictions ($RMSD$ = 181.53) when predicting the students' performance on the 4th event.

Correct Attempts: Canonical and Non-Canonical Learning Profile

Separating the students' performances into those who displayed canonical and non-canonical learning profiles, reveals two different sets of completion times. The performance of students who displayed a canonical learning profile was found to be monotonically improve over the course of the three events (Figure 2-A). Students who the non-canonical learning profile, on average displayed non-monotonic improvement in their performance over the four events (Figure 2-B). Additionally, it is seen that the accuracy of PPE's prior and non-prior predictions varied based on the performance of the students' learning profile. When predicting the performance of students' who showed a canonical learning profile, PPE's prior and non-prior predictions became more accurate as PPE was calibrated to additional events before generating a prediction, during the 2nd (Prior: $RMSD$ = 117.97; Non-prior: $RMSD$ = 310.42), 3rd (Prior: $RMSD$ = 82.90; Non-prior: $RMSD$ = 132.78), and 4th(Prior: $RMSD$ = 56.31;Non-prior: $RMSD$ =91.52) event (Figure 2-A). However, PPE's accuracy decreased when it was calibrated to each additional event when predicting performance of students' whose performance was found to have a non-canonical learning profile. When predicting the performance of students' who showed a non-canonical learning profile, PPE's prior and non-prior prediction accuracy decreased as PPE calibrated to additional events, during the 2nd (Prior: $RMSD$ = 41.26; Non-prior: $RMSD$ = 95.67), 3rd, (Prior: $RMSD$ = 101.62; Non-prior: $RMSD$ = 154.06) and 4th(Prior: $RMSD$ = 95.80; Non-prior: $RMSD$ = 190.80) event (Figure 2-B). Although, PPE's prediction accuracy varied based on the students' learning profile, PPE's prior performance predictions were more accurate than PPE's non-prior predictions.

Number of Incorrect Attempts

Examining the average students' number of incorrect attempts on the 2nd event given a students' previous performance on the first event (Figure 2), a large difference is observed in the predicted average number of incorrect attempts in PPE's prior (M = .47, SD = 1.25) and non-prior

(M = 2.19, SD = 4.58; t(10496) = 38.97, p < .01) predictions, compared to the students' average number of incorrect attempts (M = .38, SD = 1.09). Looking at the RMSD between PPE's predictions and the students' performance, PPE's prior (*RMSD* = 1.51) predictions had a lower RMSD than PPE's non-prior informed predictions (*RMSD* =4.87).

When predicting the students' average number of incorrect attempts (M = .39, SD = 1.63) on the 3rd event, again a significant difference between PPE's prior (M = *.49, SD = 1.63*) and non-prior predictions is observed (M = 1.20, SD = 3.58; t(10496) = 20.92, p < .01). However, unlike when predicting performance on the 2nd event, the RMSD of PPE's prior informed predictions increased (*RMSD* = 2.14). While as well as PPE's non-prior (*RMSD* = 3.91) decreased slightly.

Finally, when predicting the students' number of incorrect attempts on their 4th event, given their performance on the previous three events, a similar pattern of predictions is seen. A significant difference was observed between PPE's prior (M = .52, SD = 2.13) and non-prior predictions (M = 1.24, SD = 3.79; t(10496) = 22.62, p < .01), compared to the students' average performance was observed (M = .39, SD = 1.63). Additionally, the RMSD between the PPE's prior (RMSD = 2.22) and non-prior predictions (RMSD = 3.87) were not seen to improve. However, the PPE's prior informed predictions were lower than PPE's non-prior informed predictions.

Incorrect Attempts: Canonical and Non-Canonical Learning Profile

Separating the students' performance into those who displayed canonical and non-canonical learning profiles, two different sets of the students' number of incorrect attempts are seen. From students who displayed a canonical learning profile, number of incorrect responses decreased over the course of the four learning events (Figure 2-C). Conversely, students who displayed a non-canonical learning profile on average displayed a non-monotonic performance over the four events (Figure 2-D). The accuracy of PPE's prior and non-prior predictions varied based on the type of learning displayed by the students. When predicting the performance of students who showed a canonical learning profile, PPE's prior and non-prior predictions became more accurate when PPE calibrated to additional events, during the 2nd (Prior: *RMSD* = 1.04, Non-Prior: *RMSD* = 4. 69), 3rd, (Prior: *RMSD* = .68 Non-Prior: *RMSD* = 2.27) and 4th(Prior: *RMSD* = .56 Non-Prior: *RMSD* =1.29) event (Figure 2-C). However, PPE's accuracy decreased when it calibrated to additional events of students with a non-canonical learning profile. When predicting the performance of students' who showed a non-canonical learning profile, PPE's prior and non-prior predictions became less accurate as PPE calibrated to additional events, during the 2nd (Prior: *RMSD* = 2.03; Non-Prior: *RMSD* = 5.14), 3rd (Prior: *RMSD* = 3.30; Non-Prior: *RMSD* = 5.57), and 4th(Prior: *RMSD* = 3.34; Non-Prior: *RMSD* = 5.96) (Figure 2 –D). Although, prediction accuracy varied based on the students' average performance based on the learning profile of the student, PPE's prior performance predictions were more accurate than PPE's non-prior predictions.

Discussion

The primary goal of this paper was to describe our assessment of the accuracy of PPE predictions of performance in the tutoring data available on DataShop, both with and without the use of informative priors. We find evidence that incorporating prior data into PPE's predictions at a lower (individual student) level of aggregation, slightly improves prediction accuracy, depending on the performance measure, the event being predicted, and the student's learning profile.

When predicting a student's completion time on the 2nd, 3rd, and 4th event, we found that PPE's prior informed predictions were more accurate than PPE's individualized predictions. Additionally, we found that PPE's predictions varied based on the student's learning profile. When predicting the performance of students' who were found to have a canonical learning profile, the accuracy of PPE's increased as PPE was calibrated to additional events. However, the opposite results were observed when predicting the performance of students' who were found to have a non-canonical learning profile. Here it was observed that PPE's ability to predict performance depended on the variability of the students performance history in their performance. When variability in a student's performance history was low and improved regularly (i.e., canonical learning profile), PPE was better able to predict their future learning. When variability was high and a student's performance history showed both improvement and forgetting (i.e, non-canonical learning), the increased uncertainty in performance hindered the PPE's predictions from accurately predicting future performance. Although, the benefit of using priors was observed in PPE's predictions in each of these cases.

These results are partially consistent with results from Collins et al. (2016), where we found an initial benefit of using prior predictions to generate initial performance predictions of the 2nd event, as was found when predicting the student's completion time. Without information from prior data, PPE's parameters must be estimated with sparse data from the student's prior performance during the first event. Because the model is under constrained in this case, the parameter estimates are likely unreliable.

Additionally, when predicting the average completion time and the number of incorrect attempts, a benefit of using a priors was found. When predicting a student's future performance, PPE is able to utilize information from other students who have previously performed the skill before, allowing for a better estimate of the student's future performance will be. These findings are in line with our

previous findings that PPE's prior predictions benefit PPE's predictions beyond the 2nd event.

Conclusion

The benefits of using prior data are not new to cognitive science. However, within the context of the PPE line of investigation, little previous research has been conducted on how prior data can be used to inform predictions, especially within the context of early performance predictions of individual students. In summary, we find evidence that our previously proposed method of incorporating information from prior data into PPE's free parameters (Collins et al. 2016), can add some benefit to prediction accuracy when attempting to predict the performance of individual students on particular skills. The results suggest that prior data is a useful source of information about the performance of individual students when generating predictions with PPE. Future work should attempt to incorporate information from prior data to generate initial performance predictions in order to decrease overall training or education time.

Acknowledgements

The authors would also like to thank the Oak Ridge Institute for Science and Education (ORISE) who supported this research by appointing Michael Collins, to the Student Research Participant Program at the U.S. Air Force Research Laboratory (USAFRL), 711th Human Performance Wing, Human Effectiveness Directorate, Warfighter Readiness Research Division, Cognitive Models and Agents Branch administered by the ORISE.

References

Anderson, J. R., & Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. *Advances in instructional psychology, Educational design and cognitive science*, 5. 1-33.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993) Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science* , 4, 316-321.

Beck, J. E., & Chang, K. M. (2007, July). Identifiability: A fundamental problem of student modeling. In *International Conference on User Modeling*, 137-146, Springer Berlin Heidelberg.

Collins, M. G, Gluck, K. A., & Jastrzembski, T. S. (2015). Datashopping for performance predictions. *Proceedings of the Foundations of Augmented Cognition*, Los Angeles, California, (pp. 12-23).

Collins, M.G., Gluck, K.A., Walsh, M., Krusmark, M., Gunzelmann, G., (2016, July) Using prior data to inform model parameters in the predictive performance equation. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Philadelphia, PA

Cen, H., Koedinger, K., & Junker, B. (2006, June). Learning factors analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems* (pp. 164-175). Springer Berlin Heidelberg.

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

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias⁄variance dilemma. *Neural Computation, 4*, 1-58.

Jastrzembski , T. S. , Gluck , K. A. , & Gunzelmann , G. (2006) . Knowledge tracing and prediction of future trainee performance . In*: The proceedings of the interservice/industry training, simulation, and education conference* . Orlando, FL :National Training Systems Association (pp. 1498 – 1508).

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C.K.I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1321-1329). La Jolla, CA: NIPS Foundation.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and Their Acquisition*, 1, 1-55

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2), 101.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: Application of the SAM model. Cognitive Science, 27, 431–452.

Ritter, S., Harris, T. K., Nixon, T., Dickison, D., Murray, R. C., & Towle, B. (2009). Reducing the Knowledge Tracing Space. *International Working Group on Educational Data Mining*.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2010). *Handbook of educational data mining*. CRC Press.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734–760.

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., and Wintersgill, M. (2005). The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence and Education, 15* (3).