# Prediction Intervals for Performance Prediction

**Tiffany S. Jastrzembski (tiffany.jastrzembski@us.af.mil)**
**Kelly Addis (kelly.addis@mesa.afmc.af.mil)**
**Michael Krusmark (michael.krusmark@mesa.afmc.af.mil)**
**Kevin A. Gluck (kevin.gluck@us.af.mil)**
Warfighter Readiness Research Division, Air Force Research Laboratory
6030 S Kent Street, Mesa, AZ 85206 USA

**Stuart Rodgers (stu@agstechnet.com)**
AGS TechNet, 10887 Miriam Lane
Dayton, OH 45458 USA

**Abstract**

The Predictive Performance Equation (PPE) is a mathematical model of learning and forgetting developed to capture performance effectiveness across training histories, and to generate precise, quantitative *point predictions* of performance by extrapolating the unique mathematical regularities indicative of the learner. This equation is implemented in the Predictive Performance Optimizer (PPO) cognitive tool, designed to help learners and instructors make principled training decisions through examination of the learning and retention tradespace. Because the point predictive nature of the model implies a high degree of certainty, decision-makers could be misled into making less than optimal decisions in applied settings; and with regards to basic science, the model lacks prediction error and uncertainty which would more accurately represent the predicted range of human performance. Implementation of prediction intervals into a point predictive model of human performance is unprecedented in the psychological literature. We must balance the competing factors of reduced performance variation as practice accumulates, and greater prediction uncertainty as time spans increase. In this paper, we explore new methodologies for incorporating prediction intervals into quantitative predictions of future performance.

**Keywords:** point prediction; mathematical model; prediction interval; knowledge retention; skill retention

## Introduction

The Predictive Performance Equation (PPE) is a mathematical model of learning and forgetting developed to capture performance effectiveness across training histories, and to generate precise, quantitative *point predictions* of performance. This is accomplished by extrapolating unique mathematical regularities indicative of the learner from training history, while additionally accounting for the spacing at which knowledge and skills were trained to estimate the stability of performance across time. This equation is based upon robust findings in the psychological literature, and designed with the intent to be relevant in applied learning domains. As such, the PPE is implemented in the Predictive Performance Optimizer (PPO)–a cognitive tool designed to help learners and instructors make principled training decisions through examination of the learning and retention tradespace.

What the PPE currently lacks is a measure of uncertainty, because it contains no noise or error parameter in its current form. If the model is run 100 times, it will produce the same answer again and again. We know that if a human performs a task 100 times a range of performance values will be produced due to the usual suspects (e.g., distractions, fatigue, fluctuating motivation, random noise) coming into play. Thus, the point predictive nature of the model could be misleading due to the high degree of accuracy implied in its predictions. Therefore, it is necessary to incorporate principled measures of uncertainty, or *prediction intervals (PIs)*, around model point predictions. This provides the likely *range* of performance that is expected, and equips decision-makers with a more thorough picture.

Unfortunately, implementation of PIs into a *hybrid* point predictive model of human performance (to be detailed in the next section) is unprecedented in the psychological literature. By hybrid, we are referring to the notion that one step of the model functions by *calibrating* parameters to available historical data, while the other step *extrapolates* mathematical regularities beyond known data, to make true a priori predictions of performance for practical applications and purposes (e.g., Kahrs & Marquardt, 2007; Psichogios & Ungar, 1992).


Figure 1: Example of prediction uncertainty in the meteorological domain.

Other disciplines, including meteorology, econometrics, and the physical natural sciences, have well-established methods for incorporating uncertainty into time-series model predictions, such that in general, prediction uncertainty increases as time increases (see Figure 1). We may think of this trend as an expanding cone of uncertainty as lead time increases.

In the human performance domain, this is also a fair assumption to make. As the length of time between known data and a prediction increases, uncertainty would be expected to increase (see Figure 2).

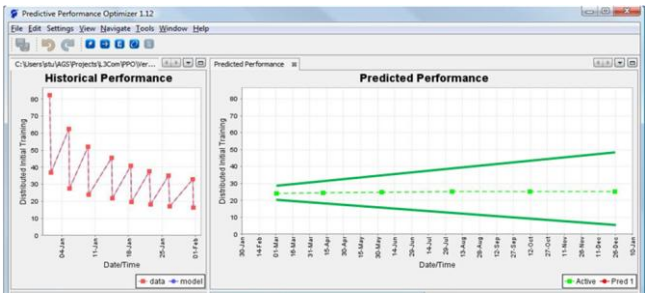



Figure 2: Notional training historical data and predicted refreshers to maintain performance from 1-10 months out.

What meteorological and econometric disciplines do not have to contend with is the fact that as practice accumulates, variability in human performance decreases (e.g., Ericsson, 1996; Rabbitt & Banjeri, 1989). Thus, model uncertainty should *decrease* as practice amasses (see Figure 3).


Figure 3: Expected levels of uncertainty for 3 regimens immediately following a 45-day lag and within a 2-4 day training block.

Furthermore, if multiple predictions are made, as shown in Figure 3, uncertainty is *conditionally* dependent on all previous model predictions. Thus, prediction uncertainty *n*-steps ahead of known empirical training history should generally grow incrementally larger-and prediction uncertainty should additionally be greater after a 12-month lag compared to a 1-month lag.

Thus, we are in the unique predicament of requiring a PI calculation method that balances the competing factors of reduced performance variation as practice amasses, and greater prediction uncertainty as lead time increases. Furthermore, to adhere to both basic and applied science demands, we need to ensure our methods are based on principle, while concurrently providing *useful* and *relevant* guidance for decision-making purposes. Before we turn our attention towards the new methodologies we are exploring to achieve alignment with these trends, we must first detail the nature of the hybrid point predictive human performance model.

## The Performance Prediction Equation

The PPE is built upon the strengths of the General Performance Equation (GPE) (Anderson & Schunn, 2000), which handles effects of recency and frequency very well, but is ill-equipped to handle effects of massed versus distributed practice. As such, the PPE formally extends the GPE by capturing effects of spacing, while providing the additional capability to predict performance at later points in time in an a priori fashion. The PPE is expressed as:

$$Performance = S \cdot St \cdot N^{c} \cdot T^{-d};$$

(Equation 1a)

where free parameters include *S*, a scalar to accommodate any variable of interest, *c*, the learning rate, and *d*, the decay rate. Fixed parameters include *T*, the true time passed since the onset of training, and *N*, the discrete number of training events that occurred over the training period. The term *St*, defined in Equation 1b, is short for Stability Term and is responsible for capturing effects of spacing by calculating experience amassed as a function of temporal training distribution and true time passed.

$$St = \left[ \frac{\sum lag}{P} \cdot \frac{P_i}{T_i} \cdot \frac{\sum_i^j \left( lag\,max_{i,j} - lag\,min_{i,j} \right)}{N_i} \right];$$

(Equation 1b)

*Lag* is computed as the amount of wall clock time passed between training events and *P* is computed as the true amount of time amassed in practice. As such, experience and training distribution attenuate performance by affecting knowledge and skill stability at the macro-level of analysis.

## Descriptive Adequacy across Data

We have validated the descriptive adequacy and predictive validity of this mathematical model across multiple types of previously published datasets available in the cognitive/experimental psychology literature, including empirical studies spanning knowledge acquisition, knowledge retention, skill acquisition, and skill retention. Goodness-of-fit measures across those domains have achieved an average $R^2$ of 0.98 (see Jastrzembski & Gluck, 2009, for additional information).

These results are encouraging. However, the datasets available in the psychological literature are from simple laboratory tasks, possessing few data points over an extensive retention period (e.g., Bahrick et al., 1993, study measured performance at seven points over the course of eight years), or measuring performance at short timescales (e.g., Glenberg, 1976, examined monotonic versus non-monotonic effects within one paired-associate training session). These datasets are useful to include in a larger test harness of empirical data to thoroughly validate model mechanisms, but their ecological validity is questionable.

Thus, it is necessary to validate against empirical data from more applied realms - where the interplay of knowledge and skill are often inextricably linked, extended lags between practice opportunities are on the

order of several weeks to multiple months, and knowledge and skill decay across extended lags can have a real impact on mission success. These features often characterize the nature of military training, where resources are both costly and scarce. As such, we validated PPE in a team coordination Unmanned Air Systems (UAS) reconnaissance task (Cooke, 2005), and with F-16 simulator air-to-air combat data collected in the Distributed Missions Operations testbed at the Air Force Research Laboratory (see Jastrzembski, et al., 2009). These highly complex datasets possess significantly longer inter-stimulus intervals than those found in the literature, and provide excellent opportunities to evaluate the incorporation of uncertainty within training blocks and across extended lags, where the need to provide estimates of uncertainty have very clear ramifications.

## Predictive Performance Equation Methodology

We will now explain the two distinct, non-stochastic sequential steps in our performance prediction methodology. The first step in using PPE deals with calibrating, or optimizing (using maximum likelihood estimation), the learning and decay parameters to the unique mathematical regularities of the learner, identified by tracking training history. The second step is extrapolating the mathematical regularities to make true a priori predictions of performance at specified future times. We make this distinction because it is commonplace for modelers in the cognitive science community to use the term *prediction* when *fitting* empirical datasets, often in a post-hoc manner; whereas we use the term *calibration* to refer to that fitting process, and prediction for out of sample calculations.

With regard to the UAS reconnaissance study (Cooke, 2005), teams of three individuals were required to coordinate to fly a UAS and attain pictures of targets. They completed five 40-minute missions on the first day of training (the training baseline used for model calibration), and returned either one or three months later to complete an additional three missions (used to validate model a priori predictions) (see Figure 4).

The design of the DMO study was similar in nature, but required teams of four F-16 pilots to fly air-to-air combat missions over a more extensive training baseline (one to two hour-long missions trained each day over for five days), allowing us to examine skill acquisition and decay patterns both within days (where prediction uncertainty should decrease) and across days (where prediction uncertainty should increase). Teams were reassessed either three or six months later and completed three hour-long missions over the course of two training days (see Figure 5 for individual team level analysis).

The need to incorporate valid PIs around model point predictions becomes extremely evident in the following potential use cases, as PPO is indeed intended to help decision-makers make informed training decisions. As shown in Figure 6, PPO may be used to help determine how many additional practice opportunities unique learners (an F-16 pilot team in this case) need to achieve a desired level of performance (denoted as achievement of a wing standard of 0.015 in this particular case).

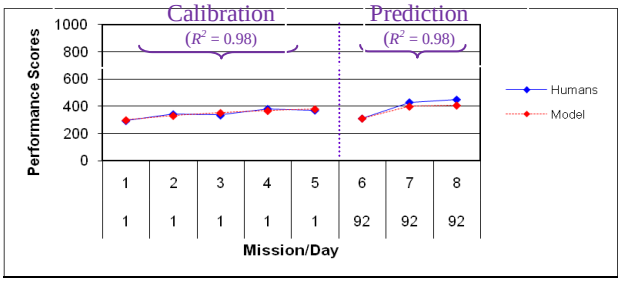



Figure 4: Aggregate team performance in a UAS task, with a three month lag.

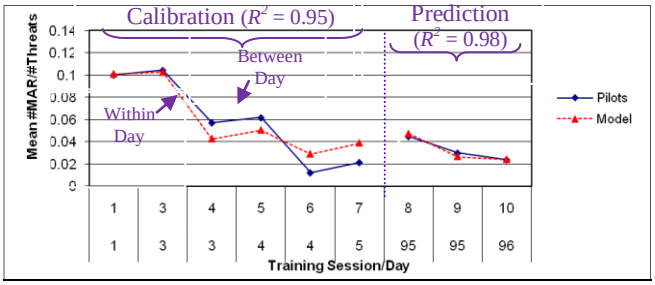
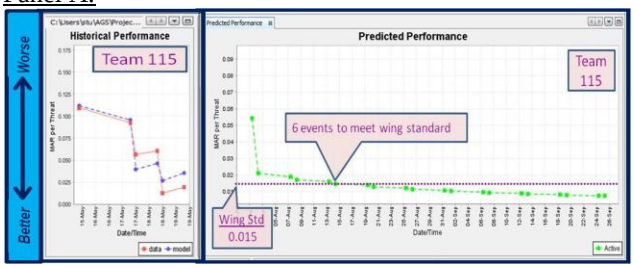



Figure 5: Number of times enemy airspace was violated by an individual F-16 team, with a lag of three months.

PPO takes in the historical data for each unique team, optimizes the learning and decay parameters to the mathematical regularities inherent in the training history, and makes customized team performance predictions by extrapolating those learning trends into the future. Thus, Team 115 (shown in Figure 6, Panel A) is predicted to require six additional training events to achieve the desired performance level, while Team 112 (shown in Figure 6, Panel B) is predicted to require 20 more events.

Panel A:




Panel B:

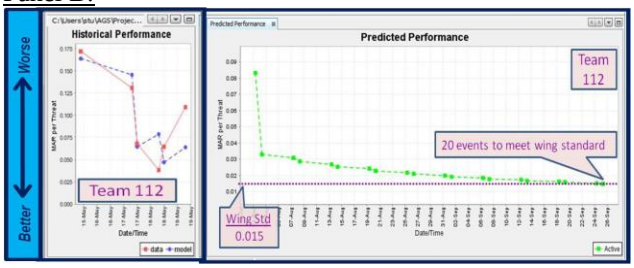



Figure 6: Model predictions for two unique F-16 pilot teams to achieve the same criterion.

In line with statistical principles, as PPO makes multiple time-series dependent predictions, significant *uncertainty* will build for predictions made farther and farther ahead in time from actual historical data. Thus, in the example above, Team 115's predicted attainment of criterion is actually more certain than Team 112's, simply due to the fact that criterion is reached with fewer timesteps ahead from the historically calibrated data.

Another potential use case that nicely demonstrates the need to incorporate "risk" into model point predictions is revealed by PPO's capability to examine performance implications across a multitude of potential future training regimens.

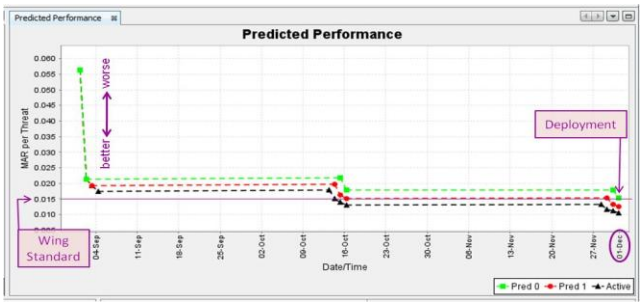



Figure 7: Future training regimen comparisons to identify which training routine best meets desired goals.

The graph revealed in Figure 7 is calibrated upon the historical F-16 pilot team performance data shown in Figure 6, Panel A, and depicts predicted levels of performance under three distinct training regimens. The green line depicts two training opportunities given in each training block (occurring every 45 days), while the red line reveals three, and the black line reveals four. Noting that a desired performance effectiveness level of 0.015 is to be reached by the intended deployment date, the learner or instructor may easily inspect and assess the efficiency and effectiveness each potential future training regimen will likely provide.

As shown in Figure 7, the red and black lines both achieve the desired performance level by deployment, while the green line does not. However, PIs for the black line should theoretically be smaller than those in the red line - because more training opportunities are provided meaning performance variability should be reduced. Thus, less risk would be involved in deploying trainees who completed the black training regimen.

Given the potential ramifications these types of prospective use case decisions entail, it becomes very clear why the incorporation of prediction uncertainty measures is needed. Further, equipping PPE with these measures will better aid decision-makers' understanding both learning and training needs, as well as the risks.

## Prediction Interval Calculation Methodology

As previously expressed, there is no precedent for incorporating PIs into a human performance point prediction model of this nature. As such, we have developed and are investigating new methods to achieve our goals of both reducing variability as practice amasses, and increasing variability at longer lead times.

### Extrapolation of Residuals

The first method we are investigating involves extrapolating residuals from calibrated model predictions and human empirical data to model point predictions. Residuals are often used to add uncertainty to models in other disciplines, like econometrics (see Chatfield, 2001, for a review); but as previously mentioned, other disciplines do not have the added phenomenological complexity of uncertainty decreasing as practice increases, nor do they have good solutions for estimating *how much* larger PIs should be after lags of increased length. Thus, in order to base a PI method on residuals in the human performance domain, a good deal of innovation will be required to ensure estimates stay true to expected human performance trends.

As such, we have modified the residuals by the stability term (see Expression 1) and will illustrate PI incorporation based on this method later in this paper.

$$S * St * N^{c} * T^{-d} \pm (z_{\alpha/2} * E[RMSD] * St_{i+h});$$

(Expression 1)

### The Coefficient of Variation

The second method we have developed and are continuing to investigate deals with adding variability into the learning and decay parameters themselves. The amount we have chosen to vary parameters by is the coefficient of variation (CV), selected because it is a unitless measure of deviation between model predictions and human empirical data, generally ranging between zero and one (Schweickert, et al., 2003), and it has previously been used to incorporate stochasticity into other types of cognitive and task performance models (Patton, et al., 2009; Patton & Gray, submitted; Schweickert, et al., 2003). It is calculated across historical training calibration data using Equation 2:

$$CV = RMSD/model\ mean;$$

(Equation 2)

and integrated into PPE in the following way (see Expression 2):

$$S * St * N^{c \pm c*CV} * T^{-(d \pm d*CV)};$$

(Expression 2)

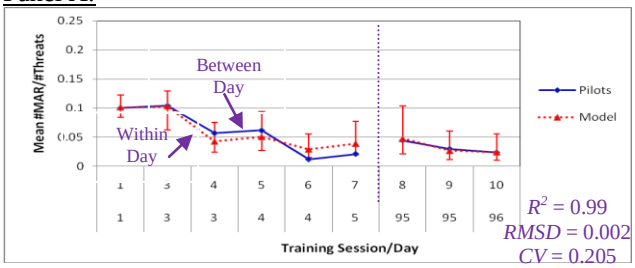
thus producing upper and lower PI bounds.

Desirable qualities of this measure include a readily available mapping to the learning and decay rates, which also range from zero to one; and greater variability being added into models that produce lower quality calibrated fits to empirical data, producing larger PIs as a result.

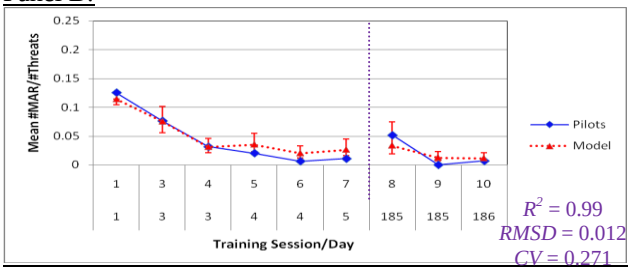
**Prediction Interval Utility in the Applied Domain**

We now illustrate the PI incorporation across four unique F-16 pilot teams, possessing differences in learning regularities and quality of calibration fit – leading to differences in PI spans as a result (see Figures 8 and 9).

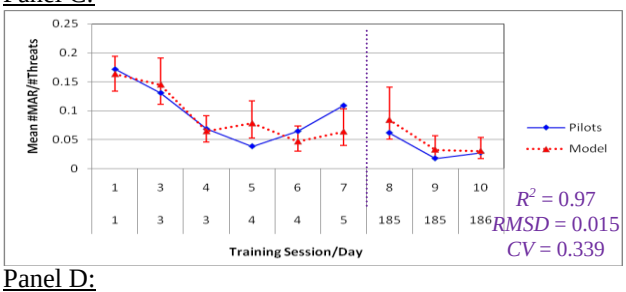
Panel A:




Panel B:




Panel C:




Panel D:

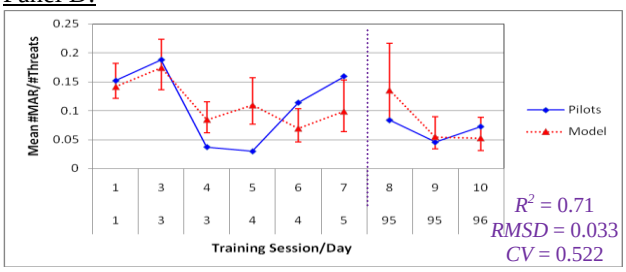
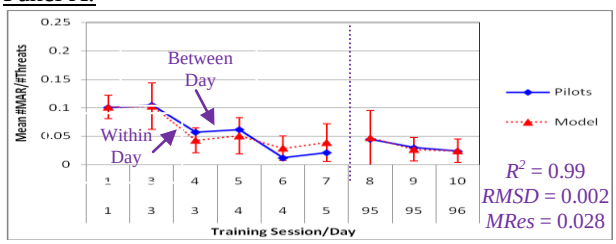



Figure 8: CV PI incorporation for F-16 pilot teams.

As revealed by Figures 8 and 9 (Figure 9 displays identical empirical data displayed in Figure 8, Panels A and D), each method produces larger PIs *between* training days and smaller PIs *within* training days – thus, mapping nicely onto human empirical findings showing that performance variation decreases as practice amasses. They also reveal wider PI bands following the three or six month lag relative to other predicted points; thereby aligning with the notion that longer lead time predictions are more uncertain than predictions at shorter lead times.

An added unexpected, but very desirable effect, of the CV method was that the PI bands are asymmetrical in nature – thereby diverging from standard symmetrical estimates of confidence or error (as revealed by the residual-based method). This is pleasing in cases where human performance is bounded by a floor or ceiling, (ceiling performance was zero on the y-axis in Figures 8 and 9). Thus, there is more room to err (the higher end of the y-axis) and less room to gain (performing closer to zero), mapping nicely to CV-based error bars having longer upper than lower whiskers.

Panel A:




Panel B:

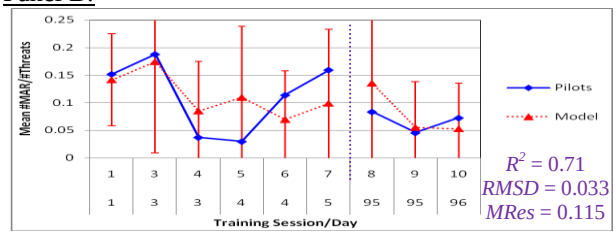



Figure 9: Residual-based PIs across unique F-16 teams.

Comparison of these PI methods to empirical data reveal that utilization of residuals, compared to the CV-based method, tends to produce larger error bars in general (it is more liberal, but covers more of the data), produces error bands outside the bounds of possible performance (below zero in this case), and is more sensitive to noisy data (see Figure 9, Panel B – the same empirical data as Figure 8, Panel D). This raises concerns for how useful a residual-based approach will be as a decision-making guide. As such, additional modifications are being examined.

**Resolution of Data** In our last set of analyses, we will limit our discussion to the CV PI methodology, due to limitations of the residual-based method described above. Using data collected in the UAS reconnaissance task (Cooke, 2005), we applied PIs to models aggregated at different grains of analysis. Given the intended utility of the PPO as a principled training decision guide, it is important to understand the implications of using a predictive model at the aggregate, team, and individual learner levels of performance (see Jastrzembski, et al., 2006), as aggregate data, by definition, reduces noise through averaging procedures that smooth out the shape of human performance curves. Thus, data will always be noisier at finer and finer grains of resolution, implying PIs should be wider and wider as aggregation decreases. We inspect the ability of the CV PI method to align with this phenomenon as shown in Figures 10-12 below.

As we might expect, PIs for the first point prediction after the lag are indeed larger after the long delay ($PI_{range}$ = 146) compared to the short delay ($PI_{range}$ = 129), revealed in Figure 9, showcasing the fact that predictions at longer lead times will be less certain than predictions at shorter lead times. This effect is generated in PPE because the upper and lower CV bounds are placed in the learning and decay exponents, which interact with the number of training opportunities accumulated, as well as the actual amount of time passed.

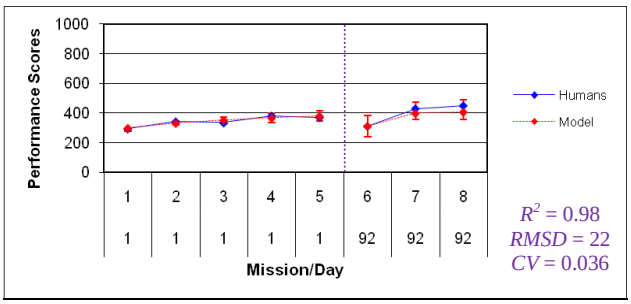



Figure 10: Aggregate performance across all teams in the UAS reconnaissance task, with lags of 30 or 91 days.

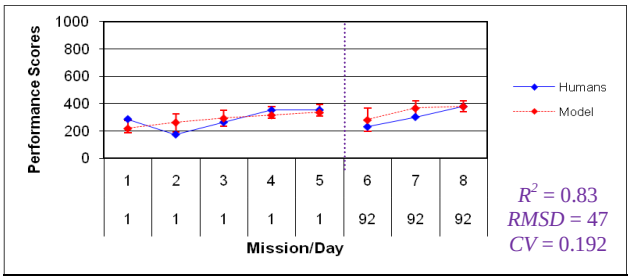



Figure 11: Individual UAS team performance.

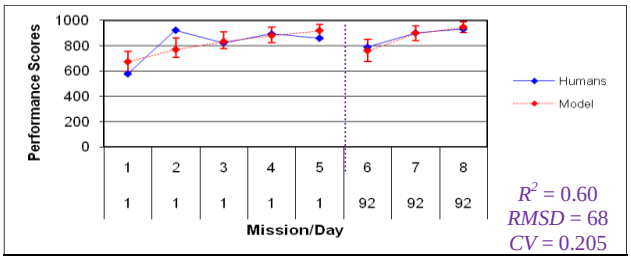



Figure 12: Individual UAS team member performance.

Finally, we note that the CV increases as we move from aggregate to team to individual levels of performance, as expected (see Figures 10-12). This is a useful property to note because it shows that decisions may be riskier at finer grains of resolution. One way to help circumvent this problem at finer grains of analysis is to in fact accumulate larger training histories to calibrate PPE upon, allowing variability and noise to be smoothed.

These illustrative exercises help lend credence to the notion that use of this newly developed CV PI calculation method may have merit as being a useful way to help guide training decisions in a way that nicely accounts for the competing trends of reduced performance variability expected with increases in practice, and increased prediction uncertainty expected for longer lead times.

## Conclusions

The incorporation of estimates of uncertainty into model point predictions is a necessary extension to our point predictive model in order to provide learners and instructors with relevant and useful guidance concerning the amount of predictive uncertainty that should be expected at specific future points in time and under competing future training regimens. Because there are no precedented existing methodologies to apply to this problem, we plan to further the validation effort across the two potential solutions we proposed in this paper against human empirical data, and we are hopeful this new capability will apply not only to our modeling effort, but also for others who are working on the optimization of training (e.g., Lindsey, et al., 2009; Pavlik, & Anderson, 2008; van Rijn et al., 2009).

## References


Bahrick, H., & Phelps, E. (1987). Retention of spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 344-349.

Chatfield, C. (2001). Prediction intervals for time series. In J.S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Practitioners and Researchers*. Norwall, MA: Kluwer.

Ericsson, K. A. (Ed.) (1996). *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games.* Mahweh, NJ: Erlbaum.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning & Verbal Behavior, 15*, 1-16.

Jastrzembski, T. S., & Gluck, K. A. (2009). A formal comparison of model variants for performance prediction. *Proceedings of the International Conference on Cognitive Modeling (ICCM)*, Manchester, England.

Jastrzembski, T. S., Rodgers, S., & Gluck, K. A. (2009). Improving military readiness: A state-of-the-art cognitive tool to predict performance and optimize training effectiveness. *Proceedings of the I/ITSEC annual meetings*, Orlando, Florida.

Kahrs, O., & Marquardt, W. (2007). The validity domain of hybrid models and its application in process optimization. *Chemical Engineering and Processing, 46*, 1054-1066.

Lindsey, R. Mozer, M., Cepeda, N., & Pashler, H. (2009). Optimizing memory retention with cognitive models. *Proceedings of the ICCM* annual meeting, Manchester, England.

Patton, E. W., & Gray, W. D. (submitted). SANLab-CM: A tool for incorporating stochastic operations into activity network modeling. *Behavior Research Methods*.

Patton, E. W., Gray, W. D., & Schoelles, M. J. (2009). SANLab-CM – The stochastic activity networking laboratory for cognitive modeling. *Proceedings of the 53rd HFES Annual Meetinge*, San Antonio, Texas.

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied, 14*, 101-117.

Psichogios, D. C., & Ungar, L. H. (1992). A hybrid neural network – first principles approach to process modeling. *AIChE Journal, 38(10)*, 1499-1511.

Rabbitt, P., & Banjeri, N. (1989). How does very prolonged practice improve decision speed? *Journal of Experimental Psychology: General, 118*, 338-345.

Schweickert, R., Fisher, D. L., & Proctor, R. W. (2003). Steps toward building mathematical and computer models from cognitive task analyses. *Human Factors, 45(1)*, 77-103.

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. *Proceedings of the ICCM* annual meeting, Manchester, England.