Improving Predictive Accuracy of Models of Learning and Retention Through Bayesian Hierarchical Modeling: An Exploration with the Predictive Performance Equation

Michael Collins
michael.collins.ctr@us.af.mil /collins.283@wright.edu
ORISE at Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Florian Sense
f.sense@rug.nl / floriansense@gmail.com
InfiniteTactics, LLC, Beavercreek, OH, USA
Department of Experimental Psychology & Behavioral and Cognitive Neuroscience
University of Groningen, Groningen, The Netherlands

Michael Krusmark, **Tiffany S. Jastrzembski**
{michael.krusmark.ctr, tiffany.jastrzembski}@us.af.mil
Air Force Research Laboratory, Wright-Patterson Air Force Base, Dayton, OH, USA

Abstract

Human learning has been characterized by three robust effects (i.e. power law of learning, power law of decay, and spacing), which have been validated across multiple domains and time intervals. To account for these different effects mathematical model of learning and retention have been developed. These models hold a great deal of potential for application a wide range of educational and training scenarios. However, many models are not validated according for their ability to make accurate predictions of human performance. The predictive demand of these models is made increasingly complex by the needs of training domain, needing both to predict both skill decay and reacquisition from little historical data. In this paper, we examine the predictive capability of the Predictive Performance Equation (PPE) implemented in a Bayesian hierarchical model. Through a comparison of two Bayesian hierarchical models we show how hierarchical model fit to a participant's performance across a set of items compared to only a single item improves PPE's predictive accuracy of both skill decay and reacquisition over multiple learning schedules

Keywords: Mathematical Model, Bayesian Hierarchical Model, Prediction, Skill acquisition, skill decay, spacing effect, Learning management system

Introduction

The increase in availability of personal technologies such as mobile phones, computers, and laptops are becoming a ubiquitous part of everyday lives. The availability of personal technology gives individuals the opportunity for greater access to a variety of educational resources, such as learning management systems. Learning management systems (LMS) offer personalized education curriculums and training on a wide range of topics and a platform to record performance history. The flexibility and personalization that learning management systems afford makes their utility appealing to a variety of military, education, and medical applications for key reasons. First, LMS can be used to train an individual up to a particular performance standard. Second, LMS can schedule additional training events in way so that the individual performance stays at or exceeds a defined standard. Each of these goals can be achieved by applying robust empirically grounded findings from the psychology of human learning to the construction of training content and assessment.

Three general components of human learning have been identified across a diverse range of domains and time scales and are seen as critical to human learning in education. First, the power the law of learning (Newell & Rosenbloom, 1981) reveals that performance improves over the course of repeated trainings exposures. Second, the power law of decay reveals that performance decreases non-linearly as the time between exposures increases. Third the spacing effect reveals when the exposure to a task is distributed over time (i.e., spaced) as opposed to condensed within a short duration (i.e., mass practice) individuals retain the spaced information better than if under massed practice (Carpenter, Cepeda, Rohrer, Kang, Pashler, 2012).

Multiple mathematical models of learning and retention have been developed that attempt to account for the previously discussed learning phenomena. (Pavlik & Anderson 2005; Raaijmakers 2003; Walsh, Gluck, Gunzelmann, Jastrzembski 2018). For a formal model comparison across these three models of learning and retention see Walsh et al. (2018). In this paper, we focus solely on the Predictive Performance Equation (PPE) as it has been found to account for a variety of learning and retention phenomena and has shown potential for real world application.

Predictive Performance Equation

The Predictive Performance Equation (PPE) is a mathematical model of learning and retention that makes performance predictions at an individual-level based on prior performance and the temporal schedule training. In short, PPE is composed of five equations which represent the power law of learning, the power law of decay, and the spacing effect (for a detailed description see Walsh et al., 2018). The power law of learning (Eq.1, first term) is a function of *N,* the number of exposures to a task, and the *learning rate*, which is held constant at .1.

$$M = N^{learning\ rate} * T^{-decay\ rate}\ (Eq. 1)$$

The power law of decay (Eq.1, second term) is a function of *model time* (*T*) and the *decay rate*. *Model time T* Eq. 2 is weighted according to a factor of time (Eq. 3).

$$T = \sum_{i=1}^{n-1} w_i \cdot t_i \, , (Eq. 2)$$

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

The spacing effect is represented within the *decay rate* equation (Eq. 4), which includes two free parameters defined as the *decay intercept (b)* and the *decay slope (m)*.

$$decay\, rate = b + m * \; average\left(\frac{1}{log(lag_i)}\right) \text{(Eq.4)}$$

In effect, as practice events occur more tightly spaced in time (i.e., massed), the decay rate increases, and as practice events become more distributed in time (i.e., spaced), the decay rate decreases. Finally, *M (Eq.5)* is placed within a logistic function and adjusted according to the two final free parameter τ and *s*.

$$Performance = \frac{1}{1 + exp\left(\frac{\tau - M}{s}\right)} \; (Eq. 5)$$

Model Limitations

A benefit of using mathematical models of learning and retention is that they can be used to infer an individual's current state of knowledge on a particular task (Eq.5) and predict how their knowledge will change over time based on psychological principles of memory. However, the predictive accuracy of these models of psychological models are often not assessed (Yarkoni & Westfall, 2017). Instead prior research has focused on evaluating models of spacing based on their fit to empirical data sets. (Walsh et al. 2018, Pavlik & Anderson, 2005; Ragermaker 2003). Though validating models by their ad-hoc fit to empirical datasets is an important component of model development, it does not allow for an evaluation of a model's predictive ability or utility in an applied domain, which can be difficult. In an applied setting, a model needs to accurately predict skill decay between learning sessions, as well as skill reacquisition within a new session so that adequate training prescriptions can be made. Developing these types of predictions is made more difficult by the fact that often little historical data is often available to calibrate to. This is also often the case for machine learning models, which have been shown to reach a high degree of predictive accuracy in particular learning tasks (Settles, 2018).

One way that these predictive goals can be achieved in psychological models is through the use of a Hierarchical Bayesian Model (HBM) (McElreath 2018). Bayesian implementations of psychological models have a number of benefits that can be leveraged to improve a model's predictive accuracy (Lee & Wagenmakers, 2014). First, prior knowledge about the probable model parameters can be explicitly implemented into the model. Second, Bayesian methods allow for an integration of prior information with a set of observations into a posterior distribution, from which predictions can be made. Finally, hierarchical Bayesian models allow for dependencies across parameters to be specified, allowing free parameters to be estimated at different levels of aggregation. These multi-level parameter dependencies allow for greater constraint to be placed on a model, which guards against over fitting.

In this paper, we explicitly examine the predictive accuracy of the PPE for individuals learning Japanese-English word pairs under different within and between session spacing conditions. Additionally, we compare two different Bayesian implementations of the PPE. The first implementation is an Item model, where the model was calibrated separately to performance on *each* of the Japanese-English word pair learned by a subject during the experiment. The second implementation was a subject model, where the PPE was fit simultaneously to *all* Japanese-English word pairs learned by a subject during the experiment. In the simulations we examine the PPE's predictive ability across various period of time and the additional benefit of taking account all of the subject data.

Method

Participants

Sixty-one participants were recruited from a midwestern university in this paired-associate learning study. All participants completed a total of three experimental sessions spanning a three-week period.

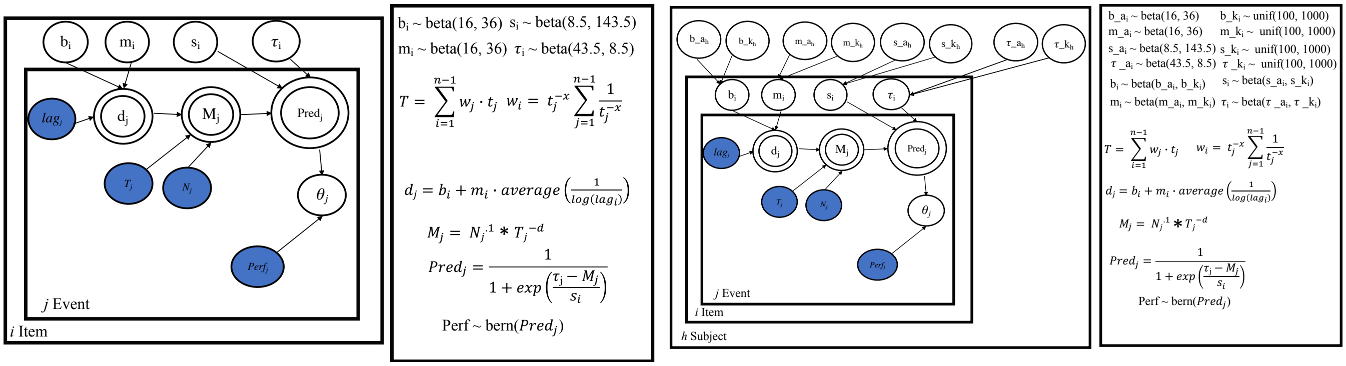
Task Stimuli

Over the course of the experiment participants memorized a set of 30 Japanese-English words. All of the words used in this study were taken from the Medical Research Council (MRC) Psycholinguistic Database manual and have been used in other previous memory studies (e.g., Pavlik & Anderson, 2005).

Experimental Design and Procedure

During the experiment, an item's training schedule was manipulated according to inter-session interval (ISI) and inter-trial interval (ITI) over the course of experimental sessions. The ISI controlled the amount of time between the 1st and 2nd experimental session. The ISIs in this study were fixed at short (5 min), medium (7 days), and long (14 days) delay. The ITI manipulated the number of trials between presentations of the same item. Two ITI consisting of a short (items repeated every 2 trials) and long (items repeated every 11 trials) delay were embedded in each experimental session.

During the study, participants, with no knowledge of the Japanese language, were given instructions for the paired associate learning task and had an opportunity to ask any questions. Once participants began the experiment, they were shown a Japanese word (e.g., "kanboku") on the screen and asked to type the English translation (e.g., "bush") to the Japanese word. Upon first presentation of a word participants were shown the English translation and asked to type the correct answer to ensure the item was studied. During all subsequent presentations, participants were asked to recall and type the English translation from memory. Participants



$b_i \sim \text{beta}(16, 36)$ $s_i \sim \text{beta}(8.5, 143.5)$

$m_i \sim \text{beta}(16, 36)$ $\tau_i \sim \text{beta}(43.5, 8.5)$

$$T = \sum_{i=1}^{n-1} w_j \cdot t_j \quad w_i = t_j^{-x} \sum_{j=1}^{n-1} \frac{1}{t_j^{-x}}$$

$$d_j = b_i + m_i \cdot average\left(\frac{1}{log(lag_i)}\right)$$

$$M_j = N_j^{.1} * T_j^{-d}$$

$$Pred_j = \frac{1}{1 + exp\left(\frac{\tau_j - M_j}{s_i}\right)}$$

$\text{Perf} \sim \text{bern}(Pred_j)$

$b\_a_i \sim \text{beta}(16, 36)$ $b\_k_i \sim \text{unif}(100, 1000)$

$m\_a_i \sim \text{beta}(16, 36)$ $m\_k_i \sim \text{unif}(100, 1000)$

$s\_a_i \sim \text{beta}(8.5, 143.5)$ $s\_k_i \sim \text{unif}(100, 1000)$

$\tau\_a_i \sim \text{beta}(43.5, 8.5)$ $\tau\_k_i \sim \text{unif}(100, 1000)$

$b_i \sim \text{beta}(b\_a_i, b\_k_i)$ $s_i \sim \text{beta}(s\_a_i, s\_k_i)$

$m_i \sim \text{beta}(m\_a_i, m\_k_i)$ $\tau_i \sim \text{beta}(\tau\_a_i, \tau\_k_i)$

$$T = \sum_{i=1}^{n-1} w_j \cdot t_j \quad w_i = t_j^{-x} \sum_{j=1}^{n-1} \frac{1}{t_j^{-x}}$$

$$d_j = b_i + m_i \cdot average\left(\frac{1}{log(lag_i)}\right)$$

$$M_j = N_j^{.1} * T_j^{-d}$$

$$Pred_j = \frac{1}{1 + exp\left(\frac{\tau_j - M_j}{s_i}\right)}$$

$\text{Perf} \sim \text{bern}(Pred_j)$

Figure 1. The two graphical representations of the item (left plot) and subject model (right plot).

were given a maximum of 7 seconds to type their answer during each trial. If a participant could not generate a response within 7 seconds, then their answer was considered incorrect. At the end of each trial participants were given feedback (correct or incorrect) and given 2 seconds to study the correct answer.

Hierarchical Bayesian Implementation of PPE

To examine the PPE's predictive ability, two Bayesian hierarchical models of PPE were developed – Item and Subject model. Each model made predictions of the participants' performance on the Japanese - English word pairs over the course of the experiment (Figure 1). Each model was represented as a graphical model (Lee & Wagenmakers, 2011) to allow each variable, variable type and dependencies across variables to be observed. All observed variables (participants' performance - Perf, time variables - lag$_i$, *T*, *N*) are represented as shaded circles. Estimated parameters (*b, m, a, tau*) are represented as unshaded circles. Stochastic variables are represented by a single open circle, while deterministic variables are represented by two open circles. The multiple panes within each figure represent redundancies within the model for each participant, items (Japanese-English word pair), and trials

Both of the Item and Subject models share a similar structure, differing only in the constraints placed on each of the parameter estimates for each item. The Item model estimates PPE's free parameters (b_i, m_i, a_i, tau_i) individually for each of the Japanese-English word pairs attempted by an individual. PPE's free parameters sampled from these distributions are then combined together with the individual's unique time variables (lag_i, T_j, N_j) and PPE's equations to calculate activation (M$_j$) for a particular trial. The estimated activation for a particular trial is then transformed into PPE's logistic distribution to estimate a participant's performance ability on a particular item for a particular trial. PPE's performance estimate is then placed within a likelihood function a Bernoulli distribution (θ), which is then compared to the subject's performance (i.e., correct or incorrect) during a given trial.

The Subject model uses a similar structure to the individual-item level model, with one key difference. Both models estimate individual parameters for each subject on a particular item (b_i, m_i, a_i, tau_i,), but unlike the individual Item model, the Subject model constrains parameters at participant level while generating estimates for each item. Thus, the Subject model estimates the mean (b_a_h, m_a_h, a_a_h, tau_a_i) and degree of certainty (b_k_h, m_k_h, a_k_h, $\tau\ \_k_h$) for each of the free parameters for each subject's performance across all of the learned items. The additional parameters estimated for each participant across all of the Japanese-English word pairs, allows for the parameter estimates of a particular item to be constrained by the subject's performance of all other items.

Model Fitting Procedure In these simulations we were interested in examining PPE's predictive ability across periods of time consistent with real -would applications. For this reason, both the Subject and Item models were independently fit and used to predict performance during two segmented portions of the experiment. The first fit and prediction period was over the 1st and 2nd experimental session. Each model calibrated to participants' performance during the 1st session (10 trials), and used the resulting parameters to generate predictions of the participants' performance during the 2nd session. During the second fit and prediction period both models were calibrated to the performance during the 1st and 2nd experimental session to predict performance during the 3rd session. When calibrating to the participants' performance, the Item model was applied separately to the participants' performance on each Japanese-English word pairs while the Subject model was calibrated simultaneously to all Japanese- English word pair. Though both models were calibrated to different portions of a subject's performance, a single item (Item model) or all items (Subject model), both models used PPE's estimated free parameters (b$_i$, m$_i$, s$_i$, and τ_i) during each calibration section to make predictions for each future repetition of each item for all participants.

Results

The two models' fits and predictive accuracy were evaluated using three formal evaluation metrics. First, we assessed the correlation and root-mean squared error (RMSD) between the participants' average performance and each model's average calibrations and out of sample prediction for each of the 6 unique ISI and ITI conditions. An evaluation of the

Table 1. The Subject and Item Model's correlation(*r*), *RMSD*, and *AUC* when calibrated to the 1st session predicting the 2nd session across 6 unique learning schedules.

Spacing Condion		Subject					Item				
ISI	ITI	r 1st session	RMSD 1st session	r 2nd Session	RMSD 2nd Session	AUC 2nd Session	r 1st Session	RMSD 1st Session	r 2nd session	RMSD 2nd Session	AUC 2nd Session
5 min	Short	0.99	0.04	0.93	0.10	0.90	0.99	0.12	0.91	0.11	0.80
5 min	Long	0.99	0.07	0.72	0.02	0.89	0.96	0.13	0.71	0.17	0.79
7 days	Short	0.99	0.02	0.22	0.06	0.95	0.99	0.11	0.99	0.25	0.88
7 days	Long	0.99	0.08	0.99	0.05	0.94	0.98	0.12	0.99	0.24	0.84
14 days	Short	0.99	0.01	0.99	0.06	0.96	0.99	0.12	0.99	0.25	0.90
14 days	Long	0.98	0.08	0.99	0.05	0.93	0.96	0.13	0.99	0.22	0.87

participants' overall average performance and each model's prediction allows for an overall evaluation of how well each model could account for the participants' performance. In addition, we used the area under the curve (AUC) metric. AUC can be interpreted as the probability that the model will rank a randomly sampled item that is correct higher than a randomly chosen item that is incorrect. Each of these three metrics was computed individually for within sample performance (calibration) and out of sample (predicted) performance as well as individually for each experimental condition (see Table 1 & 2). Evaluations of each of these measures allows for an examination of the characteristics of each model is well-suited to explain and predict.

Predicting 2nd Session

Fit to Calibration Phase Both the Item and Subject models were found to fit the average performance of the participants during the first session fairly well (Table 1) and possessed high correlations to performance within the first session. However, evaluation using RMSD revealed a different picture. For each of the experimental conditions, the Item model was found to have a higher RMSD compared to the Subject model. As shown in Figure 2, when models are calibrated to the participants' performance during the 1st session (Trials 1 – 10), the Item model estimates the participants' performance to be lower than their average performance. The Item model's under estimation of performance was a result of its greater uncertainty in the participants' performance, as seen in the model's 95% HDI. The Subject model by comparison, not only obtained average fit to the participants' average performance more accurately but did so with greater certainty, as indicated by the tighter 95% HDI. These differences in each model's ability to calibrate to the participants' performance translated into differences in each model's predictive ability.

Out of sample performance Both models revealed high correlations for participants' average performance and each model's average predictions during the 2nd session across the three different ISI (5 min, 7 days, and 14 days). However, the difference in the RMSD between Item and Subject model during the calibration persisted. Again, the difference in the RMSD between the participants' performance during the 2nd session and each of the model's predictions was a result of each model's uncertainty in the participants' ability. The uncertainty in predictions of the Item model was found to increases over time, relative to the Subject model (Figure 2). The higher uncertainty in the Item model's predictions decreased in the Item model's ability to predict the subjects' performance during the 2nd session.

Finally, differences in the *AUC* between the predictions made by both models and the participants' performance on each individual item was found. Both models had a high *AUC* scores across each of the 6 spacing conditions (Table 1). The high AUC across all conditions revealed that despite the difference in average performance predictions between the Subject and Item models, both models were able to account for the *relative* performance of participants on the individual Japanese-English word pairs during the second session. Nonetheless, *AUC* for the Subject model was higher compared to the Item model across all conditions, highlighting the better predictive ability of the Subject model. Taken together, these results show that the Subject model made more precise predictions—both relative and absolute—compared to the Item model.

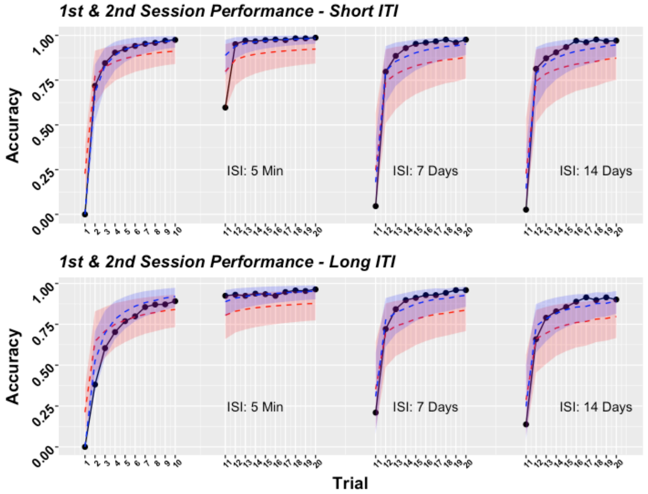


Figure 2. The average accuracy of participant's recall of Japanese - English word pairs (connected black points) on the short (top panel) and long (7 bottom panel) Inter-Trial-Interval (ITI) and three unique Inter Session Interval (ISI - 5 minutes, 7 days, and 14 days) schedules. The subject (dashed blue line) and item (dashed red line) mean and 95% Highest Density Interval (HDI) of each model's calibration (Trials 1-10) to performance during the 1st session and predictions of the 2nd session (Trials 11-20).

Table 2. The Subject and Item Models (*r*), *RMSD*, and *AUC* when calibrated to the 1st and 2nd session predicting the 3rd session across 6 unique learning schedules.

Spacing Condition		**Subject**					**Item**				
ISI	*ITI*	*r 1st & 2nd session*	*1st & 2nd session RMSD*	*r 3rd Session*	*RMSD 3rd Session*	*AUC 3rd Session*	*1st & 2nd Session r*	*1st & 2nd Session RMSD*	*r 3rd Session*	*RMSD 3rd Session*	*AUC 3rd Session*
5 min	Short	0.96	0.06	1.00	0.03	0.95	0.97	0.06	1.00	0.24	0.80
5 min	Long	0.97	0.08	0.98	0.12	0.94	0.95	0.09	0.98	0.34	0.76
7 days	Short	0.99	0.05	0.99	0.06	0.93	0.98	0.06	0.99	0.13	0.82
7 days	Long	0.98	0.05	0.99	0.11	0.92	0.93	0.09	0.99	0.22	0.85
14 days	Short	0.99	0.04	1.00	0.11	0.91	0.98	0.06	1.00	0.08	0.81
14 days	Long	0.99	0.05	0.99	0.12	0.90	0.95	0.09	0.99	0.22	0.83

Predicting 3rd Session

Fit to Calibration Phase For the second prediction, both the Subject and Item models were calibrated the participant's performance on each item during the first two sessions. As seen in Table 2, both models outperformed the fits observed when calibrating to the 1st session alone (cf. Table 1). Unsurprisingly, the additional data over the course of two sessions allowed for better performance tracking for individual participants (Figure 3).

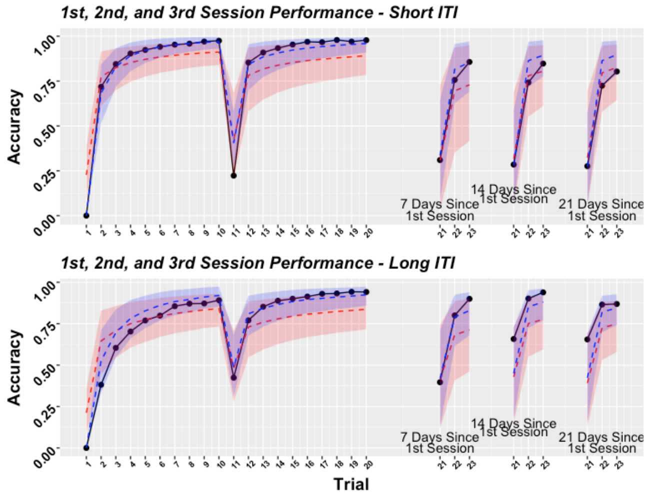


Figure 3. The average accuracy on the Japanese - English word pairs given on the short (top panel) and long (7 bottom panel) Inter-Trial-Interval (ITI) and three Inter Session Interval (ISI - 1 hour, 7 Days, and 14 days) schedules. The subject (blue line) and item (red line) mean and 95% Highest Density Interval (HDI) of each model's calibration (Trials 1-30) to performance during the 1st & 2n session and predictions of the 3rd session (Trials 21-23).

Out of sample prediction Each of the model's predictions of the participants' performance within the 3rd session revealed a similar pattern. High correlations were found between the participants' average performance and each of the model's average predictions (Table 2). As observed during the calibration period, models differed slightly in the *RMSD* between their predictions and participants' performance. Both models over predicted initial performance for items in two of the three ISI levels on the long ITI manipulations, but the Item model under predicted skill acquisition. In contrast, both models predicted initial skill decay on the 3rd session across all ISI levels in the short ITI manipulation, but the Subject model was found to slightly over predict skill acquisition. These differences decreased the differences between both model's predictions of the 3rd session.

Finally, despite the differences between models both models are found to have a fairly high *AUC* when predicting the subjects' performance on the 3rd session. However, again across each of the six experimental conditions, the *AUC* of the predictions from the Subject model are found to be higher relative to the Item model. Taken together, evaluation of the predictions for the 3rd session confirm and reinforce the results of the predictions for the 2nd session.

Individual Item Predictions

Up to this point, each model's ability to predict the participants' performance over time has been evaluated. However, each model's ability to account for individual differences has not been measured. To address this question, the average performance and each model's average prediction ±95% HDI for a given subject during each calibration and prediction event was calculated (Figure 4). Across all participants, both the Item and Subject model were able to match the participants' average performance. However, a large difference in the average 95%HDI was found between the Item and Subject model. Across all participants, the Item model was less certain of each participant's average performance compared to the Subject model over the set of Japanese-English word pairs. These differences in the models' certainty in the participants' average performance led to differences in the accuracy of its predictions. When predicting performance during both the 2nd and 3rd session, the Subject model was able to accurately predict the subjects' average performance, with a majority of the participant's performance falling within the predicted 95% HDI. The same level of accuracy was not observed for the Item model's predictions, showing consistent underprediction of the participants' performance.

Discussion

In this paper, we explored the predictive capability of the PPE, which holds potential for real world application. However, often models of learning and retention are often not

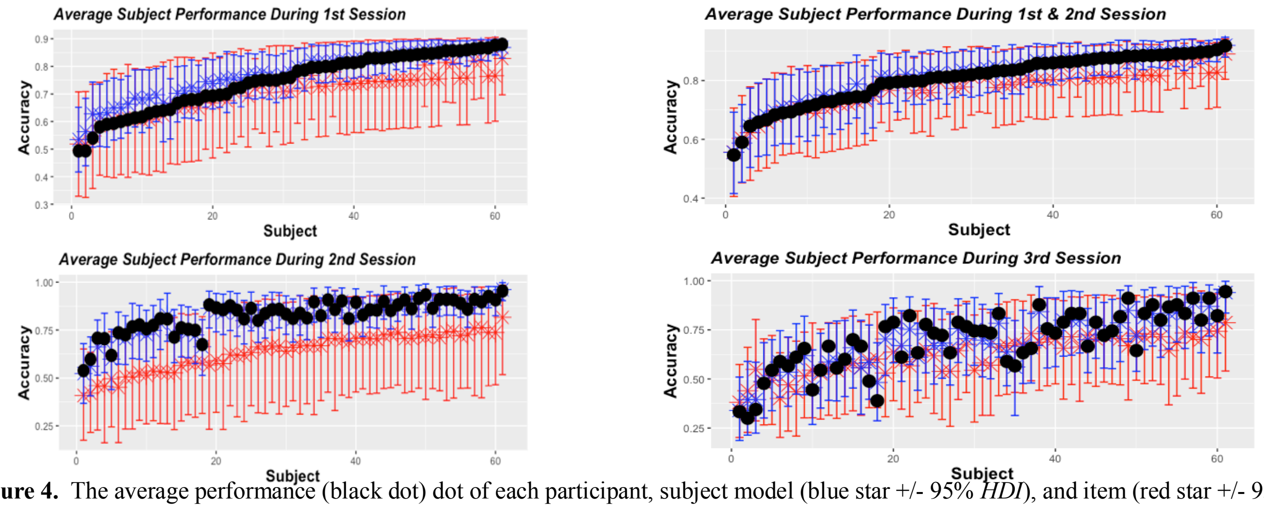


Figure 4. The average performance (black dot) dot of each participant, subject model (blue star +/- 95% *HDI*), and item (red star +/- 95% *HDI*) of the two calibration (Calibrate 1st session Predict 2nd session – Left panels, Calibrate 1st and 2nd session Predict 3rd session – right panels) and prediction simulations

If the PPE are used to be used in in real-world applications, then it must be able to make use of a limited amount of performance data from an individual to extrapolate its estimates of performance into the future. In this paper, we explored using Bayesian hierarchical modeling to simultaneously fit all of a participant's data compared to the Item model where the standard PPE was fit only to the performance of a single word pair. Overall our results showed that the PPE captured the relative trends observed in the human data across a variety of different spacing manipulations as measured by the AUC metrics. These results show that both the Item and Subject model predicted performance under conditions where little information was available. Differences in the average predictions of each model stem from the confidence in each models' predictions. Due to the fact that the item model only calibrated to the performance of a single item during the 1st session (10 trials) or 1st and 2nd session (20 trials), it had a higher degree of uncertainty in its estimate of the individuals knowledge of a given word pair compared to the Subject model that simultaneously fit the performance of all word pairs learned by a participant. While the Subject model calibrated to all of a participant's performance allowed for more certain performance estaminets and accurate predictions.

Given the results presented in this paper, two avenues of future research are seen. First, we evaluated the predictive accuracy of the PPE under laboratory conditions, where the training schedule was tightly controlled. Future research should explore applying these methods to real world data. Second, exploration in PPE's ability to account for performance at an item level over time should be explored. In this paper, we explored predictive accuracy of each model at an aggregate level. Human performance at an aggregate level often follows a power law, but individual item level performance is found to follow an exponential function. An evaluation of PPE's predictive accuracy at this lower level of analysis, might warrant further model modification.

In conclusion, the results presented in this paper show that PPE is capable of making valid performance predictions over various periods of time. Furthermore, in cases where little data is available, the model's predictions can be improved by using a hierarchical modeling approach, conditioning estaminets of performance of a single item based on a set of multiple items, allowing for more precise estimates of performance. By utilizing these statistical approaches, psychological models of learning and retention can better predict both skill decay and requisition, allowing psychological models to meet the needs of real-world application.

References

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

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.

Lee, M. D., and Wagenmakers, E. J. (2014). Bayesian Cognitive Modeling: A Practical Course. Cambridge: Cambridge University Press.

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

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018, June). Second language acquisition modeling. *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 56-65).

McElreath, R. (2018). Statistical rethinking: a Bayesian course with examples in R and Stan. Place of publication not identified: Chapman and Hall/CRC.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive science*, *42*, 644-691.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122