Detecting Learning Phases to Improve Performance Prediction

Michael G Collins[1,2] (michael.collins.ctr@us.af.mil)
Caitlin Tenison[3] (ctenison@ets.org)
Kevin A Gluck[1] (kevin.gluck@us.af.mil)
John Anderson[4] (ja0s@andrew.cmu.edu)

[1]Air Force Research Laboratory, [2]ORISE at AFRL, [3]ETS, [4]Carnegie Mellon University

Abstract

Models of learning and retention make predictions of human performance based on the interaction of cognitive mechanisms with temporal features such as the number of repetitions, time since last presentation, and item spacing. These features have been shown to consistently influence performance across a variety of domains. Typically omitted from these accounts are the changes in cognitive process and key mechanisms used by people while acquiring a skill. Here we integrate a model of skill acquisition (Tenison & Anderson, 2016) with the Predictive Performance Equation (PPE; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2019) using Bayesian change detection (Lee, 2019). Our results show this allows for both better representation of an individual's performance during training and improved out-of-sample prediction.

Keywords: Mathematical model, Bayesian model, Hidden Markov model, Skill acquisition, Learning, Strategy change, Spacing effect, Change detection, Prediction

Since the research of Ebbinghaus over a century ago, psychologists have studied human learning and forgetting. This research has resulted in three core empirical phenomena. First, the law of practice (Newell & Rosenbloom, 1982). With sufficiently frequent practice on a task, performance improves over time. Second, the law of decay. As the time between instances of exposure increases, individual memory starts to decay and performance gets worse. Third, the spacing effect. When exposures between learning opportunities are distributed over time (spaced practice) individuals are slower to acquire the information, but retain the information better than if given the same number of exposures within a short time (massed practice; Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012). To explain and predict these factors of human learning, formal models of learning and retention have been developed (see Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018, for a detailed model review and comparison). In the current paper, we focus on the Predictive Performance Equation (PPE), a model of learning and retention (Walsh et al., 2018).

Predictive Performance Equation (PPE)

PPE is a mathematical model of learning and retention that makes performance predictions at the individual level based on (1) prior performance and (2) the learning schedule of an individual. PPE is composed of five equations representing three psychological factors of learning (power law of learning, power law of decay, and spacing). The first factor is the power law of learning (Eq.1, first term), which is a function of *N*, the number of exposures to a task, *a* which represents one's prior task knowledge, and learning rate, *c*, which is held constant.

$$M = (N + a)^c * T^{-d} \ (Eq. 1)$$

The second factor is the power law of decay (Eq.1, second term). Temporal decay is represented using *T* (Eq. 2), which weights time by (Eq. 3), and exponent decay parameter, *d*.

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

$$w_i = t_i^{-x} \sum_{j=1}^{n-1} \frac{1}{t_j^{-.75}} (Eq. 3)$$

The third factor is spacing effect or temporal distribution of practice over time, which is represented within the decay parameter (Eq. 4).

$$d = b + m \cdot average\left(\frac{1}{log(lag_i)}\right) \ (Eq. 4)$$

Spacing is accounted for using two free parameters *b* and *m* and cumulative average lag time. As practice events occur together (i.e., massed), the decay increases. As practice becomes distributed (i.e., spaced), the decay decreases. Finally, activation (M) is placed within a logistic function and adjusted according to a *threshold* parameter, τ.

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

Model Limitations

A limitation of PPE is that it assumes an individual's performance is expected to improve or decrease according to a single continuous function over time. This is based on empirical findings of aggregate performance curves, which often reveal smooth performance curves following power laws. Research examining learning in individuals find this an artifact of averaging the performance of multiple individuals (Heathcote, Brown, & Mewhort, 2000; Gray & Lindstedt, 2017). An individual's performance can appear to have "sporadic" performance variation. These "sporadic" performance variations sometimes appear random but have been shown to reflect changes in an individual's strategy (Gray & Lindstedt, 2017; Tenison & Anderson, 2016)

These findings are consistent with research suggesting skill acquisition occurs in phases, where an individual uses

different strategies and/or representations to complete a task (Fitts & Posner, 1967; Tenison, Fincham, & Anderson, 2016). One such theory of skill acquisition, based on the ACT-R architecture, proposes that individuals go through three phases while acquiring a skill (Anderson, 1982). In the first phase, the *Computational phase*, the individual solves a problem by applying general problem-solving rules to achieve the solution. In the second, *Associative phase*, the problem is solved through the direct retrieval of various portions of the problem. In the third phase of learning, the *Autonomous phase*, the individual has created a stimulus-response rule for a given problem. Learning during these phases is driven by knowledge compilation and declarative strengthening. It can be modeled by three separate power law functions and their associated parameters (i.e, intercept, slope, asymptote). Prior work has used hidden Markov models (HMMs) to estimate these parameters and identify where phase shifts occur (Tenison & Anderson, 2016).

Currently, PPE does not represent or make predictions based on these learning phases. Prior research with PPE focused on accounting for average performance on memory retrieval during word association tasks and overall performance metrics of complex tasks (Gluck, Collins, Krusmark, Sense, Maaß, & van Rijn, 2019; Jastrzembski, Gluck, & Gunzelmann, 2006). In each of these cases, the assumption was a continuous performance curve moderated by features of the learning schedule (i.e., number of attempts, time between trials, spacing). Assuming a continuous learning curve is reasonable in these cases because average performance will follow standard features of learning. However, the argument for this assumption of a continuous performance curve weakens when accounting for performance at lower levels of aggregation. At the individual person learning individual skills level of analysis, there are often discontinuities in performance. PPE interprets these "sporadic" changes as noise, leading to (1) decreased confidence in an individual's performance, (2) a less accurate representation of the learning profile, and/or (3) unrealistic out-of-sample performance predictions. To mitigate these limitations a substantive mechanistic means of interpreting discontinuities in individual learning profiles is required.

In this paper, we propose and evaluate a theoretical and methodological integration to achieve just that. We use a Bayesian change detection procedure (Lee, 2019; Lee, Gluck, & Walsh, 2019) to identify when identifiable performance discontinuities occur. We then use information about these change points to infer changes in phase during learning and make predictions about subsequent performance. We refer to this novel implementation as TAPPED, which is an integration of Tenison and Anderson's (2016) skill acquisition model, Walsh et al.'s (2018) PPE, and Lee's (2018) change detection procedure. To foreshadow, the TAPPED model identifies similar learning phases similar to Tenison and Anderson's (2016) HMM-based models of skill acquisition and is found to have superior predictive accuracy compared to PPE.

Method

Participants

We used Amazon Mechanical Turk to recruit 101 participants (Gender: Female = 49, Age: M = 31.4, SD = 6). All participants were paid $10 for participation in both experimental sessions and $.02 for each correct problem.

Task Stimuli

During the experiment participants completed a set of novel mathematics problems (Pyramid problems). Each problem is composed of two numbers separated by a "$" symbol (e.g., 2$4). The first number is referred to as a base. The second number is referred to as the height. The base of the problem represents the first term in the additive sequence. The height of the problem represents the number of sequential numbers that must be added to the base. For example, if a participant was given the problem 3$3, then they would have to add together the number 3 + 4 + 5 + 6 = 18. In this experiment participants were given problems with bases ranging from 3 – 6 and heights from 4 – 11.

Procedure

The experiment consisted of two experimental sessions, with a 66 hour lag in between them. During the first day, problems were displayed on the screen and participants were instructed to type in their answer. All participants received feedback on whether they were correct or incorrect. Participants went through 10 practice blocks, with each block including 40 items each. Each block consisted of 40 items, with each item in one of four spacing conditions. Items in Spacing group 4 were presented 25 times with 3 problems in-between. Spacing group 8 were presented 25 times with 7 intervening questions. Spacing group 16 were presented 25 times with 15 intervening problems. Spacing group 32, were presented 12 times with 31 intervening presentations. The second experimental session was given 66 hours after the first session and were tested on the items they practiced on Day 1.

Hidden Markov Model

We fit an adaptation of the Tenison and Anderson (2016) power-law skill acquisition model to the response latencies for the items solved during the 10 practice blocks completed on Day 1. We refer to this model as the *Phase HMM.* Only an overview of the HMM model is provided here, a detailed description can be found in Tenison and Anderson (2016). The HMM consists of three states, each representing the three phases of skill acquisition. Within each phase, we have a state representing each practice opportunity the participant may have had within that phase. After each stimulus, participants either transition to the first state of the next phase or the next state within the current phase. The HMM predicts that participants' reaction times follow phase-specific power functions, where the opportunity count for each power function is determined by the current state within that phase. This HMM structure enables joint estimation of a participant's state as well as their response latency.

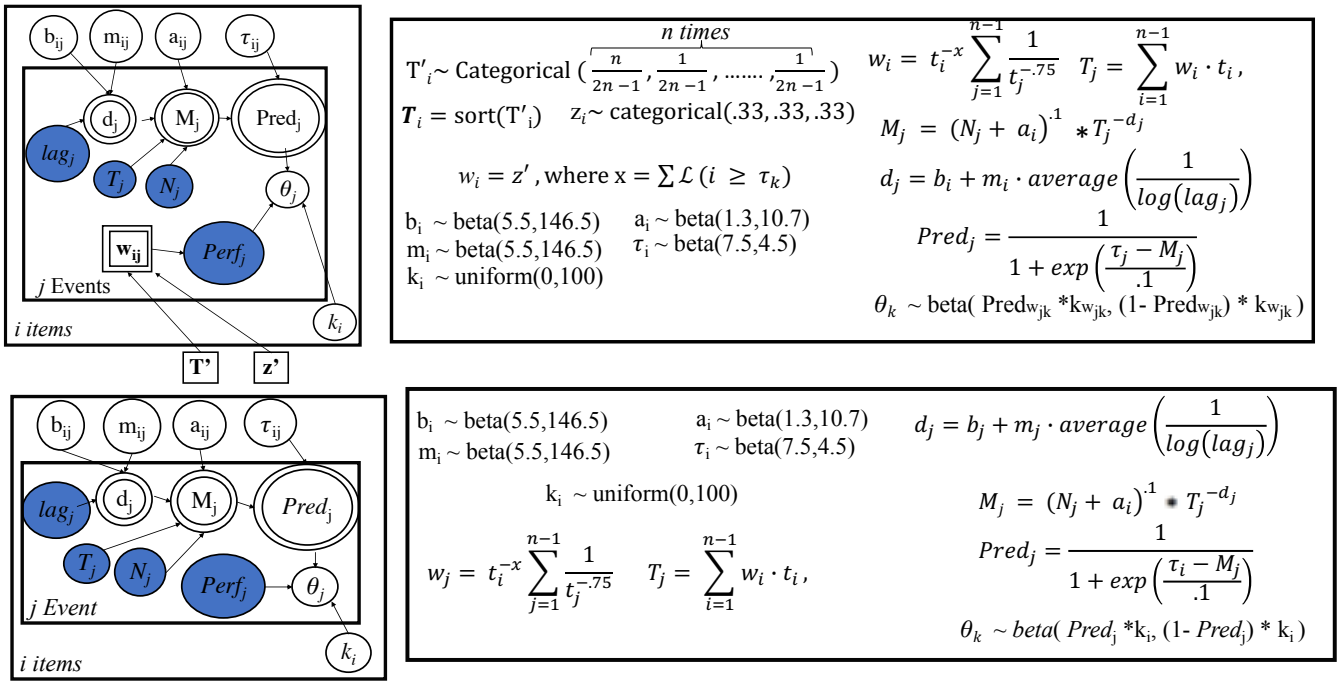

Figure 1. The Bayesian model diagrams for TAPPED (top) and PPE (bottom).

In fitting the Tenison and Anderson model (2016) to this current experiment we made two changes. First, we adjusted the model to account for general learning which occurs over the course of the task and is therefore shared across spacing groups (Eq 6). To do this we extended the skill specific model to capture general learning using the general learning equation derived from ACT* (Pirolli and Anderson, 1985). Second, we expanded our model fitting procedure to identify which parameters should be shared between items of different spacing groups. We fit eight models to the data, exploring whether sharing learning rate (α), transition probability (π_{12} π_{23}), and scale (β_{phase}) across spacing groups improved the fit of the model.

$$\mu_{ret} = \beta_{phase} {n_{specific}}^{-a_{specific}} \times {n_{general}}^{-a_{general}} \; (Eq.6)$$

Table 1. Parameters for best fitting HMM

Spacing group	$\beta_{computation}$	$\beta_{associative}$	$\beta_{autonomous}$
4	9.97	3.6	1.8
8	11.5	4.6	2.0
16	12.0	5.1	2.1
32	12.0	4.6	1.9
Parameters shared across spacing group			
$a_{gen.}$	$a_{spec.}$	π_{12}	π_{23}
-.05	-.07	.12	.09

All models were compared using *BIC*. We found that the model that estimated unique scale parameters for each spacing group shared transition parameters and a shared learning rates fit best (Table 1; *BIC* = 119,517.2). This model fit the data better than more complex models with all unique parameters (*BIC*= 119,530.7) and simpler models with all shared parameters (*BIC* = 119,672.5). We list the parameters of our final model in Table 1. For each practice opportunity of each item, the model generates a likelihood of being in each state of the HMM. This can be translated into discrete Phase labels. We use these labels in our comparison with the TAPPED model.

Bayesian Models

The Bayesian implementations of TAPPED and PPE are each represented as a graphical model (Koller, Friedman, Getoor, & Taskar, 2007) (Figure 1). A graphical model format allows each variable, variable type, and the dependencies across variables to be observed. All observable variables (e.g., participant's response on a given trial – RT) are represented in shaded circles. PPE's estimated free parameters (*b*, *m*, *a*, τ) are represented as unshaded circles. Stochastic variables are represented with a single open circle, while deterministic variables are represented with two circles. The multiple panes represent redundancies for the different participants (i), and events (j).

The PPE (bottom graph) and TAPPED (top graph) (Figure 1) models are similar in their overall structure. Both models are run over a participant's (*i*) performance on a particular pyramid problem (item – *j*). For each participant the model estimates values for each of PPE's free parameter values (*b*, *m*, *a*, τ). The difference is in the number of different free parameters each model uses to account for participant performance during the 1st day. The PPE model (Figure 1 –

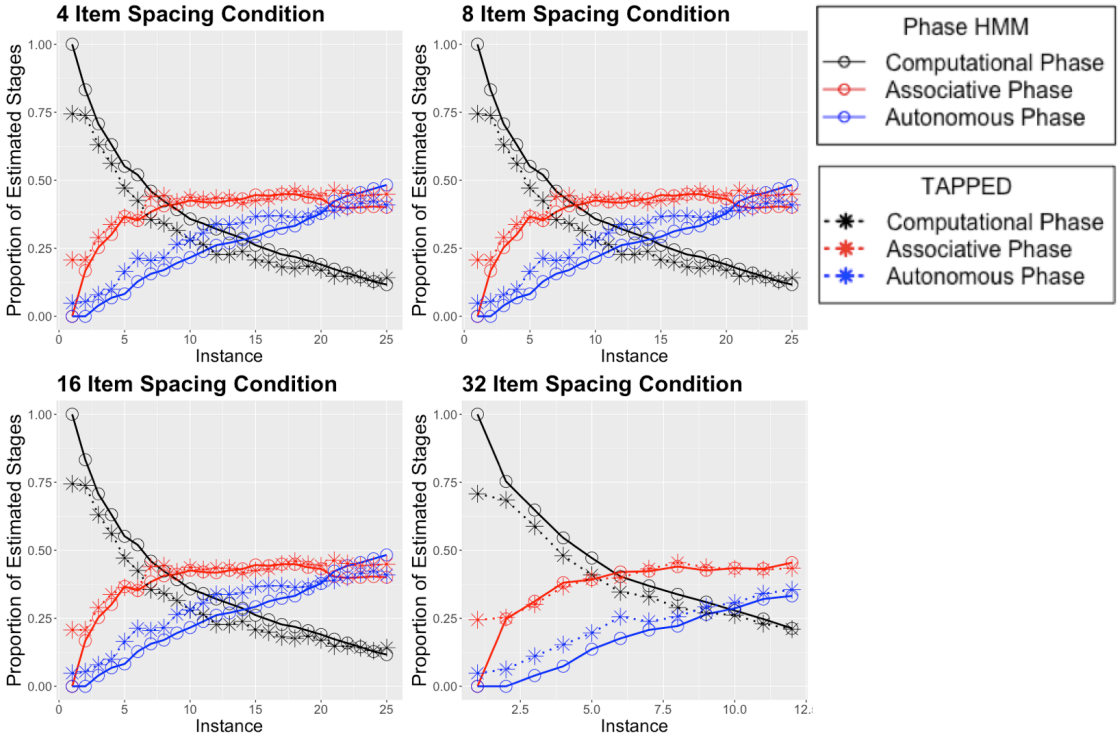

Figure 2. The round by round proportion of each phase of learning (Computational — black, Associative — red, and Procedural — blue) estimated with the Phase HMM (Open circles) and the TAPPED model (stars).

bottom figure) estimates only one value of each of PPE's free parameters for the participant's performance on a given item during the 1st day. However, the TAPPED model (Figure 2) can estimate up to 5 different values for each of PPE's free parameters, depending on the number of change points estimated by the model (*T*). Change points are estimated using a spike-and-slab prior (*T*) (Lee, 2019). After parameters are estimated for a portion of the learning curve (w_{ij}) from the prior distribution, they are combined with the participants' observed time variables (lag_j, *T*, N_j), and PPEs fixed equations
to create a performance estimate ($Pred_j$) for a given trial. The performance estimate ($Pred_j$) is then combined with a
precision parameter (k_i) in a beta distribution to develop a prior distribution for the likelihood function. Because PPE is developed to predict accuracy, reaction times (*Perf*) were transformed to a proportion by dividing their reaction time by a maximum reaction time of 7 minutes. The opposite transformation was used to get PPE's predictions represent reaction time for all of the preceding results by multiplying PPE's predictions by 7 minutes.

Additionally, during the student's 1st day of performance, the TAPPED model also inferred the phase (i.e., computational, associative, procedural) of learning that each individual was in before and after each unique change point (*z*). Each of the three learning phases (declarative, associative, and procedural), was identified based on response times estimated from the Tenison, Fincham, and Anderson (2016) empirical data distribution. The estimations of the individual's phase were then used to make predictions of the participant's subsequent performance. Predictions the participants performance of the 2nd experimental session were generated by using parameters estimated for the highest learning phase obtained during Day 1 performance.

Results

First we will examine the similarity between the inferences of the participants' phases of learning estimated by Tenison and Anderson (2016) (Phase HMM) and our TAPPED model. A comparison between the two models' results allows us to evaluate the extent to which they reach converging inferences about learning phase. Second we will use the estimation of phases from the participants' performance during the 1st day to make theory-driven predictions of the 3rd day and compare these to the PPE.

Behavioral Effects

For this analysis, we only considered spacing groups with the same number of total practice opportunities (i.e. SG-4, SG-8 and SG-16). Rather than rely on accuracy, which is high across all spacing groups, we used response latency to verify the impact of spacing on performance. A repeated measures analysis of variance (ANOVA) run on mean latency data (log transformed) revealed a significant main effect of spacing group ($F(2,174)=97.6$, $p < .001$) and practice opportunity ($F(1,87)=684.7$, $p < .001$). The interaction between spacing group and practice was not significant ($F(2,174)=1.7$, $p = .2$). During the initial practice period, items that experience

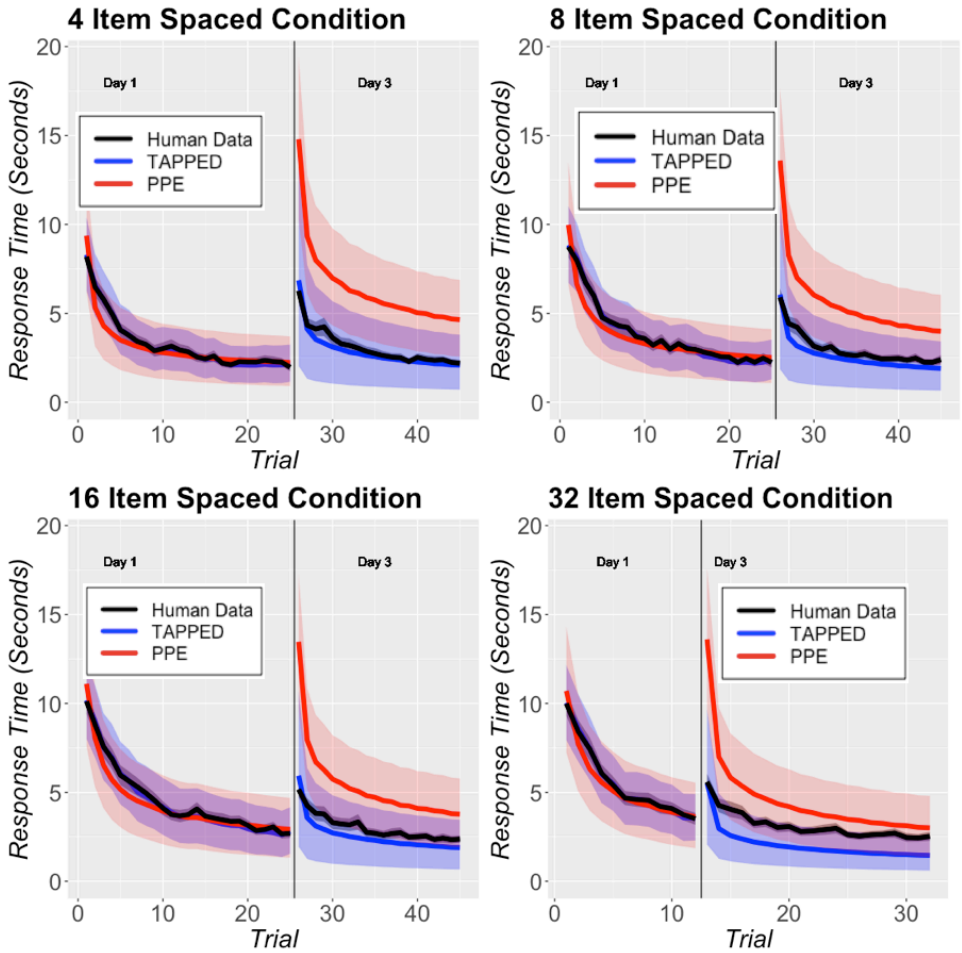

Figure 3. The mean performance of participants +/- 95% CI (black line and shaded ribbon) and the TAPPED +/- 95% HDI and PPE +/- 95% HDI calibration and prediction for the 4 spacing conditions.

An greater delay between practice opportunities take longer to solve than those with less delay between practice opportunities. We ran a repeated measures ANOVA to explore the impact of spacing on the response latency three days after the initial learning period. For the first opportunity on Day 3, we see a significant main effect of training ($F(2,174)=8.8, p<.001$) on problem solving latency (log transformed). The benefit of spaced practice is present in the faster response latency for spaced items, SG-4(M = 6.2s, SD = .3), SG-8 (M = 5.9s, SD = .3), SG-16 (M = 5.1s, SD = .3). These analyses confirm the presence of the spacing effect within our experiment.

Phase Comparison

To compare phases of learning inferred by the Phase HMM and the TAPPED model, the per round proportions of each phase (i.e., computational, associative, procedural) during the course of the first day in each of the 4 spacing conditions were compared (Figure 2). Across the four spacing conditions, a high degree of similarity is seen in the phases of learning estimated by the Phase HMM model and the TAPPED model (r = .87, RMSD = .10).

The greatest divergence between the two models is seen during the initial performance events (Rounds 1-3). Phase HMM model assumes that participants must sequentially go through all three phases of learning starting with the computational phase. This is why that model interprets 100% of participants as starting in the Computational phase. TAPPED does not share these assumptions, allowing for any phase of learning to be estimated at any point in time during the experiment for a given change point. Despite these differences, the TAPPED model converges to inferences similar to those of Phase HMM's model. This is evidence that the Bayesian change detection method of Lee (2019) is functionally approximating the output of the HMM used by the Phase HMM.

Day 1 – Calibration

To compare how well each model calibrated to each of the participant's performance during the 1st day, the PPE and TAPPED model are compared across the four spacing conditions (Figure 3) evaluation of each model's fit reveals two findings. First, the average performance estimate of both the PPE and TAPPED model fit participant average

performance quite well during the 1st day across each of the 4 spacing conditions.

Table 2. The correlation (*r*) and root mean squared deviation (RMSD) and the percent the participant's performance that falls within the predicted 95% HDI (% Pred) of the fits of TAPPED and PPE during the first experimental session.

	TAPPED			PPE		
Spacing	*r*	*RMSD*	*% Pred*	*r*	*RMSD*	*% Pred*
4	.93	1.32	93%	.76	2.14	84%
8	.94	1.46	92%	.77	2.69	82%
16	.93	1.68	90%	.77	2.87	77%
32	.98	1.05	96%	.81	2.99	77%

Although differences are observed in model fits of participants' individual learning curves, across the four spacing conditions, TAPPED has a higher correlation and lower RMSD with the individual participants compared to PPE (Table 2). TAPPED calibrates much more closely to individual performance during the 1st day. These better performance fits result from the fact that TAPPED selectively calibrates to separate portions of the participant's learning profile. While in contrast, PPE attempts to account for all the participant's first day performance with a single performance curve.

Day 3 – Predictions

Based on each model's calibration to the performance of participants during the 1st day, performance predictions were generated for each individual participant on the 3rd day (Figure 3). Over this period of time, the uncertainty each model has in the individual's performance increased. This uncertainty is reflected in the increase in the size of each model's 95% HDI. Though the uncertainty in predictions increases, there are differences between the two models' predictions. Over each of the 4 spacing conditions, PPE predicts slower initial and ongoing performance on the 3rd day – much slower than the human data. In contrast, TAPPED predicts faster initial and ongoing performance, much closer to the actual human experiment results.

Table 3. The correlation (*r*) and root mean squared deviation (RMSD) and the percent the participant's performance that falls within the predicted 95% HDI (% Pred) of the predictions of TAPPED and PPE during the second experimental session.

	TAPPED			PPE		
Spacing	*r*	*RMSD*	*% Pred*	*r*	*RMSD*	*% Pred*
4	.30	3.70	75%	.39	6.43	51%
8	.36	3.25	73%	.43	5.23	54%
16	.40	3.15	74%	.50	4.50	54%
32	.36	3.13	61%	.49	3.88	64%

The out-of-sample correlation and RMSD increase in both models across each of the 4 spacing conditions, relative to Day 1. PPE (Table 3) has a higher *r* but higher RMSD compared to TAPPED (Table 3). The overall decrease in the correlation between the models' predictions of the 3rd day is expected. However, in addition to how well the model captured learning during the 3rd day, we are interested in the accuracy of each model's predictions. To investigate this we calculated the percentage of response times that fell within a model's predicted 95% HDI. With the exception of the SG32 spacing condition, TAPPED had a greater predictive accuracy compared to PPE (Table 3). Here it is seen that TAPPED, despite having a lower correlation compared to PPE, was able to better predict the participants' actual performance data. This result suggests TAPPED's additional complexity is warranted, given its ability to predict out-of-sample performance.

Discussion

The results presented in this paper revealed several interesting findings. First, a comparison of the inferences of learning phase by TAPPED to Phase HMM during the 1st day of performance found a high degree of similarity. Differences between the two model's inferences about learning phases stemmed from the assumptions about the sequences of phase transitions over time. This particular Phase HMM assumes a strict sequential transition between learning phases and does not allow for regression back to previous learning phases. In contrast, TAPPED holds no such assumptions. These differences between the two models' assumptions lead to a particular disagreement in inferences of the participant's initial phases of learning during the 1st several trials. This added flexibility of TAPPED provides a more realistic model of skill acquisition in which forgetting can occur between problems. While this matters less in a highly focused training paradigm where forgetting is less likely, in more spaced and varied training paradigms this is a strength over the Tenison and Anderson (2016) model. Despite the differences in the assumptions of learning phase transitions between the two models, the high degree of similarity of the classification of learning phase over the 1st day suggest that both models are capturing similar aspects within the data.

We also compared TAPPED to PPE, contrasting how each model accounts for participant performance during the 1st day and predicting participant performance on the 3rd day. TAPPED was able to better fit participants' performance during the 1st day compared to PPE. This is due to the fact that the TAPPED has a greater number of parameters and was able to selectively fit to the performance curve of the individual. Despite TAPPED's additional complexity, this model better predicts performance on the 3rd day. This increase in accuracy comes from the fact that the TAPPED model demarcated and classified changes in a participant's performance and used information from the participant's most recent stage of learning to make a prediction.

In contrast, PPE developed predictions based on all of a participant's data from the first day, leading to predictions of much slower performance on Day 3. The only exception to these regularities were seen in the longest spacing conditions, where the performance of both models is nearly equal. The

improved performance in PPE could have been due to the nature of the spacing manipulation. The highly spaced nature of the 32-item spacing condition could have decreased abrupt changes due to changes in phase in an individual's performance allowing for the PPE to better capture and predict performance during the 2nd experimental session. However, further exploration is needed and future research should address the similarities and differences between TAPPED and PPE when fitting and predicting items on a longer spaced schedule. Furthermore, future research should investigate using TAPPED in a more complex learning tasks where individuals might go through successive iterations of the three learning phases addressed in this paper or vary in their proficiency in which they learn a particular skill. Understanding these subtle fluctuations or differences in performance is important for being able to predict performance at the individual item level.

In summary, models of learning and retention often account for performance at a given time based on an individual's prior performance and temporal features of study and practice history. Often these models do not represent the cognitive mechanisms or changes in cognitive mechanisms individuals use when acquiring a particular skill and how these particular mechanisms might interact with the presentation history of the learned material. Our results suggest that detecting and modeling learning phases can improve predictive validity.

Acknowledgements

This research was supported by National Science Foundation grant DRL-1420008 and by the U. S. Air Force Research Laboratory's 711th Human Performance Wing, through the Personalized Learning and Readiness Sciences Core Research Area. MGC's participation was enabled through an appointment to the Oak Ridge Institute for Science and Education (ORISE) Student Research Participation Program. The views expressed in this paper are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government..

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369–406.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*(3), 369-378.

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Belmont, CA: Brooks/Cole

Gray, W. D., & Lindstedt, J. K. (2017). Plateaus, dips, and leaps: Where to look for inventions and discoveries during skilled performance. *Cognitive science*, *41*(7), 1838-1870.

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

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review*, *7*(2), 185-207.

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In the *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (pp. 1498-1508). Orlando, FL: National Training Systems Association.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. *Introduction to statistical relational learning*, *43*.

Lee, M. D. (2019). A simple and flexible Bayesian method for inferring step changes in cognition. *Behavior research methods*, *51*(2), 948-960.

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision, 6*(4), 335–368

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R Anderson (Ed.), *Cognitive skills and their acquisition* (pp. I-55). Hillsdale, NJ: Erlbaum.

Pirolli, P. L., & Anderson, J. R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *39*(2), 240.

Tenison, C., & Anderson, J. R. (2016). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(5), 749–767.

Tenison, C., Fincham, J. M., & Anderson, J. R. (2016). Phases of learning: How skill acquisition impacts cognitive processing. *Cognitive psychology*, *87*, 1-28.

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.