Using Prior Data to Inform Model Parameters in the Predictive Performance Equation

Michael G. Collins[1] (michael.collins.74.ctr@us.af.mil)
Kevin A. Gluck[1] (kevin.gluck@us.af.mil)
Mathew Walsh[2] (mmw188@gmail.com)
Michael Krusmark[1] (michael.krusmark.ctr@us.af.mil)
Glenn Gunzelmann[1] (glenn.gunzelmann@us.af.mil)

[1]Air Force Research Laboratory, 2620 Q Street, Building 852
Wright-Patterson Air Force Base, OH 45433
[2]TiER1 Performance Solutions100 E Rivercenter
Blvd #100, Covington, KY 41011

Abstract

The predictive performance equation (PPE) is a mathematical model of learning and retention that attempts to capitalize on the regularities seen in human learning to predict future performance. To generate predictions, PPE's free parameters must be calibrated to a minimum amount of historical performance data, leaving PPE unable to generate valid predictions for initial learning events. We examined the feasibility of using the data from other individuals, who performed the same task in the past, to inform PPE's free parameters for new individuals (prior-informed predictions). This approach could enable earlier and more accurate performance predictions. To assess the predictive validity of this methodology, the accuracy of PPE's individualized and prior-informed predictions before the point in time where PPE can be fully calibrated using an individual's unique performance history. Our results show that the prior data can be used to inform PPE's free parameters, allowing earlier performance predictions to be made.

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

Introduction

A common characteristic of training and education programs is that instructors have little or no information about the ability of specific students arriving to a particular class. Without information, instructors must wait until a certain amount of the curriculum has been completed before they can identify who possesses adequate or inadequate knowledge about a given topic, and before they can administer informed training interventions (e.g., removing or adding requirements for additional practice). To make more effective decisions about adaptive education or training interventions, instructors must anticipate the likely effects of specific actions. This would be enabled by models that can predict future performance if a particular training intervention were to be implemented.

In the field of cognitive science, mathematical models of learning and retention have been developed to predict individuals' future performance in a variety of different domains (Anderson & Schunn, 2000; Jastrzembski, Gluck, & Gunzelmann, 2006; Pavlik, & Anderson, 2008). These models can potentially be applied to education and training situations to support more accurate predictions of students' future performance.

One such model is the Predictive Performance Equation (PPE). PPE has been used to predict aggregate group performance and the performance of individuals on declarative (know-what) and procedural (know-how) tasks (Jastrzembski et al., 2006). Prior research has validated PPE with performance data collected in laboratory settings (Jastrzembski et al. 2006), training settings such as Air Force F–16 pilot testbeds (Jastrzembski et al. 2010), and educational settings involving classroom learning and tutoring systems (Collins, Gluck, & Jastrzembski, 2015).

The Predictive Performance Equation

PPE is a model of learning and retention that predicts future performance on the basis of three factors: (1) total amount of practice; (2) elapsed time since practice occurred; and (3) how practice was distributed across time. In general, performance increases with amount of practice (Factor 1), and decreases with elapsed time since practice occurred (Factor 2)(Anderson, 1995). The third factor, distribution of practice over time, is central to research on the spacing effect (for a review, see Cepeda, Vul, Wixted, & Rohrer, 2006). This research has shown that separating practice repetitions by a delay (i.e. *spacing*) slows acquisition but enhances retention. The spacing effect is one of the most widely replicated results in psychology research, and its potential implications for education and training are substantial.

PPE has three free parameters that are calibrated based on historical performance data (Equation 1).

$$\text{Performance} = S * St * N^{c} * T^{-d} \qquad (1)$$

The three free model parameters are S (scalar), used to accommodate the performance measure of interest (e.g.,

error rate, percent correct, response time, etc.), *c* (learning rate), and *d* (decay rate). The model's fixed parameters are determined by the timing and frequency of events in the protocol, such as *T*, the amount of time passed since the onset of training, and *N*, the number of training events that occurred in the training period. *St* (Equation 2) is the stability term that "captures the effects of spacing, by calibrating experience amassed as a function of temporal training distribution and true time passed" (Jastrzembski, Addis, Krusmark, Gluck, & Rodgers, 2010, p. 110).

$$ST = \frac{T_{diff_i}}{PT_i \cdot T_i} + 1 \qquad (2)$$

St is 1 for the first event and is calculated for all other events based on the elapsed time between the current and previous events (T_{diff_i}), the total amassed practice time (PT_i) and the elapsed time between the current event and the first event (T_i).

The individualized approach to using PPE involves gathering data from an individual during a series of calibration sessions, finding the values of PPE's free parameters that maximize the correspondence between the model's output and the individual's observed performance during the calibration sessions, and using the parameterized model to predict the individual's future performance. The basic idea is that although the structure of the model is invariant across individuals, cognitive processes and the psychological parameters that control them, such as rate of learning and rate of forgetting, may vary. Once PPE has been calibrated based on an individual's training history, it can be used to make personalized performance predictions and training prescriptions (Jastrzembski et al., 2010).

Motivation for Analyses

One limitation of PPE is that is has shown to be necessary to calibrate to a minimum of three instances of prior performance before generating a prediction. A minimum of three data points are needed to estimate values for PPE's three free parameters (*S*, *c*, *d*). With fewer than three points, parameter estimates for PPE are unlikely to be accurate because they are under constrained. Multiple different combinations of parameter values may account equally well for the existing data. Consequently, out-of-sample predictions will be highly uncertain at best, and highly inaccurate at worst.

PPE is moderately complex. When such a model is fit to a small number of data points, it is likely to capture the structure of data in addition to noise. This causes poor out-of-sample prediction (i.e., generalization, Geman, Bienenstock, & Doursat, 1992). When PPE is fit using too few data points, its parameter estimates may reflect unrealistic assumptions about psychological processes (e.g., complete learning or forgetting), subsequently causing it to fail to account for the future performance of a sample.

The model's inability to make valid out-of-sample performance predictions after calibrating to fewer than three events reduces its utility early in a training regimen. This creates a lag period during which students start to complete part of the curriculum and personalized performance predictions are not available. One source of information that can help inform predictions of initial performance is data collected from others who performed the same task in the past. Indeed, educational data mining (EDM) attempts to make use of previously collected student data for exactly this purpose. EDM researchers use various data mining techniques (e.g., clustering analysis, rule mining, Bayesian models) to leverage existing student data to predict the future performance of students and to inform educational decisions (Romero & Ventura, 2010). Despite the fact that EDM and PPE have similar goals, these approaches differ in two keys ways. First, EDM seeks to use machine learning techniques to discover regularities in new and often large datasets. The enterprise is mainly data-driven. PPE, in contrast, is based on a psychological model of how various factors impact human learning and retention (Bahrick, Bahrick, Bahrick, & Bahrick 1993; Newell & Rosenbloom 1981). Thus, PPE is theory-driven. The insights and constraints afforded by a psychological theory may enable more effective use of educational data sets (Walsh & Lovett, *in press*). Second, PPE generates precise point predictions of future performance using measurements of accuracy, error rate or response time. Little attention in EDM has been placed on generating such precise performance predictions. Instead, these techniques mainly involve understanding existing data, or generating educational predictions at higher levels of aggregation (e.g., final grades, test scores).

Although these approaches may differ, EDM research highlights the usefulness of prior data when attempting to understand how a particular set of individuals will learn over time, based on how others behaved in the past. Currently, PPE has no way of incorporating any information from prior data into its predictions. However, taking such information into account could be useful when there is not yet enough historical performance data for calibration.

In this paper, we report the results of an evaluation of a new method for applying PPE to decisions about the timing of practice opportunities, using prior data to inform its learning and decay parameters when generating predictions of initial performance. We call these *prior-informed* predictions to distinguish them from *individualized* predictions based on calibration of PPE to an individual's own data, however sparse. Using real-world tutoring data, we compared the accuracy of PPE's prior-informed and individualized predictions of both initial performance (i.e., 2nd and 3rd event) and the first event after initial performance (i.e., 4th event), allowing for a comparison of the two methods when given both an inadequate and adequate amount of prior data for calibration.

Method

All of the data used in this report was obtained from Learnlab.com's DataShop (Koedinger, Baker, Cunningham, Skogsholm, Leber, Stamper, 2010), which is an online data

repository. Datashop contains a collection of publicly available datasets from different math, science, and English classroom and tutoring studies. For this paper, the data consists of performance measures from homework assignments of students from six classes, all from different semesters, using the ANDES tutoring system at the United State Naval Academy (USNA).[1] These datasets were chosen because they contain the largest number of individuals from multiple semesters collected from the same domain currently available on Datashop.

A single semester's worth of data from the USNA on DataShop is referred to as a dataset[2], which is composed of a record of the performance of individuals who attempted to solve a set of problems in a specific domain during a particular period of time. Each dataset contains a record of the performance of each individual student across that curriculum's content. The 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 comparing 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*.

In the analyses presented here, we examined the aggregate performance of students as they attempted to complete an individual step over the course of the semester (i.e., sample). A sample consisted of a selection of two or more students from a single dataset who each had a minimum of four opportunities to attempt a particular step, and had an equivalent sample of students (i.e., two or more) from the remaining 5 datasets who also had a minimum of 4 opportunities to complete the same step. Using these criteria, 307 samples were identified and used for this analysis.

We examined the data at a step level 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.

Procedure

For the analysis presented here, we systematically controlled which of the 5 datasets were used to inform predictions made about the performance of samples from the 6th datasets, selecting one dataset at a time (prediction dataset) and using data from the remaining 5 datasets as prior data to inform our prior-informed predictions. Next, from the predicted dataset, a single sample of students who completed the same step (prediction sample) were chosen. Then a second sample of individuals who completed the same step from the 5 remaining datasets (prior sample) was selected to inform predictions (prior sample) of the 2nd, 3rd, and 4th event. Finally, the error rate (performance measure) was calculated, as measured by the percent of incorrect attempts on their first opportunity to solve a step during an event for both the prediction and prior sample.

The various timing variables were then calculated for both samples based on the observed number and distribution of practice repetitions (Equation. 2). PPE then calibrated to the prior sample's performance (i.e., percent incorrect) on the first two events, obtaining a set of learning and decay rates (c_{prior} and d_{prior}) that were then generalized to the predicted sample. By calibrating PPE to the performance of the first event of the predicted sample, fixing the learning and decay rate to the parameters generated from the prior sample (c_{prior} and d_{prior}), allowing only the scalar to fluctuate, a prediction of the predicted sample's performance on the 2nd event was obtained.

This procedure was again repeated to generate predictions of performance on the 3rd and 4th event, by increasing the number of calibration events PPE used to generate learning and decay parameters from the prior sample and the event that PPE calibrated its scalar to before predicting the sample's performance on the next event.

Along with each of PPE's prior-informed predictions, predictions of each sample's performance on the 2nd, 3rd and 4th event were generated using the individualized PPE procedure, calibrating up to the event preceding the one being predicted (e.g., if predicting the 3rd event PPE would be calibrated using data from up to the 2nd event) and then generating a prediction for the remaining events.

Results

To evaluate the accuracy of PPE's predictions, we calculated the root mean square deviation (RMSD) between the sample's performance and both PPE's individualized and prior-informed predictions at each event (2nd, 3rd, and 4th). The R^2, a common metric for model assessment was not computed, because we only examined the prediction accuracy of a single event – thus, R^2 could not be computed. After the RMSD was calculated between each prediction and the sample's performance on each event, the average RMSD between the sample's performance and PPE's individualized and prior-informed predictions was computed (Figure 1).

Comparing the average RMSD between each of the model's predictions and the samples' performance on each event, we observe that there was little difference in the accuracy between individualized and prior-informed predictions on the 3rd and 4th event. A difference in the accuracy between the two methods of predictions emerges when predicting the performance of the 2nd event given the

[1] The exact studies which were exported from DataShop and used in this paper are cited in the acknowledgement section, per the guidelines on the Datashop website.

[2] All of the definitions listed in this paper came from the DataShop's online glossary and can be found on https://pslcdatashop.web.cmu.edu/help?page=terms

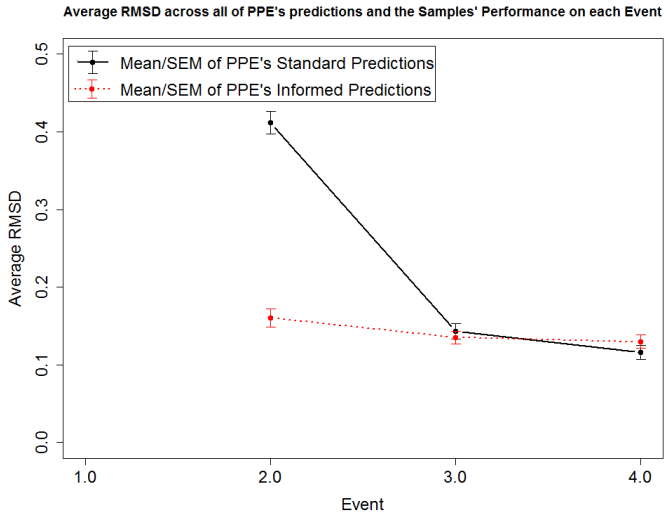

Figure 1. The average RMSD and plus and minus the standard error of the mean of the mean (SEM) between the samples' performance and the prediction of both PPE's individualized and prior-informed predictions.

performance of the first event (Figure 1). The average RMSD of PPE's individualized predictions increases dramatically, because calibrating to only the first event does not offer any information about the learning and decay rate of a sample. In comparison, the average RMSD of PPE's prior-informed prediction of the 2nd event increases only slightly, because it uses learning and decay rates informed by the prior data, allowing it to make an initial assumption based on prior data about the rate at which the sample will learn over time (Figure 2).

Calibrations Given Inadequate Prior Data

The first key question was whether PPE's prior-informed predictions were as accurate as the predictions generated by PPE's individualized approach when it had the opportunity to calibrate to three events before generating a prediction. To evaluate this question, we applied a paired t-test between the RMSD of each of PPE's individualized predictions and the sample's performance on the 4th event and the RMSD of each of PPE's prior-informed prediction and the sample's performance on the 3rd event.[3] We then applied the same test using PPE's individualized predictions and the sample's performance on the 4th event and the RMSD of each of PPE's prior-informed prediction and the sample's performance on the 2nd event. Both t-tests found that the predictions of the sample's performance on the 4th event, generated using PPE's individualized method, came from a different distribution than PPE's prior-informed predictions of both the 3rd ($t(306) = -1.91$, $p < .05$) and 2nd ($t(306) = -3.23$, $p < .001$) event (Figure 1).

These results reveal that PPE's prior-informed predictions of the 2nd and 3rd events were less accurate than when PPE calibrated to the performance of three events before generating a prediction. As predictions are generated earlier and earlier a loss of some prediction accuracy is expected and although PPE's prior-informed predictions were found to be less accurate as PPE's individualized predictions of the 4th event, the mean differences between the average RMSD between PPE's prior-informed and individualized predictions was not found to be extremely large.

Calibrations Given Adequate Prior Data

Finally, another paired t-test was run between the RMSD of the two methods when predicting a sample's performance on the 4th event. No significant difference was observed between the accuracy of these two predictions when predicting the sample's performance on the 4th event ($p > .05$). This result is not surprising in light of the overall average of PPE's free parameters used by these two methods as more data became available (Table 1). As PPE calibrates to additional data, the differences between parameters used by the individualized and prior-informed predictions begins to decrease.

Discussion

Our goal was to develop a method for PPE that could be used to generate predictions of a sample's initial performance when little historical performance data is available to calibrate the model. We accomplished this by calibrating PPE to a sample of individuals who performed

[3] All results were obtained using a Kolmogorov-Smirnov test, because the RMSDs were not normally distributed.

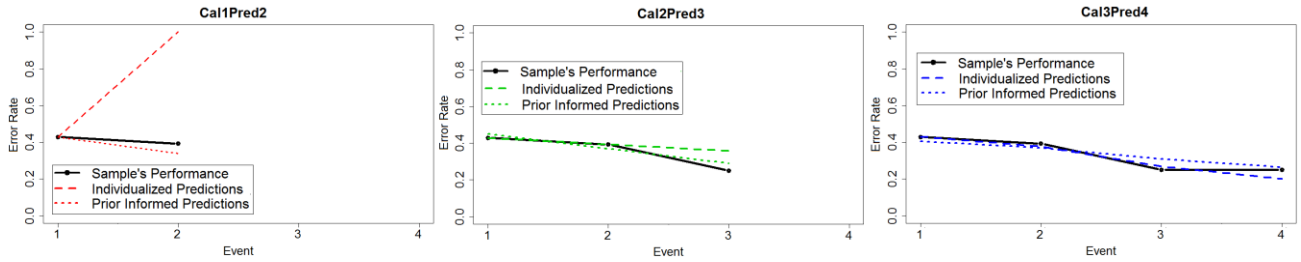

Figure 2: Predicted error rates based on prior-informed and individualized parameter estimates of a single sample from the USNA Dataset. Using the first event to predict the second event (Cal1Pred 2 – red line), two events to predict the third event (Cal2Pred3 – green line) and three events to predict the fourth event (Cal3Pred4 – blue line).

Table 1. The mean and standard deviation of the free parameters used to generate PPE's prior-informed and individualized predictions of each event.

	PPE Free Parameters								
	Learning Rate (*c*)			Decay Rate (*d*)			Scalar (*S*)		
Prediction Method	2nd Event M(SD)	3rd Event M(SD)	4th Event M(SD)	2nd Event M(SD)	3rd Event M(SD)	4th Event M(SD)	2nd Event M(SD)	3rd Event M(SD)	4th Event M(SD)
Prior-informed Predictions	.68(.27)	.58(.29)	.52(.29)	.10(.15)	.11(.18)	.09(.15)	4.15(4.77)	2.92(3.77)	2.65(3.80)
Individualized Predictions	.40(.17)	.51(.32)	.44(.29)	.23(.08)	.06(.17)	.08(.19)	.56(.15)	3.05(4.53)	2.81(4.65)

a particular task and then generalizing the estimated learning and decay rates to a different sample of individuals who performed the same task. This method allows for performance predictions to be made earlier, which may be needed in some training and education scenarios, if waiting until data is available to fully calibrate the model is not practical, such as when educational opportunities are spaced far apart or only a few educational events will be completed.

In our analyses, we first wanted to compare the predictions of initial performance generated using the individualized approach and a novel approach informed by the data of other students. A comparison of the average RMSD of these two methods revealed little difference in the performance predictions of the 3rd and 4th events. The greatest difference was seen in the accuracy of the prediction methods during the 2nd event. Here, the benefit of using learning and decay rates from a prior sample was apparent.

We also examined if PPE's prior-informed predictions could maintain the same level of accuracy as those generated by the individualized approach on the 4th event, after it had calibrated to three previous events. A paired t-test found that there was a significant difference between the accuracy scores generated from PPE's individualized method of the 4th event and PPE's prior-informed predictions of the 3rd and 2nd event.

A decrease in accuracy of predictions of events early in the learning process would be expected, due to the greater uncertainty about the parameter values controlling the process. Still, the accuracy of PPE's prior-informed predictions of the 2nd event are a large improvement in comparison to those generated using PPE's individualized approach.

Comparing the accuracy of the two methods when predicting a sample's performance on the 4th event, after enough data were available for PPE to adequately calibrate its parameters, revealed no difference between the individualized and prior-informed predictions. As PPE calibrated to additional events the difference in the free parameters used by individualized and prior-informed predictions began to diminish leading to similar predictions, and decreasing the need to rely on learning and decay parameters generalized from prior data.

Conclusion

The notion that prior data can improve a model's ability to make predictions is not a new. However, until now, PPE has relied solely on calibrating to historical performance data to determine the values for its free parameters, which has limitations when little historical performance data are available. As shown by the analyses presented here, the benefits of using prior data are substantial. Using prior data to inform predictions of a sample's performance on the 2nd event improved the prediction accuracy by 25% compared to predictions generated through PPE's individualized approach. We find that the learning and decay parameters estimated using prior data add a significant benefit to the predictive ability of PPE, when used under conditions when little historical performance data is available and the individualized calibration approach is unlikely to find a set of parameters that will characterize the future performance of a sample.

Besides using the only the aggregate learning and decay parameters from the prior data other approaches exist for combining existing data with an individual's unique performance history in order to better calibrate parameter estimates for the individual. The hierarchical Bayesian method involves estimating population *hyperparameters* that define the distributions from which an individual's parameters are drawn (MacKay, 2003). For example, PPE's parameters (*S*, *c*, and *d*) may vary across individuals, forming normal distributions at the level of the sample. The mean and variance of the distribution of each parameter are its hyperparameters. This approach balances the tension between maximizing the fit of a model to an individual's data, and maximizing the fit to the group's data. The chief advantage of this approach is that gradually assigns greater weight to an individual's own performance as more data becomes available. This enables a smooth transition from group-based parameter estimates to personalized estimates as the length of an individual's training history increases. More work is needed in order to compare this to the approach explored in this paper.

In conclusion, methods which attempt to generate earlier performance predictions, such as the one discussed here, can provide instructors the ability to better gauge what the expected performance of sample might be compared to their expected performance given a particular training intervention. The ability to make these performance comparisons earlier and with less performance data from students has the potential to boost early learning and improve the overall educational outcome.

Acknowledgements

We would like to thank Carnegie Learning, Inc., for providing the Cognitive Tutoring data supporting this analysis. We used the "USNA Physics Fall 2006", USNA Physics Fall 2007", "USNA Introductory Physics Fall 2009", "USNA Physics Spring 2007", "USNA Physics Spring 2008", "USNA Introductory Physics Spring 2010", accessed via DataShop. 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, Cognitive Models and Agents Branch.

References

Anderson, J. R. (1995). *Learning and Memory*. New York: Wiley.

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.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354–380.

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

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*: Proceedings of the interservice/industry training, simulation, and education conference*. Orlando, FL, 1498 – 1508.

Jastrzembski, T. S., Addis, K., Krusmark, M., Gluck, K. A., & Rodgers, S. (2010) Prediction intervals for performance prediction. In: *Proceedings of 10th annual International Conference on Cognitive Modeling*, Philadelphia, PA, 109-114.

Jastrzembski, T. S., Juvina, I., & McKinley, A. (2013) Neurobiological extensions to a mathematical model for performance enhancement observed under conditions of noninvasive brain stimulation. In: *Proceedings of 12th annual International Conference on Cognitive Modeling*, Ottawa, Canada, 131-136.

Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010). A data repository for the EDM community: the pslc datashop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, England: Cambridge University Press.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6), 601-618.

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.

Walsh, M. M., & Lovett, M. C. (*in press*). The cognitive science approach to learning. In S. Chipman (Ed.), *Oxford handbook of cognitive science*. Oxford: Oxford University Press.