Extending the Predictive Performance Equation to Account for Multivariate Performance

Michael G. Collins (Collins.283@wright.edu)
ORISE at Airforce Research Laboratory, Wright-Patterson
Air Force Base, Dayton, Ohio USA

Florian Sense (florian.sense@infinitetactics.com)
InfiniteTactics, LLC,
Dayton, OH, USA

Michael Krusmark (michael.krusmark.ctr@us.af.mil))
L3Harris Technologies
Melbourne, FL, USA

Tiffany S. Jastrzembski (tiffany.myers.1@us.af.mil))
Air Force Research Laboratory, 711th Human Performance Wing
Wright-Patterson Air Force Base, Dayton, OH, USA

Abstract

Adaptive scheduling systems aim to estimate the ability of an individual in order to prescribe a personalized training schedule. These adaptive systems are often founded on regularities of human memory such as a learning, forgetting, and the spacing effect. One such model which has been developed to both account for regularities of memory and be used in applied contexts is the Predictive Performance Equation (PPE). One limitation of the PPE is that it is only able to account for and incorporate information about a participant's accuracy on a task and cannot take into account additional performance measures such as reaction time. To expand the PPE, we propose a simple extension to the model, allowing it to account for both accuracy and reaction time measures. Our paper reports the extension to the PPE as well as a formal model comparison to another model of learning and retention (Pavlik and Anderson, 2005). The results of our model comparison reveal that the extended PPE can both better account and predict an individual's performance than Pavlik and Anderson (2005) model.

Keywords: memory, learning, decay, spacing effect, mathematical modeling, model comparison, model identifiability, reaction time, adaptive scheduling

Introduction

Adaptive scheduling systems aim to develop technologies that track the performance and estimate the ability of an individual over time in order to prescribe when an individual should be trained on a skill(s) again to maintain or achieve a predetermined level of competency. These adaptive scheduling systems can take many forms, including educational tutoring systems, medical skills training, or fact based learning systems. The benefit of these adaptive systems is that they hold potential for reducing the overall amount of training time required for an individual to reach and maintain some objective performance criteria by tailoring the time of training to the individual and their current estimated ability.

To interpret the historical performance of learners and make prescriptive judgments about their future performance, cognitive models of learning and retention have been developed based on three stable findings of human memory. The first is the power law of learning, where performance is observed to improve over repeated instances of practice (Newell & Rosenbloom, 1981). The second is the power law of forgetting, where performance is observed to decrease as the time between instances of practice increases (Rubin & Wenzel, 1996). The third is the spacing effect, where the retention of information is improved by distributing practice over time compared to the equivalent amount of practice conducted within a shorter period of time (i.e.,massed schedule) (Bahrick, Bahrick, Bahrick, & Bahrick, 1993). To date, multiple models of memory have been developed which account for these empirical phenomena across a wide range of situations (Pavlik & Anderson, 2005; Walsh et al., 2018; Raaijmakers, 2003). In this paper, we focus on two specific models of learning and retention, the Predictive Performance Equation (PPE) (Walsh et al., 2018) and Pavlik and Anderson's (2005) (P&A) spacing model.

Accuracy and Response Time Measures

The main focus of the research with PPE and other models of learning and retention has been to account for an individual's objective performance, be it correct or incorrect response on an item or overall accuracy on a task(s). The primary attention to accuracy is understandable since ability to recall has been the primary focus of the memory literature and is often the focus of improvement in a training scenario. However, other dependent measures of performance can be informative from a training and education perspective, such

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

as response time (RT). RT has been shown to follow the same qualitative patterns as accuracy (Tenison & Anderson, 2016), is often closely associated with the ability to recall from memory (i.e., memory strength)(Sense, Jastrzembski, Krusmark, Martinez, & Van Rijn, 2019), and can be used as a measure of cognitive processing (Tenison & Anderson, 2016). For these reasons, RT can be an informative metric to assess an individual's ability under situations where a participant's accuracy has plateaued, which is common in situations of skill acquisition or within a long training scenario.

As of now, PPE lacks a formal way of simultaneously accounting for both an individual's RT and accuracy measures. Previous attempts to handle RT with the PPE either have tried to create a new performance metric which combines RT and accuracy into a single metric (Sense, et al. 2019) or normalize RT data (Collins, Tenison, Gluck, & Anderson, 2020). In contrast to the PPE, another model of learning and retention, which can take into account all three of the memory characteristics previously discussed and simultaneously account for accuracy and RT data is Pavlik and Anderson's (2005) spacing model. Here we examine both the PPE and P&A model and then propose an amendment to the PPE based on the formulation of the P&A model that allows the PPE to take into account accuracy and RT measures.

Predictive Performance Equation

The standard PPE is composed of six individual equations, containing 5 free parameters. At the center of the PPE is the Activation term M_i (Eq. 1), which is a product of a learning term ($a + N^c$) and a forgetting term (T^{-d}). The learning term is a function of the number of exposures to a task (N), free parameter a used to represent prior knowledge on a task, raised to a constant learning term (c).

$$M_i = (a + N)^c * T_i^{-d} \quad \text{(Eq .1)}$$

The forgetting term is a function of model time (T, Eq. 2), which is a weighted (Eq. 3) average of the elapsed time (t_i) between exposures to a task.

$$T_i = \sum_{i=1}^{n-1} \quad w_i * t_i \quad \text{(Eq .2)}$$

$$w_i = t_i^{-x} \sum_{j=1}^{n-1} \quad \frac{1}{t_j^{-x}} \quad \text{(Eq. 3)}$$

The decay rate (d) of the forgetting term (Eq. 5) is a function of the Stability term (St_i., Eq. 4), which is a cumulative average of time between task exposures. In addition to the stability term, two free parameters augment the intercept (b) and slope (m) of the decay parameter.

$$St_i = \left(\frac{1}{n-1} * \sum_{j-1}^{n-1} \quad \frac{1}{ln\,(lag_j + e)}\right) \text{(Eq. 4)}$$

$$d_i = b + m * St_i (Eq. 5)$$

Finally, to generate a prediction of performance, PPE's activation term (M_i) is nested within a logistic function (Eq. 6), which is controlled by two additional free parameters, τ and s, controlling the intercept and slope of predicted performance.

$$Prob_i = \frac{1}{1 + exp(\frac{\tau - M_i}{s})} \quad \text{(Eq. 6)}$$

Pavlik and Anderson (2005)

Pavlik and Anderson's (2005) spacing model is composed of two main components: an activation term, A_i, (Eq. 7) and decay rate, d_i, (Eq. 8) and contains two transformation functions to estimate the probability of recall (Eq. 9) and response time RT (Eq. 10). The activation term (Eq. 7), is a log sum of the time between the current time (t) and each previous presentation of an item (t_j) offset by an estimated base level constant ($Base_i$).

$$A_i = log\,(\sum_{j=1}^{n} \quad (t_i - t_j)^{-d_j}) + Base_i \text{ (Eq . 7)}$$

The strength of each previous presentation is proposed to decay according to the current decay rate (d, Eq. 8). The amount of decay applied to each instance depends on an item's previous activation (A_{i-1}), which is offset by two free parameters controlling the intercept (α) and slope (c) of the decay parameter.

$$d_i = \alpha + ce^{A_{i-1}(t)} \text{ (Eq. 8)}$$

To generate predictions of accuracy, P&A model uses the same function as PPE (Eq., 6), the Activation term (A_i) is nested with a logistic function, augmented by two free parameters τ and s (Eq. 9).

$$Prob_i = \frac{1}{1 + e^{\frac{\tau - A_i}{s}}} \text{ (Eq. 9)}$$

To generate predictions of response time (RT) the Activation term is nested in an exponential function, which is manipulated by an intercept parameter F (Eq.10).

$$RT_i = Fe^{-A_i} \text{ (Eq. 10)}$$

Modification to the PPE The formulation of both the PPE and P&A share a great deal of similarity. Each model has an underlying activation term (Eq. 1 & 7), which is then nested within a logistic transformation function to estimate the performance of an individual. The largest differences between the two models lies in their assumptions about how each model summarizes the passage of time and the construction of their decay term. See Walsh et al. (2018) for a detailed comparison and discussion of the theoretical differences between the two models. Besides the differences between each model's underlying formulations, another difference is in the dependent measures each model can account for: P&A's model is able to account for both

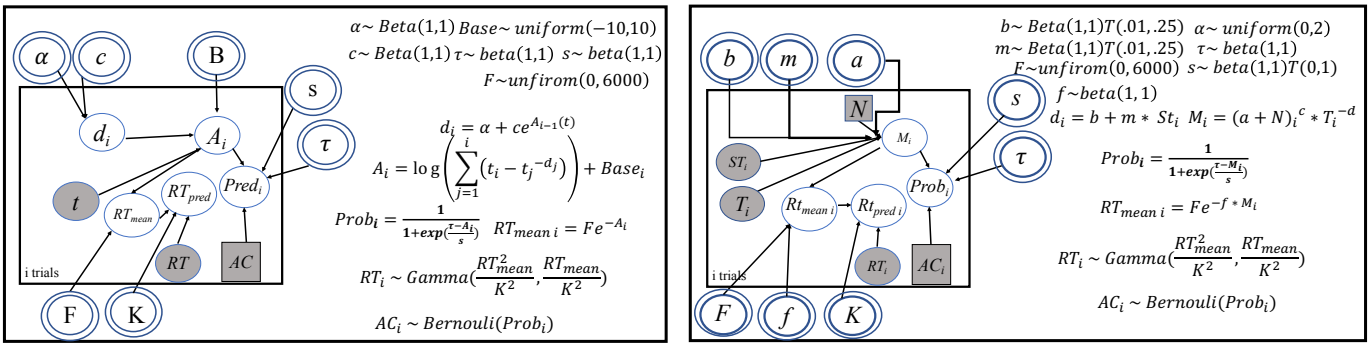

Figure 1. A graphical representation of the Bayesian implementation of Pavlick and Anderson (2005) spacing model (left plot) predictive performance equation (right plot).

accuracy and reaction time measures; while PPE can only account for accuracy.

However, due to the similarity between the two models, we explore generalizing P&A's model to account for RT data (Eq. 10) to the PPE (Eq. 12). To allow the PPE to take into account RT time data, we nest PPE's activation term within an exponential function, augmented by an intercept parameter (F) and an additional slope parameter (f) (Eq. 12). The slope parameter (f) was added because it was found that PPE without a slope parameter had difficulty accounting for response time data. We believe that this is necessary because PPE's activation term is not on an unbounded scale as Pavlik and Anderson's (2005) activation term which can range from $-\infty$ to ∞.

$$RT_i = Fe^{-f*M_i} \text{ (Eq. 12)}$$

In order to validate the new extension of the PPE we conduct a formal model comparison between the PPE and P&A's model using data from a spacing study (Walsh, Krusmark, Jastrembski, Hansen, Honn, & Gunzelmann, 2022).

Method

Participants Participants (N = 38) who were part of the control condition in a larger fatigue study were recruited for this study (Walsh, Krusmark, Jastrzembski, Hansen, Honn, & Gunzelmann, G, 2022).

Task Stimuli During the study participants completed a standard memory recognition task. Participants learned the matching pair of a two digit numbers and random line drawings, called droodles (Nishimoto, Ueda, Miyawaki, Une, & Takahashi, 2010).

Experimental Design The experiment was conducted over a period of three days where participants learned number and droodle pairs over 17 individual learning schedules (Table 1). Days 1 and 2 were the learning phase of the experiment where each digit/droodle pair was presented 20 times. Day 3 was testing phase where pairs were each presented 2 times. Of the 17 learning schedules there were 5 different types of schedules, with 3 digit/droodle pairs used per schedule. First were the long schedules (Long two, four and ten), where items were presented over the course of both days and the number of droodle pairs presented during each session were manipulated. The second type of schedule was the Massed schedule, where items were presented during one session, early or late in the day having either a 36 or 12 hour retention period. The third type of schedule was Level, where items were presented evenly across all sessions during Day 1 or Day 2. The fourth and fifth type of schedule was Ramp and Taper, where the number of item presentations either increased (Ramp) or decreased (Tamper) over the 4 sessions. The final type of learning schedule was Short which presented items 10 times over 2 sessions (Short 2) or 5, 10, and 5 times over three sessions (Short 3).

Table 1. A visual representation of the the17 individual schedules, showing the time, number of presentations per session, that Participants learned digit droodle pairs over the course three days

	Day 1					Day 2						Day 3
Overall Sessions	1	2	3	4	5	6	7	8	9	10		11
Sessions per day	1	2	3	4	5	1	2	3	4	5		1
Time	9:00	13:00	15:00	19:00	21:00	9:00	13:00	15:00	19:00	21:00		9:00
Level	4	4	4	4	4							2
Massed Early	20	-	-	-	-							2
Massed Late	-	-	-	-	20							2
Taper	6	5	4	3	2			24 hour				2
Ramp	2	3	4	5	6							2
Short Two	-	10	-	10	-							2
Short Three	-	5	10	5	-							2
Level						4	4	4	4	4	12 hour lag	2
Massed Early						20	-	-	-	-		2
Massed Late						-	-	-	-	20		2
Taper						6	5	4	3	2		2
Ramp						2	3	4	5	6		2
Short Two						-	10	-	10	-		2
Short Three						-	5	10	5	-		2
Long Ten	2	2	2	2	2	2	2	2	2	2		2
Long Two	-	-	10	-	-	-	-	10	-	-		2
Long Four	-	5	-	5	-	-	5	-	5	-		2

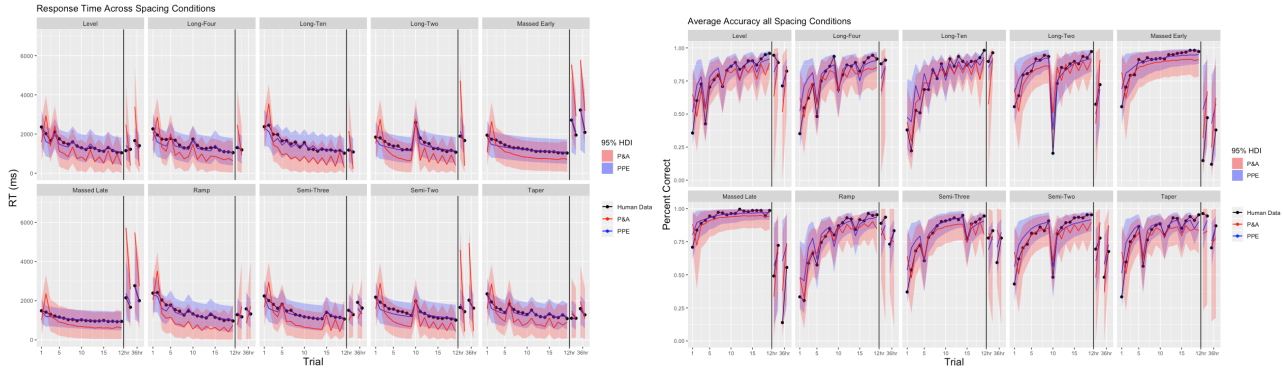

Figure 2 The average performance (response time – left plot, accuracy – right plot) of participants (solid black line) and model fit (trials 1-19) and predictions after a 12 and 36 hr delay (trials 20-23) of the PPE(blue line & ribbon) +/0 95% HDI and P&A +/- 95%HDI model (red line and ribbon).

Procedure To complete the experiment participants sat down in front of a computer screen, where the droodle was displayed. Participants were given up to 6 seconds to type with their dominant hand the droodle's associated two digit number. Once the participant entered their response, they received feedback on whether their answer was correct or incorrect, in which case they had two seconds to examine the correct answer. After receiving feedback participants were shown the next stimuli.

Bayesian Models In this paper both PPE and P&A were implemented as Bayesian models in JAGS. Each model is represented in a graphical format (Figure 1. P&A – left plot, PPE- right plot;), showing each variable type (circles - continuous variables, squares - discrete variables) each model's free parameters (unshaded circles) and fixed variables (shaded circles and square), along with variable type (stochastic elements - double circles and deterministic elements single circles). A comparison between the two models again reveals a great deal of similarity between each model's underlying formulation. Each model's activation term (PPE - M_i, P&A - A_i) is the deterministic product of fixed inputs (PPE - N_i, T_i, ST_i, ; P&A - t, t_j,) and free parameters (PPE - b, m, a, P&A - α , c, $Base$). The activation term of both models is nested within a logistic function to estimate the probability of a correct response augmented by the τ and s parameters. The probability of a correct response served as a parameter for a Bernoulli distribution to account for the participant's accuracy on each trial. To account for the participant's response time, each model's respective activation term is nested within an exponential function, augmented by an intercept parameter (F) with an addition of a slope parameter (f) for PPE only. The mean response time is then combined with a free parameter K controlling the variance of the distribution to determine the shape and rate parameter of a gamma distribution to account for the participant's observed RT measure on each trial. To compare each models' ability to both fit and predict the participants' performance over the course of the experiment, each model was calibrated to the performance of each individual number - droodle pair over the first two days of the experiment and then predicted the participants' performance on the 3rd day after a 12 or 36 hr lag (Table 1). Each model was run with 3 independent chains for 9000 iterations with a 2000 iteration burn in period. All chains were visually inspected for convergence.

Results

Here we evaluate how well the PPE and P&A model were able to calibrate and predict the participants' accuracy and response times (RT) of performance across the 17 individual schedules. To assess each models' performance, we compared each model's aggregate performance (Figure 2) and individual fit statistics across the 17 learning schedules for both dependent measures (i.e., accuracy and RT) (Figure 3). Finally, we computed Bayes Factors for both accuracy, RT, and for both dependent measures of the model's predictions of the participants recall session.

Accuracy: From a visual inspection of each model's fit and predictions of the participant's accuracy measure (Figure 2 – right panel), little difference is observed. During the acquisition period (trials 1-19), both PPE and P&A are seen to qualitatively fit the average accuracy of each of the training schedules. However, one consistency seen across the model fits is that PPE inferred higher average levels of initial performance early in the acquisition phase compared to the P&A model, but then later on P&A model is seen to under fit the participants' performance compared to the PPE. Though little difference is observed between the acquisition period, larger differences are seen between the models' predictions after the 12hr (trial 20 & 21) and 36hrs (trial 22 & 23) lag . Over a majority of the schedules, the P&A model consistently under predicts participants' performance relative to the PPE, with the exception of the Massed Early and Late schedule. Additionally, the 95% Highest Density Intervals (*HDI*) of the

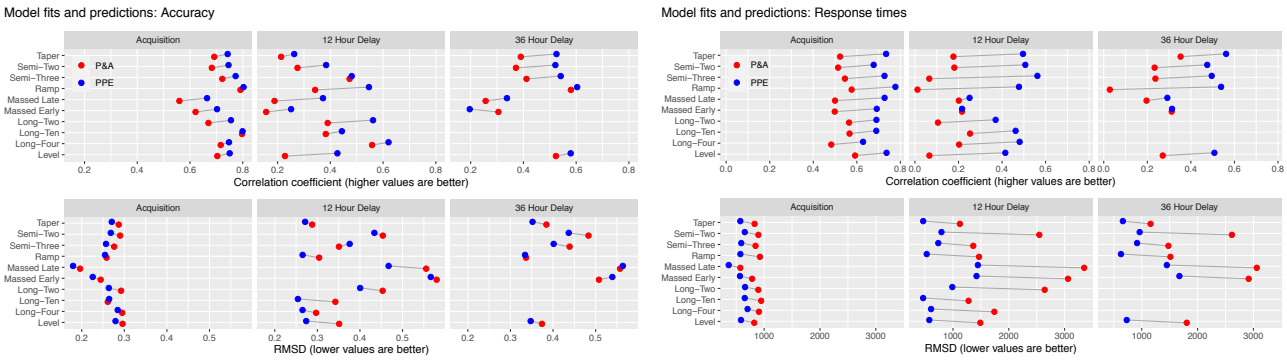

Figure 3. The correlation (r) (upper panel) and RMSD (lower panel) between the PPE (blue dot) and Pavlick & Anderson's (2005) (red dots) and the participants' individual performance during the models' fits (acquisition) predictions of the 12 and 36 hour delay for the participants accuracy measures (left panel) and response time measures (right panel).

P&A model is consistently larger than the PPE's. One interpretation of the model's *HDI* is that P&A is more uncertain about the participants' expected performance compared to the PPE. Turning to the model assessment at the individual level (Figure 3) it can be seen that across all schedules when both fitting and predicting participant's accuracy measures, PPE was found to have a higher correlation (*r*) and lower *RMSD* compared to P&A model across all 17 schedules.

Reaction Time An examination of each model's fit and predictions of the participants RT measures warrants similar conclusions when compared to each models' performance when accounting for participants' accuracy (Figure - 2 left panel). Both models fit the qualitative pattern seen in the participants' average response time measures. One clear visual difference that is observed between the two models are that the P&A model generates faster response times consistently across all schedules, while PPE better captures the average RT across all 17 schedules. Looking at each model's out of sample predictions (12 - 36 hour lag), it is consistently seen that the P&A model over predicts participants' response times across each of the schedules, with the exception of the Taper schedule, and under predicts response times in the Ramp and Semi-three schedule. Again, in contrast, PPE's out of sample predictions does a better job at predicting average RT across each of the schedules. These findings are again supported when looking at the descriptive statistics evaluating each model at the individual participant level, for each of the learning phases (acquisition, 12 hrs and 36 hrs recall) (Figure 3), PPE was found to better fit and predict participants RT measures compared to the P&A model.

Bayes Factor Finally, in order to quantify the differences between the predictions of both models, a Bayes factor for each individual participant across the 17 schedules was computed. For a holistic comparison, three different BF were compared comparing the models' predictions for accuracy, reaction time, and overall performance (accuracy and reaction time). Our results found that the majority of the participants' data (81%) were much more likely under the PPE model (BF > 3), while 7 of the participants' data (19%) were more likely under the P&A model (BF < 1/10). Comparing the two models based on each participant's reaction time data and their overall performance all participants were found to be more likely to come from the PPE model than the P&A model (BF > 10).

Discussion

In this paper we compared two models of learning and retention and their ability to account for and predict participants' accuracy and response time data. Both the PPE and Pavlik and Anderson's (2005) model have been shown to account for the same psychological phenomena (i.e., power law of learning, power law of forgetting, and spacing effect) and have been used in real world applications (Pavlik & Anderson, 2008; Sense, Van der Velde & van Rijn, 2021 Gluck, Collins, Krusmark, Sense, Maaß, & van Rijn, 2019). One difference between the two models is that PPE lacks a formal way to account for an individual's RT data. This limitation makes PPE a one dimensional model of memory that bases its prescriptions only on an individual's historical accuracy. To rectify this limitation, an additional transformation function, generalized from Pavlik and Anderson's (2005) spacing effect model and the ACT-R architecture (Anderson, 2007), was added to the PPE, allowing the PPE's activation term (M_i) (Eq.,1) to account for RT accuracy. To evaluate the PPE's new formulation, PPE was compared to Pavlik and Anderson's (2005) model by assessing each model's ability to both fit and predict a participants' performance (RT and accuracy) across a variety of learning schedules.

The results of our model comparison were clear: when accounting for both of the participants' dependent measures (accuracy and RT), PPE both better fit and predicted (Figure 3) a majority of performance metrics across each of the 17 schedules in the experiment. The results of this comparison were further supported by the evaluation of the Bayes factor

for each participant. When looking at each model's accuracy predictions, PPE was favored over P&A model for 31 out of 38 participants. While the Bayes factor for both predictions of reaction time and overall metric, unanimously favored PPE. These results highlight that despite the complexity of the PPE, its ability to better predict the participant's performance suggests the additional complexity of PPE warranted.

The results reported here in this paper between the PPE and P&A model are consistent with a previous model comparison by Walsh et al. (2018). However, the previous comparison only compared the two models based on their ability to account for accuracy. The differences in the performance between these two models likely stems from the construction of each model's decay term. Pavlik and Anderson's (2005) model formulates the most recent activation values within the decay term (Eq. 8, 9, 10) and each of the transformation functions, making the activation term entirely responsible for accounting for both the observed performance (Eq., 9,10) and for memory decay (Eq.. 8). In contrast, PPE casts the decay term as a function of an item's previous presentation history (Eq., 4), giving for PPE two separate mechanisms one for recall (Eq. 5) and decay (Eq. 6) (Walsh et al., 2018). This formulation of the PPE is in line with Bjork and Bjork's (1992) theory of use and disuse.

Finally, the results reported in this paper have several implications for using PPE within an adaptive learning system. First, being able to take into account both dependent measures of performance (accuracy and RT) could improve PPE's ability to make finer grained assessment of an individual's ability on a skill or piece of factual knowledge. For example, PPE could use the estimated performance of both response time and accuracy across a range of learned material to better determine what material would benefit the most from study. Such information can be extremely useful when an individual has to learn a large amount of information (e.g., foreign language learning) and the adaptive system has to attempt to prescribe material which would benefit the most from practice. Second, accounting for both accuracy and response time using PPE could better allow an adaptive system to determine when an individual has reached an appropriate level of ability or reached mastery on some material. Determining performance mastery using several different performance values could improve an adaptive systems ability to insure that an individual is trained to a level of ability that is required when the individual must apply the knowledge in a real world situation.

Limitations and Future Research

Though the results from the model comparison presented here are promising, several limitations should be noted. One limitation is how the model accounted for response time. We used a gamma distribution for our likelihood function which has been used to account for response time data in other studies (Tenison & Anderson, 2016). Though one criticism has been that the parameters of a Gamma distribution are difficult to interpret. Future research should explore using other candidate distributions, such as log normal or Weibull distribution to account for response time data. Second, Pavlik & Anderson (2005) have suggested offsetting the decay intercept between sessions arguing that decay between sessions is slower then within sessions. We choose not to implement this between session offset of the model to allow more direct comparison between the models as been in previous model comparisons (Walsh, et al. 2018).

Finally, the work presented in this paper have implications for two direct lines of future research. First, in this paper we fit each model to the individual items learned over the course of the experiment for each participant. However, these models could naturally be extended to a hierarchical format. A hierarchical model implementation would allow for more constrained implementation of PPE and allow for parameters to be estimated at multiple levels of aggregation. Second, the extended multivariate version of the PPE has the potential to expand the capability of PPE to account for more complex skill acquisition tasks, where accuracy often plateaus and reaction time is the primary informative dependent measure (Tenison & Anderson, 2016).

Conclusion

In the area of adaptive scheduling, accounting for a participant's accuracy and response time to learned items has been highlighted as being important to prescribe schedules of practice for an item. In our previous research, the PPE has primarily focused on accounting for measures of accuracy. The evaluation of this extension shows that PPE can be extended to both account for and predict accuracy and response time measures. Furthermore the PPE's extension evaluated here, generalized from Pavlik and Anderson, (2008), was originally taken from the ACT-R architecture. This provides further theoretical grounding for the PPE and links PPE to a broader theory of memory. In conclusion, our results add further empirical evidence that PPE can be used as a prescriptive adaptive scheduling tool, which can account for multiple dependent measures.

Acknowledgements

MGC's participation was enabled through an appointment to the Oak Ridge Institute for Science and Education (ORISE) Student Research Participation Program.

References

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(5), 316-321.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. From learning processes to cognitive processes: Essays in honor of William K. Estes, 2, 35-67.

Collins, M. G., Tenison, C., Gluck, K. A., & Anderson, J. (2020). Detecting Learning Phases to Improve Performance Prediction. In Proceedings of the 18th International Conference on Cognitive Modeling.

Gluck, K. A., Collins, M. G., Krusmark, M. A., Sense, F., Maaß, S., & van Rijn, H. (2019). Predicting performance in cardiopulmonary resuscitation. In Proceedings of the 17th International Conference on Cognitive Modeling (pp. 53-58).

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving Adaptive Learning Technology through the Use of Response Times. Grantee Submission.

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

Nishimoto, T., Ueda, T., Miyawaki, K., Une, Y., & Takahashi, M. (2010). A normative set of 98 pairs of nonsensical pictures (droodles). *Behavior research methods*, *42*(3), 685-691.

Pavlik Jr, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognitive Science, 29(4), 559-586.

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. (2003). Spacing and repetition effects in human memory: Application of the SAM model. Cognitive Science, 27(3), 431-452.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. Psychological review, 103(4), 734.

Sense, F., van der Velde, M., & van Rijn, H. (2021). Predicting university students' exam performance using a model-based adaptive fact-learning system. *Journal of Learning Analytics*, *8*(3), 155-169.

Sense, F., Jastrzembski, T. S., Krusmark, M., Martinez, S., & van Rijn, H. (2019). An Integrated Trial-Level Performance Measure: Combining Accuracy and RT to Express Performance During Learning. In CogSci (pp. 1029-1034).

Tenison, C., & Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. Journal of Experimental Psychology: Learning, Memory, and Cognition, 42(5), 749.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., … & Zhou, R. (2018). Mechanisms underlying the spacing effect in learning: A comparison of three computational models. Journal of Experimental Psychology: General, 147(9), 1325.

Walsh, M., Krusmark, M., Jastrzembski, T., Hansen, D. A., Honn, K. A., & Gunzelmann, G. (2022, January 23). Enhancing Learning and Retention through the Distribution of Practice Repetitions across Multiple Sessions. https://doi.org/10.31234/osf.io/dmf4p