Predicting Performance in Cardiopulmonary Resuscitation

Kevin A. Gluck[1] (kevin.gluck@us.af.mil)
Michael G. Collins[2] (michael.collins.ctr@us.af.mil),
Michael A. Krusmark[3] (michael.krusmark.ctr@us.af.mil)
[1]Air Force Research Laboratory, [2]ORISE at AFRL, [3]L3 Technologies at AFRL
Wright-Patterson Air Force Base, Ohio, USA
Florian Sense (f.sense@rug.nl), Sarah Maaß (s.c.maass@rug.nl), Hedderik van Rijn (d.h.van.rijn@rug.nl)
Department of Experimental Psychology, University of Groningen
Groningen, The Netherlands

Abstract

Cardiopulmonary resuscitation (CPR) is a real-world basic lifesaving skill that requires a complex combination of declarative memory and psychomotor skill. It is also simple and brief enough to be practical for laboratory use. Here we describe a repeated measures study with increasing lags between sessions. At the time of the writing of this initial manuscript submission, the final session of CPR performance data has not been run. This paper documents our participant-level performance predictions for that final session, using the Predictive Performance Equation (PPE; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018). With the final lag period for that final experiment session at approximately one year for every participant, we will be able to assess predictive accuracy of PPE over an ecologically relevant timeframe.

Keywords: skill acquisition; retention; learning; memory; CPR; performance prediction; mathematical model

Introduction

Cardiopulmonary resuscitation (CPR) is an essential component of first aid training. The Basic Life Support CPR procedure, as laid out by the European Resuscitation Council (ERC) guidelines (Soar et al., 2015), includes an assessment of the so-called victim (check response, check breathing) and a series of steps (alert emergency services, hand positioning) before the actual chest compressions and rescue breaths are administered in a cycle of 30 compressions and two rescue breaths. In addition to its being a critical life-saving skill, CPR is a useful domain for studying human performance. It is a task that combines declarative knowledge and psychomotor skills, and clear performance standards are available. High fidelity training and assessment equipment, such as the Laerdal QCPR manikin used in this study, record and store detailed performance measures automatically.

Crucially, CPR certification entails periodic retraining to ensure performance remains above criterion. For medical professionals, retraining is typically completed every other year. This interval is often considered suboptimal as performance is likely to drop below the criterion during this period (Stross, 1983). Furthermore, the American Heart Association (AHA) recognizes that large individual differences in CPR performance exist, which complicates the prescription of ideal methodology and frequency of training (Nolan et al., 2015). Consequently, cognitive-psychological theories of learning and retention over realistic time-frames could provide a benefit to public health and safety by accurately predicting when someone should be provided with additional training to remain above performance criteria.

To validate learning and retention theories for this purpose, we initiated the collection of a CPR dataset (Sense, Maaß, Gluck, & van Rijn, 2019, https://osf.io/m8bxe/). A benefit of studying CPR performance is that there exist certain sub-populations who have been trained previously on this task before entering the lab. Specifically, part of the requirements to obtain a German driver's license is to demonstrate CPR performance above criterion. Therefore, German students with a driver's license are a suitable population to test long-term retention of procedural and declarative knowledge because they had CPR training in the past, typically had no retraining, and there will be natural variation in time since last presentation.

Mathematical models of learning and retention can help describe fluctuations in CPR performance over time based on individuals' prior performance. Sometimes the motivation in research and application of these models is to optimize repetition schedules *within* individual learning sessions (van Rijn, van Woudenberg, & van Maanen, 2009; Sense & van Rijn, submitted). Earlier research has shown that an ACT-R-based cognitive model can use response accuracy and latency on a trial-by-trial basis to predict when each studied item is likely to be forgotten and ensure rehearsal before that moment. This improves retention of the studied facts (Sense & van Rijn, submitted) and allows the estimation of a learner's rate of forgetting (Sense, Behrens, Meijer, & van Rijn, 2016).

Other times the motivation is to predict performance over longer time periods *between* sessions. This is our primary interest in the analyses reported here. A model that has shown some promise regarding its predictive validity over those longer between-session intervals, regardless of the relative mix of declarative or procedural knowledge involved, is the Predictive Performance Equation (PPE; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, 2018). PPE is a set of equations capturing key human performance dynamics. First, activation increases with the number of learning events (N). This is implemented as a power law of learning, with the *learning rate* fixed at 0.1 based on prior empirical evidence and model fits (Equation 1). Because participants enter the study at different experience levels, we add to N a free

parameter, *a*, to represent each individual's past CPR experience. Second, performance drops as a function of elapsed time among practice events (*T*; Equation 1). This is implemented as a power function of forgetting, with the decay rate determined by the function expressed in Equation 4, below. In PPE the effects of learning and retention on activation is multiplicative, such that:

$$activation = (N + a)^{learning\ rate} \cdot T^{-decay\ rate} \quad (1)$$

Third, PPE captures the spacing effect, such that retention is better and more stable when practice is distributed over time. This is implemented in the forgetting function through *T* and the *decay rate*. *T* is computed as the sum of the weighted age of each practice event,

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

where the weight, w_i, is an exponential decay function of time,

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

Thus, *T* weighs practice repetitions so that more recent events carry more weight, and the variable *x*, which is fixed at 0.6, controls the degree of the weighting.

The *decay rate* is computed as a function of the complete history of lags between successive practice opportunities (lag_j):

$$decay\ rate = decay\ intercept + decay\ slope \cdot average\left(\frac{1}{log(lag_i)}\right) \quad (4)$$

Finally, in PPE performance is computed as a logistic function of activation:

$$Performance = \frac{1}{1 + exp\left(\frac{threshold - activation}{scalar}\right)} \quad (5)$$

PPE parameters are estimated separately for individuals based on their performance histories. For each individual, we compute the optimal values of the *decay intercept*, *decay slope*, *threshold*, and *experience (a)* parameters that maximizes the likelihood of individual performance trajectories. These parameter values are then used to generate out-of-sample predictions of performance at future points in time.

An increasingly common modeling practice in environments with sparse and noisy data is to seed a model's parameter values with priors. This avoids over fitting and improves out-of-sample prediction (Yarkoni & Westfall, 2017). There has been some previous exploration of this approach in the context of PPE (Collins, Gluck, Walsh, & Krusmark, 2017; Collins & Gluck, 2018). Here we use priors generated from an independently completed CPR study (Jastrzembski et al, 2017). In that research, CPR compression data from four sessions separated by either one day, one week, one month, or three months were used to calibrate the model and generate model parameters for temporal predictions at either three or six months in the future. These model parameters were used in the current study to inform parameters generated during the model fitting process. The generalization of priors allows PPE to use available prior information about human performance on CPR.

The current study was devised to assess the accuracy of personalized performance predictions. The ERC's guidelines (Soar et al., 2015) state that "*The intervals for retraining will differ according to the characteristics of the participants*". The availability of predictive models that take an individual's performance profile over time into account permits such personalized predictions. Ideally, this would make retrainings more efficient and reduce the interval during which a medical professional might perform below criterion. The viability of personalized refresher schedules crucially depends on the accuracy of the predictions: Requiring people to retrain too early is a waste of resources but requiring them too late can cost lives.

Method

Participants

Fifty participants (age range = [18, 27]) were recruited for a first learning session, 40 took part in the second session, and 35 participated in the third session. All participants held a valid German driver's license.

Procedure and Stimuli

The full experiment protocol includes four sessions in which CPR performance data are collected. At the time these model predictions were run, participants had completed three experimental sessions, with the fourth session upcoming. In addition to assessing CPR performance in all sessions, a set of computerized laboratory tasks more typical of experimental cognitive psychology were also administered in some of the sessions. These are documented elsewhere in detail (Sense, Maaß, Gluck, & van Rijn, 2019, https://osf.io/m8bxe/) and are not a focus in this paper. A graphical summary of the CPR-specific experiment protocol is provided in Figure 1.

Session 1: Test 1.1. At the beginning of each session, participants signed the informed consent forms. In the first session participants then entered the experimentation room where a Laerdal Resusci Anne QCPR manikin was lying on the ground. Participants read the following instructions: "*You volunteered for community service to help elderly neighbors with chores in their homes. When you enter the house of Mr. Johnson, you find him on the living room floor. There are no signs of bleeding or open wounds and no one else is in the house. Based on your first aid training, take the steps*

*necessary in this situation on the manikin to **assess** and **react upon** the situation.*"

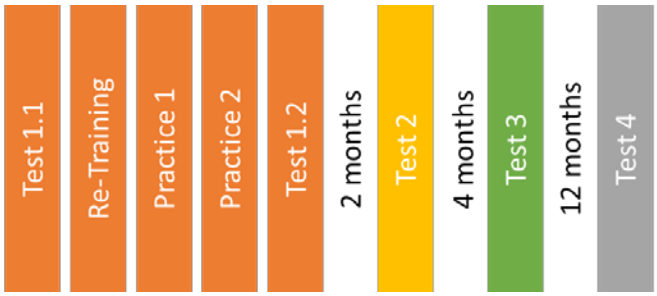


Figure 1. Overview of the experiment protocol.

This scenario was chosen to sketch a hypothetical scenario that required participants to perform CPR on the manikin. They were asked to perform the necessary actions required according to the ERC guidelines (Soar et al., 2015). This means participants were supposed to alternate between 30 compressions and two rescue breaths. Participants were stopped after administering four cycles of compressions and rescue breaths (i.e., 30-2, 30-2, etc.) to avoid fatigue. We refer to this procedure (i.e., initial steps followed by four rounds of 30-2) as one run-through of CPR.

Performance scores were based on Laerdal's proprietary scoring algorithm using the European guidelines (ranging from 0 to 100%; a score of 75% or higher is considered "proficient").

Retraining. After the initial assessment, participants were re-trained. First, participants watched a short instructional video (see https://osf.io/9er6g/) demonstrating the initial steps, as well as instructions on how to correctly apply chest compressions and rescue breaths. This video was specifically made for this research project.

Subsequently, participants had the opportunity to practice compressions on the manikin with its live feedback option enabled for one minute. That is, for each compression participants could track their depth and frequency and adjust if necessary. Then participants practiced giving rescue breaths until the live feedback indicated that two correct breaths had been given in a row. Following retraining, participants completed a basic lab task. As noted earlier, due to space limitations, details about the basic lab tasks will not be discussed.

Practice 1 and 2. Participants were instructed to "*Perform the complete procedure you saw in the video, with four rounds of compressions and rescue breaths*" twice while their performance was scored.

After the run-throughs of CPR, participants completed questionnaires to gather demographic information, the date their driver's license was issued, and the approximate number of months between obtaining their license and completing their CPR training. The time between the mandatory training and obtaining the driver's license ranged from 1 to 60 months (mean = 9.92 and SD = 12.71). Participants then completed two more basic lab tasks.

Test 1.2. Following the computerized tasks, participants were asked to complete another run-through of CPR. If the score of this test was below 75%, participants were re-trained until they reached criterion.

Session 2: Test 2. Participants completed a run-through of CPR. If performance was below 75%, they repeated the run-through. Participants also completed the full set of basic lab tasks.

Session 3: Test 3. Participants completed another run-through of CPR. If performance was below 75%, they repeated the run-through. Participants also did one minute of chest compressions without live feedback from the manikin, then rescue breaths until two consecutive ventilations were correctly performed.

Session 4: Test 4. Participants will complete another run-through of CPR. Then participants will watch the short instructional video again (as in Session 1) and complete another run-through of CPR.

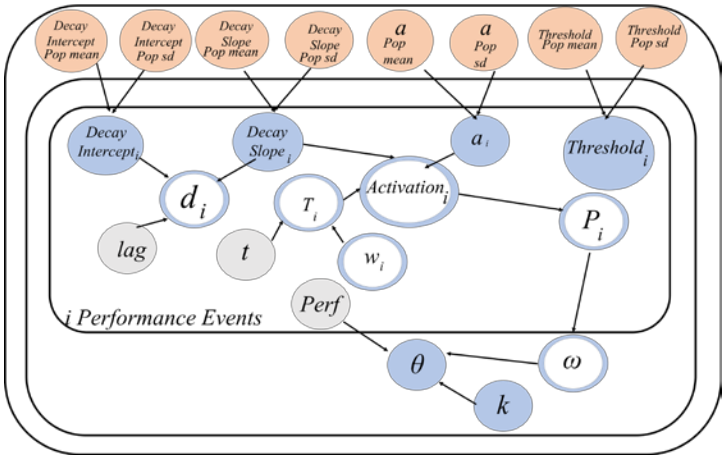
Predicting Future CPR Performance

As noted earlier, individual participant single performance event-level predictions present a small data challenge, especially earlier in the protocol. To manage and avoid overfitting to unexplained individual event-level variation we use Hierarchical Bayesian Modeling (Kruschke, 2014; Lee & Wagenmakers, 2013) to bias the PPE parameters with priors from a previous CPR study and generate posterior predictive distributions for each participant remaining in this study through Session 3.

In predicting CPR performance on Session 4, PPE's free parameter values were estimated using the model shown in Figure 2. For each participant, the model estimates probability distributions for PPE's free parameters (*decay intercept, decay slope, a, threshold*) that best characterize performance over the first three sessions. The estimates are based on a set of hyperparameters (decay intercept_{pop} decay slope_{pop}, a_{pop}, threshold_{pop}). that were estimated from individual-level CPR performance data collected in a previous CPR study (Jastrzembski et al, 2017).

Unique parameter distributions are sampled for each individual participants from the hyperparameters to derive a distribution of values for each free parameter. The sampled set of parameters are then combined with the student's fixed time variables (t, N) and are transformed into performance predictions ($Perf_{i:n}$). The average of these performance predictions ($Perf_{i:n}$) is represented by variable omega (ω) and is then combined with free parameter k, to represent the model's prior beliefs of the student's performance ($\boldsymbol{\theta}$). This prior is then combined with the student's actual performance to generate a posterior estimate of performance.

Under this methodology, PPE's free parameter values are treated as a probability distribution, representing our degree of belief in a particular parameter value to generate a prediction. The final posterior probability distribution used to determine PPE's prediction is affected by two factors: (1) Prior, the beliefs about the most likely free parameter values before observing the performance of a participant (i.e., Prior



$$\text{Decay Intercept}_{\text{pop mean}} \sim \text{Normal}(.02\ .001) \qquad \text{Decay Intercept}_{\text{pop sd}} \sim \text{gamma}(854, 82536)$$

$$\text{Decay Slope}_{\text{pop mean}} \sim \text{Normal}(.003, .03) \qquad \text{Decay Slope}_{\text{pop sd}} \sim \text{gamma}(127.97, 11652.7)$$

$$a_{\text{pop mean}} \sim \text{Normal}(.33\ .04) \qquad a_{\text{pop sd}} \sim \text{gamma}(3.80, 106.91)$$

$$\text{Threshold}_{\text{pop mean}} \sim \text{Normal}(.46\ .02) \qquad \text{Threshold}_{\text{pop sd}} \sim \text{gamma}(32.88, .195.96)$$

$$\textit{Decay intercept}_i \sim \text{Normal}(\text{Decay Intercept}_{\text{pop mean}}, \text{Decay Intercept}_{\text{pop sd}})$$

$$a_i \sim \text{Normal}(a_{\text{pop mean}}, a_{\text{pop sd}})$$

$$\textit{Threshold}_i \sim \text{Normal}(\text{Threshold}_{\text{pop mean}}, \text{Threshold}_{\text{pop sd}})$$

$$\textit{Decay slope}_i \sim \text{Normal}(\text{Decay slope}_{\text{pop mean}}, \text{Decay slope}_{\text{pop sd}})$$

$$d_i = \textit{decay intercept}_i + \textit{decay slope}_i * \left(\frac{1}{n-1} * \sum_{j=1}^{n-1} \frac{1}{\log(lag_j + e)}\right)$$

$$T_i = \sum_{l=1}^{i} w_l * t_i \qquad w_l = -t_l^{-.75} \sum_{m=1} \frac{1}{t_m^{-.75}}$$

$$Activation_i = (N + a_i)^c * T^{-d} \qquad P_i = \frac{1}{1 + \exp\left(\frac{Threshold_i - Activation_i}{.1}\right)}$$

$$\kappa \sim \text{gamma}(1, 1)$$

$$\omega = \frac{\sum P_i}{\max(i)}$$

$$\theta \sim beta(\omega * (\kappa - 2), (1 - \omega) * (\kappa - 2))$$

Figure 2. The hierarchical Bayesian model used to estimate free parameter values (*Decay Intercept*$_i$, *Decay slope*$_i$, a_i, *Threshold*$_i$ - all shown in blue) for PPE prediction of performance in the 4th CPR session, given prior distributions (*Decay Intercept*$_{pop\ mean}$, *Decay Intercept* $_{pop\ sd}$, *Decay Slope*$_{pop\ mean}$, *Decay slope* $_{pop\ sd}$, $a_{pop\ mean}$, $a_{pop\ sd}$, *Threshold* $_{pop\ mean}$, *Threshold* $_{pop\ sd}$ - all shown in salmon) based on a sample of CPR performance from a different study and the participant's prior performance across the first three sessions (*Perf*). Random variables are represented as circles, deterministic variables are represented as double circles, and observed variables (*lag*, *t*, *Perf*) are in grey.

CPR performance), and (2) the performance of a particular participant over the course of three sessions (i.e., likelihood). These two factors are combined together to generate a posterior distribution for each of PPE's free parameters. It is this posterior distribution that is used to make predictions for each participant's next performance. We do this iteratively through the experiment protocol for each of the 35 participants, culminating in a prediction for their upcoming performance in Session 4.

Results

Data collection for Session 4 is scheduled for May 2019, after the submission deadline of this manuscript. Our key interest at present is in documenting our predictions for each participant's CPR performance when they return for Session 4, approximately one year after they did Session 3.

In the process of generating Session 4 CPR predictions for each participant, we ran several simulations to assess how different assumptions about participants' past CPR experience would affect predictions. Recall that the participants in the current study were German college students with a valid driver's license, which required them to successfully complete CPR training prior to getting their license. Thus, from the issue dates on the licenses, we know the approximate date of each participant's initial CPR training. Although no performance measures are available, we assumed that all participants reached criterion level of performance (i.e., 75%) during this initial training. Based on this information, we combined the 75% performance score that we assumed at the time of licensing with the data from Sessions 1 and 2, and predicted performance on Session 3. We then compared the accuracy of these predictions to predictions from the model with only data from Sessions 1 and 2 predicting Session 3. Results of this comparison showed that predictions were more accurate when we ignored the licensing data. A possible explanation for this is that we were making assumptions about the level of performance participants reached when they received their license, but not that they started with no experience. Thus, we ran the model again assuming that performance was 0 prior to their initial training, and that it increased to 75 afterwards. But again, this did not improve predictive accuracy. Predictions were more accurate when we made no assumptions about CPR performance prior to the onset of the study.

Figure 3 plots data for the fit and prediction methodology described in the previous section for each of the 35 participants. A data file documenting the raw values used to generate the graph is available at (https://osf.io/5ma29/). Performance scores are exported from Laerdal's proprietary software, which combines the compression and rescue breath performance into a single score.

On the initial test at the beginning of the first session, only two participants demonstrate proficient performance (a score of 75% or above), while many score below 25%. The CPR Retraining administered between CPR Test 1 and CPR Practice 1 results in a marked increase in performance, making the majority of participants reach criterion. Testing for a difference between those two scores with a paired Bayesian t-test yields a decisive Bayes Factor of 2.4×10^{17} in favor of a difference. The second practice marks a further increase in overall performance and the vast majority of participants retain above-threshold performance until CPR Test 2 at the end of the first session. In the eight-week interval until the second session, and performance decreases (BF_{H1} = 9.95) but many participants still exhibit near-ceiling performance.

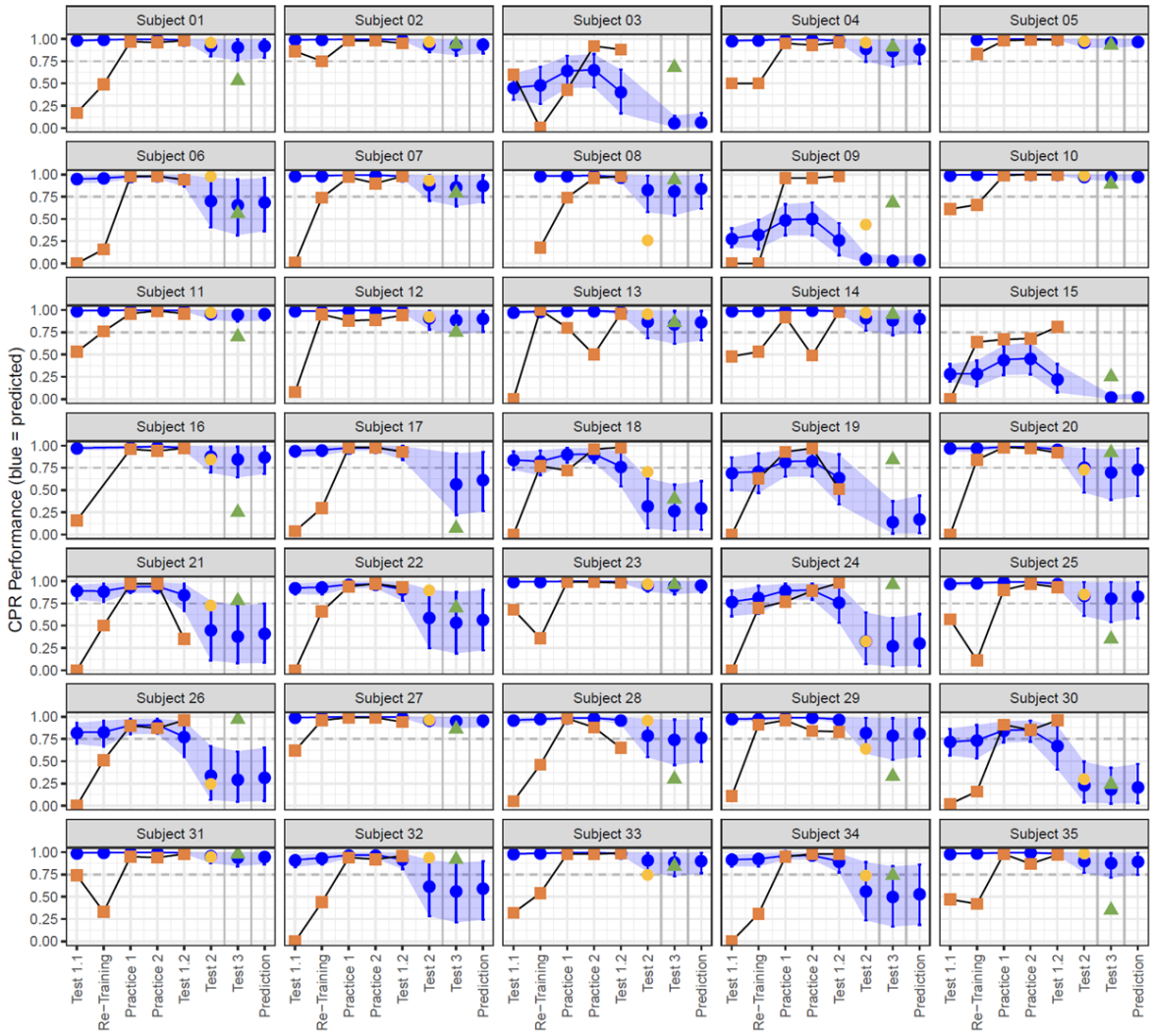


Figure 3. CPR performance for each of the 35 participants. Observations within a session are connected by lines; sessions are indicated by color and shape. Session 1 is orange squares. Session 2 is a yellow circle. Session 3 is a green triangle. The model's predicted performance is shown in blue, with mean predicted performance at each measurement indicated by the blue circle. The blue ribbons indicate the 95% highest density interval (HDI) of the posterior distribution at each instance. The predictions for the final session are the rightmost blue circles in each panel.

Another way to summarize the data and contrast the predictions with the recorded data is to compute a prediction error at each measurement event. In Figure 4, predicted performance has been subtracted from the actual performance to express the prediction error at each measurement event. The color-coding indicates the session and the root-mean-squared-error (RMSE) is listed for each event to summarize the prediction error. At the first event, most errors are negative, suggesting that predicted performance was consistently estimated to be higher than the recorded performance. This is also apparent in Figure 3, where we see human performance nearly always worse than the model's posterior predictions at Test 1.1. The RMSE decreases over the events in the first session and increasingly normally distributed around zero, suggesting that the model's posteriors become less biased: Performance is overestimated about as often as it is underestimated.

Discussion

The focus of the current work is the prediction of future CPR performance over ecologically valid periods. After completion of the first three sessions, individual predictions have been made for CPR performance in the 4th session.

As can be seen in Figure 3, the priors do a generally poor job of representing the actual performance of participants in the early trials. This is a risk in generalizing parameters from one study to another. They are different samples, with participants in the previous study starting at and maintaining higher levels of proficiency. Given that participants in the study reported here started at a lower proficiency, it is to be expected that the prior distributions based on better performers would not predict worse performer data very well. However, most participants in both studies achieved and maintained higher levels of proficiency after several trials, so

this is bias-variance tradeoff we are willing to make in the interest of what we hope will be an improvement in predictive validity in Session 4. Additionally, the use of Bayesian Hierarchical Modeling as a method of parameter estimation provides posterior predictive distributions for each individual's learning profile. The use of prediction intervals allows for a quantification of certainty in our out-of-sample predictions.

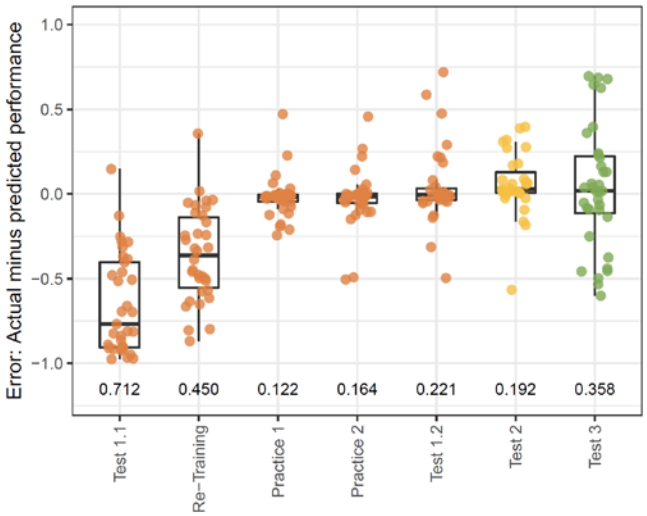


Figure 4. Prediction errors at each measurement event. Colors indicate the session; numbers at the bottom are the RMSE at each event.

The ERC's report (Soar et al., 2015) states that CPR performance is known to deteriorate within months of training and, therefore, even annual retraining might not be frequent enough for some people. Due to the fact that CPR training can be time consuming and optimal training intervals are currently unknown, they suggest that frequent "low dose" training using video instructions and hands-on practice can be as effective as instructor-led courses (Nolan et al., 2015). The work presented here confirms that there is a swift improvement in performance after such CPR retraining.

In summary, we report an experimental setup in which the learning and forgetting of CPR is assessed over ecologically relevant timeframes. We test a mathematical model's ability to predict future CPR performance using very sparse data. A first wave of predictions is presented here and an evaluation of the accuracy of those predictions will be presented at the conference.

Acknowledgments

This work was supported by EOARD grant #11926121 and by the Air Force Research Laboratory's 711 Human Performance Wing, Cognitive Models and Agents Branch.

References

Collins, M. G., Gluck, K. A., Walsh, M. M., & Krusmark, M. A. (2017). Using prior data to inform initial performance predictions on individual students, In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Jastrzembski, T., Walsh, M. M., Krusmark, M., Kardong-Edgren, S., Oermann, M., Dufour, K., . . . , & Stefanidis, D. (2017). Personalizing training to acquire and sustain competence through use of a cognitive model. In D. D. Schmorrow & C. M. Fidopiastis, *Augmented cognition. Enhancing cognition and behavior in complex environments* (pp. 148-161). Switzerland: Springer International Publishing AG.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Academic Press

Lee, M.D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Nolan, J. P., Hazinski, M. F., Aickin, R., Bhanji, F., Billi, J. E., Callaway, C. W., ... & Gent, L. M. (2015). Part 1: Executive Summary: 2015 International Consensus on Cardiopulmonary Resuscitation and Emergency Cardiovascular Care Science with Treatment Recommendations. *Resuscitation*, 95, e1-e31.

Sense, F., Behrens, F., Meijer, R. R., & van Rijn, H. (2016). An Individual's Rate of Forgetting Is Stable Over Time but Differs Across Materials. *Topics in Cognitive Science*, 8: 305–321.

Sense, F., Maaß, S., Gluck, K. A., & Van Rijn, H. (2019). Within-subject performance on a real-life, complex task and traditional lab experiments: Measures of word learning, Raven matrices, tapping, and CPR. *Journal of Cognition*, 2(1), 12. DOI: http://doi.org/10.5334/joc.65

Sense, F., & Van Rijn, H. (submitted). *Optimizing Within-Session Repetition Schedules with a Response-Latency-Based Cognitive Model Improves Retention*.

Soar, J., Nolan, J. P., Böttiger, B. W., Perkins, G. D., Lott, C., Carli, P., ... & Sunde, K. (2015). European Resuscitation Council guidelines for resuscitation 2015. *Resuscitation*, 95, 100-147.

Stross, J.K. (1983). Maintaining competency in advanced cardiac life support skills. *Journal of the American Medical Association*, 249(24), 3339–3341.

van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects. In A. Howes, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 110–115). Manchester, UK.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretical adequacy and applied potential of computational models of the spacing effect. *Cognitive Science, 42*(S3), 644-691. DOI: https://doi.org/10.1111/cogs.12602.

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