GENERALIZABILITY OF PREDICTIVE PERFORMANCE OPTIMIZER PREDICTIONS ACROSS LEARNING TASK TYPE

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science in Human Factors & I/O Psychology

By

HALEY PACE WILSON
B.S., United States Air Force Academy, 2013

2016
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

APRIL 18, 2016

I HERBY RECOMMENT THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Haley Pace Wilson ENTITLED Generalizability of Predictive Performance Optimizer Predictions Across Learning Task Type BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Human Factors & I/O Psychology.

Gary Burns, Ph.D.
Thesis Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on
Final Examination

Nathan Bowling, Ph.D.

Tiffany Jastrzembski, Ph.D.

Glenn Gunzelmann, Ph.D.

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean, Graduate School

ABSTRACT

Wilson, Haley Pace. M.S. Department of Psychology, Wright State University College of Science and Mathematics, 2016; Generalizability of Predictive Performance Optimizer Predictions Across Learning Task Type.

The purpose of my study is to understand the relationship of learning and forgetting rates estimated by a cognitive model at the level of the individual and overall task performance across similar learning tasks. Cognitive computational models are formal representations of theories that enable better understanding and prediction of dynamic human behavior in complex environments (Adner, Polos, Ryall, & Sorenson, 2009). The Predictive Performance Optimizer (PPO) is a cognitive model and training aid based in learning theory that tracks quantitative performance data and also makes predictions for future performance. It does so by estimating learning and decay rates for specific tasks and trainees. In this study, I used three learning tasks to assess individual performance and the model's potential to generalize parameters and retention interval predictions at the level of the individual and across similar-type tasks. The similar-type tasks were memory recall tasks and the different-type task was a spatial learning task. I hypothesized that the raw performance scores, PPO optimized parameter estimates, and PPO predictions for each individual would be similar for two learning tasks within the same type and different for the different type learning task. Fifty-eight participants completed four training sessions, each consisting of the three tasks. I used the PPO to assess performance on task, knowledge acquisition, learning, forgetting, and retention

over time. Additionally, I tested PPO generalizability by assessing fit when PPO optimized parameters for one task were applied to another. Results showed similarities in performance, PPO optimization trends, and predicted performance trends across similar task types, and differences for the different type task. As hypothesized, the results for PPO parameter generalizability and overall performance predictions were less distinct. Outcomes of this study suggest potential differences in learning and retention based on task-type designation and potential generalizability of PPO by accounting for these differences. This decreases the requirements for individual performance data on a specific task to determine training optimization scheduling.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Generalizability of Predictive Performance Optimizer Predictions Across Learning Task Type

Introduction

The objective of this research was to examine the consistency and generalizability of cognitive model parameters and predictions as well as individual performance based on type of task. Training is used to ensure personnel have the knowledge, skills, and tools necessary to operate in their specified work environment (Goldstein & Ford, 2002). Further, because skills and knowledge are shown to decay over periods of nonuse, retraining is necessary to ensure that all trainees are able to maintain and demonstrate a proficient level of skills or knowledge retention at any given time (Arthur, Bennett, Stanush, & McNelly, 1998). The importance of training and retraining will only continue to increase as modern technology expands the need for a skilled workforce (Arthur et al., 1998). This means that methods for enhanced efficiency of training programs and schedules will be increasingly important for the future workforce.

Efficient training programs provide a balance of proper resource allocation, such as time allocated to training, and adequate acquisition and retention of skills. For example, it is imperative that medical personnel maintain proper skill levels in order to avoid the detrimental consequences of forgetting such as poor patient outcomes and malpractice lawsuits (Wallace, Abella, & Becker, 2013). It is necessary to provide refresher training to medical personnel often enough to ensure skill maintenance over

time, while minimizing costs of training to the organization (e.g., time away from patients, cost of instructors, etc.).

Current training schedules within most organizations are calendar driven and do not account for the empirical and theoretical underpinnings of learning and forgetting. Researchers at the United States Air Force Research Laboratory argue that performance-based training schedules constructed around an individual's learning and retention rates may surpass the effectiveness of calendar-based training schedules (Jastrzembski, Gluck, & Gunzelmann, 2006). Specifically, such performance-based training systems would prescribe optimal training schedules based on individual learning needs.

The Predictive Performance Optimizer (PPO) is a cognitive modeling tool, rooted in cognitive theory of learning, that estimates an individual's learning and forgetting rates, predicts future performance on a specified task, and prescribes scheduling of training events based on historical and objective performance data (Jastrzembski, Gluck, & Rodgers, 2009). One limitation of PPO is that outcomes and predictions are only valid for the specific task and task procedures assessed in the input. Therefore, the predictions determined by PPO are not considered valid across different tasks. Understanding variance in learning and retention accounted for by task-type distinction could have implications for performance predictions and optimal schedule prescriptions across tasks.

In this study, I evaluated task performance as well as PPO optimized parameters and future performance predictions in one task and compared to results from other tasks. I also assessed PPO goodness of fit when optimized parameters from one task were

"transferred" to the other tasks. Low model prediction error would suggest potential generalizability of PPO parameter optimization and prescribed retention intervals across similar types of tasks. This finding would also suggest that type of task may be a valuable task descriptor for PPO task analysis, which could enhance the usability and generalizability of predictive performance across similar learning tasks. Below I provide a brief review of the literature on computational models of learning, theories of learning acquisition and forgetting, and model generalizability as a background to the experimental method.

Modeling Human Cognition

Cognitive models are formal, mathematical representations of complex, dynamic theories of thought and cognition (Boden, 2008). The models provide a framework to support precise, transparent, and consistent analysis and prediction of human behavior (Adner et al., 2009). Models can be simulated, or run using a given set or a distribution of parameters, and used to enhance understanding, test and evaluate assumptions, or make predictions (Vancouver & Weinhardt, 2012).

Computational Models of Learning

Cognitive modeling can help illustrate learning and forgetting for an individual or a group of individuals based on underlying mechanics of cognitive thought processing. Cognitive models are able to account for learning nuances that span human learning, and/or individual differences in learning and retention (e.g., a person's individual rate of learning and rate of forgetting). This information can provide decision makers with

empirical evidence of the advantages and disadvantages of a particular training schedule and provide insight into learning forgetting tendencies (Jastrzembski et al., 2006). A specified cognitive model of learning, applied uniformly in an organization and grounded in theory of human cognition, can be an invaluable tool for understanding and assessing training needs, methods, and schedules.

Theories of Learning and Forgetting

Humans are imperfect processors who forget or misremember things over time. Learning theories span decades of psychological research (e.g., Guthrie, 1952; Hull, 1943; Skinner, 1938; Thorndike, 1923; Tulving & Pearlstone, 1966). Cognitive theories of learning aim to assist understanding of the foundations of human learning to make predictions about human behavior (Soderstrom & Bjork, 2015). Despite more than 100 years of research on learning and forgetting, the mechanisms responsible for cognitive acquisition and degradation are not fully understood (Ebbinghaus, 1885; Cepeda, Pashler, Vul, Wixted, Rohrer, 2006). However, there are several variables that consistently account for variability of learning and retention across individuals. Such variables include practice, period of nonuse, spacing of training, type of task, task content, and individual differences (Arthur et al., 1998; Ericsson, Krampe & Tesch-Romer, 1993).

Learning and Performance. Soderstrom and Bjork (2015) define *learning* as "relatively permanent changes in behavior or knowledge that support long term retention and transfer" and *performance* as "temporary fluctuations in behavior or knowledge that can be observed and measured during or immediately after the acquisition process".

Researchers have attempted to consider this distinction when evaluating skill acquisition and decay. In a meta-analysis, Arthur, Bennett, Edens and Bell (2003) coded studies as evaluating the impact of training on learning, defined as typically paper-and-pencil or performance tests taking place outside of the work environment, or behavior, which they defined as measures of actual on-the-job performance. They concluded that overall, training had a similar impact on learning (d = .63) and behavioral (d = .62) criteria. Arthur et al.'s (1998) evaluation of skill decay suggests that learning criteria (d = -1.07) tend to degrade more over time than behavioral criteria (d = -.78). The researchers noted that training methods (including training environment and spacing), skill or task characteristic trained (task type), and evaluation criteria (learning versus performance), all relate to observable training outcomes (Arthur et al., 2003). Variation in training schedules and methods should aim to enhance both learning and performance on task. I expand on this finding and its implications in the cognitive model generalizability section of this paper.

Common Learning Phenomena

Generally, change in performance and learning initially occurs rapidly then at a diminishing rate with further practice (see Doyon & Benali, 2005; Newell & Rosenbloom, 1981). Factors that consistently affect learning and retention rates include amount of practice, elapsed time since practice, and the distribution of practice over time (Anderson, 1995). It is important for a cognitive model of learning to account for these consistent factors while maintaining enough flexibility to also account for individual

differences in learning. Although this is the normal pattern of learning, there are several training variables, including those described below, that affect learning patterns and learning retention post-training.

Spacing of training events. The *spacing effect* phenomenon reveals that distributed practice sessions lead to more stable and durable retention over time (Bahrick, 1979). Bahrick's research has shown that task performance increases quickly, but retention rates decrease when training intervals are narrowly spaced. Distributed training generally increases the amount of initial training required to achieve proficiency, but leads to substantial increases in retention rates over time (Bahrick, 1984). The spacing effect has been replicated in several studies (e.g., Peterson, Wampler, Kirkpatrick, and Saltzman, 1962; Bahrick & Phelps, 1984; Anderson, Fincham, & Douglass, 1999) consisting of various spacing intervals, complexity of tasks, and severity of massed versus spaced learning scenarios. A meta-analysis by Donovan and Radosevich (1999) quantitatively estimated that individuals in spaced practice conditions performed .46 standard deviations higher than individuals in massed practice conditions.

It is generally consistent across findings that spaced learning yields slower initial acquisition, but with lasting results and better long-term learning and retention. Findings related to the spacing effect suggest that both the spacing of training events (i.e., interstudy intervals) and retention intervals affect final-test performance (see Cepeda et al., 2006; Delaney, Verkoeijen & Spirgel, 2010 for review). Although the spacing of

training is held constant in my study, ability to track intervals between training sessions and retention intervals is an important quality of the model used for predictions.

Overlearning. *Overlearning* refers to training of a task that goes beyond the level needed to meet a level of proficiency (Arthur et al., 1998). For example, if it was determined that the required proficiency was 80%, repetition and practicing beyond that threshold would be overlearning. Driskell, Willis, and Copper (1992) conducted a meta-analysis and reported that overall, individuals engaging in overlearning performed at a level .63 standard deviations higher than individuals who did not engage in overlearning. Further analyses indicated that these effects were larger for cognitive tasks (d = .75) than for physical tasks (d = .44). Overlearning is considered one of the most important predictors of retention (Farr, 1987; Hurlock & Montague, 1982). Although overlearning is not the focus of this study, potential for overlearning exists and could affect model predictions.

Relearning. In 1979, Bahrick introduced a concept known as *successive relearning,* that refers to the effect on learning outcomes and long-term retention when training events included a combination of retrieval practice and retraining over time. Bahrick's study involved an initial learning session of a memory-recall paired associate task that participants completed until each pair had been correctly recalled. In subsequent sessions, participants relearned the items (retrained until each word pair was correctly identified). Findings showed a 56% and 83% increase in accuracy rates after two and five relearning sessions, respectively. Similar studies using memory-recall tasks within

educational environments have shown substantial effects of successive relearning on retention (up to 68% one month post-training and 49% four months post-training, as compared with around 11% in a baseline control condition; Rawson & Dunlosky, 2011; Rawson & Dunlosky, 2012; Rawson, Dunlosky, & Sciarteli, 2013).

Assessing retention using overall recall accuracy may reveal a desensitized measure of memory (Krinsky & Nelson, 1985). Research has shown that performance scores may suffer due to knowledge or skill retention under the threshold of recall, but are not necessarily indicative of complete forgetting (Bahrick, 1967; Kornell, Bjork, & Garcia, 2011). In this context, relearning partially forgotten items after a period of nonuse can be considerably faster compared to original learning (Bahrick, 1967; Ebbinghaus, 1885) or to the learning of a new set of similar items (MacLeod & Dunbar, 1988; Nelson, 1978). Additionally, an increased number of repetitions are required to reach criterion performance for new items compared to forgotten items that had been previously studied or tested (de Jonge, Tabbers, & Rikers, 2014). The current study captures relearning through successive training events consisting of identical training material. Thus, I expect to see effects of relearning within subsequent training sessions and for the cognitive model of learning to capture this type of learning nuance.

Forgetting

Generally, forgetting initially occurs rapidly during a period of nonuse, and then at a diminishing rate over time (Anderson & Schooler, 1991; Rubin & Wenzel, 1996; Wixted & Ebbesen, 1997). In a meta-analysis of skill-decay, Arthur et al. (1998)

identified several factors related to skill decay, such as length of nonuse and the type of task, that typically vary in training processes. Researchers and practitioners often use the term skill decay to explain forgetting in a training environment. The term *decay* describes the outcome or observed decrement in performance on learned knowledge or skills after a period of nonuse (Arthur et al., 1998).

The period following training in which the skill or knowledge learned is not used is referred to as the *nonuse interval.* Periods of nonuse occur in the workplace when individuals are not required to use the knowledge or skills acquired at initial training for extended periods of time. The length of this period of nonuse can substantially affect measures of learning and performance over time. For instance, research has shown that, due to memory consolidation processes, performance on psychomotor skills may significantly improve within hours post-training, such as 24 hours after the training session is complete (Dorfberger, Adi-Japha, & Karni, 2007; Savion-Lemieux, Bailey, & Penhune, 2009). However, according to Bjork's theory of disuse, long nonuse intervals typically reflect the opposite effect due to forgetting (Bjork & Bjork, 1992; Savion-Lemieux & Penhune, 2005).

Arthur et al. (1998) estimated that the effect of the nonuse interval on skill decay effect size was d = -.51, with longer intervals leading to increased skill decay. A recent meta-analysis by Wang, Day, Kowollik, Schuelke, and Hughes (2013) found similar effects, d = -.58. A closer look at Wang et al.'s (2013) results indicate that periods of nonuse less than a day resulted in almost no loss of performance (d = -.08) while nonuse

that lasted from 1 to 7 days resulted in nearly a half a standard deviation decrease in performance (d = -.42).

In contrast, some research has shown that situations indicative of forgetting (e.g., training delays, periods of nonuse, or variation of context, content, environment, or retrieval practice conditions) aid learning by creating desirable difficulties (Bjork, 1994; Bjork, 2013; Storm, Bjork & Bjork, 2005). This phenomena can be explained by the suggestion that moderate difficulties in retrieval require active cognitive processes, which connect knowledge elements that already exists in long term memory to those needed to solve a particular problem (Soderstrom & Bjork, 2015).

Retention varies as a function of learning and forgetting depending on several variables such as type of task, complexity of task, length of nonuse, and individual differences (Arthur et al., 2003). Results from forgetting research show a need for a forgetting parameter in cognitive models. Accounting for forgetting, in addition to learning, is necessary to ensure accurate and precise performance predictions based on an individual's learning consistencies.

Individual Differences in Learning

Individual differences in learning and forgetting contribute to variance in task performance and learning between individuals or groups. Several variables have been identified as important predictors of individual differences in learning and skill retention: general mental abilities, primary cognitive abilities (e.g., perceptual speed, processing speed, working memory capacity, reasoning ability, verbal ability, spatial ability),

cognitive controls, cognitive styles, learning styles, personality, motivation, and prior knowledge or experience (see Jonassen & Grabowski, 1993). These are examples of individual variances that cognitive models may be able to capture by tracking and varying based on individual performance over time.

Learning and memory research attributes between-subject variance in learning to initial level of knowledge, learning rate, and asymptotic performance (Meredith & Tisak, 1990; Rast, 2011; Kutrz, Mogle, Sliwinski, & Hofer, 2012; Rast & Zimprich, 2009). Additionally, a recent study by Bayliss and Jarrold (2015) demonstrated variance in individual forgetting rates explained by working memory performance over and above the variance explained by measures of individual processing rates and long-term memory storage capacity. An individual with strong working memory skills and high cognitive ability (i.e., higher learning rates and lower forgetting rates) will likely retain knowledge for longer periods of time. This individual would therefore require retraining less frequently than an individual with poor working memory skills or cognitive ability. Learning and retention rates affected by these types of individual variation drive the need for individually tailored training schedules, refresher training, and other interventions designed to ensure that skill retention is maintained over time. The within-subject repeated measure design of the current study aimed to capture individual differences in learning in order to assess differences in learning, forgetting and retention based on task type.

Modeling Individual Differences. There are several approaches to modeling individual differences including evaluating the model against data aggregated across all subjects, from each subject separately, or groups of subjects (Ashby, Maddox, & Lee, 1994; Lee & Webb, 2005; Nosofsky, 1986; Wixted & Ebbesen, 1997). Lee and Webb (2005) describe an approach for applying the same computational model to individuals, while allowing the model's parameters, which relate to the speed or effectiveness of cognitive processes, to vary across individuals. Their findings showed superior predictive value, enhanced understanding, and ability to account for variation in individual differences through interpretable differences in model parameterization. Based on this finding, many models of learning now account for individual differences by fitting algorithm parameters to quantitative data for individual subjects (e.g., Predictive Performance Optimizer; Jastrzembski et al., 2006; Jastrzembski et al., 2009).

Predictive Performance Optimizer

To understand the utility of training schedules, content, and manipulations, as well as to predict future performance post-training, one must consider training outcomes in terms of acquisition, retention, and transfer (Arthur, Day, Bennett, Portrey, 2013). One tool that attempts to assess the effectiveness of training is the Predictive Performance Optimizer (PPO), which "relates to predictive optimization of performance for a domain and, more particularly, to a cognitive tool aimed at tracking and predicting human performance for purposes of optimizing performance around a specified goal" (Jastrzembski, Rogers, Gluck, & Krusmark, 2013). PPO is a cognitive tool that tracks

historical and objective performance data and training schedules to identify individual or group learning and decay regularities. PPO output allows users to examine learning rates, forgetting rates, and projected future performances for individuals or groups.

This empirical evidence can assist decisions regarding training schedules (Jastrzembski et al., 2006). PPO validation across declarative and procedural knowledge and skills training makes it an appealing tool to assist optimization of training scheduling in expensive and complex areas, such as those relevant to the military, medicine, and education (Jastrzembski et al., 2009). However, a limitation of the software is that PPO predictions are task input specific and apply only to the task and data obtained in the historical performance. Thus, the performance predictions, though valid for the task used for model calibration, have not been assessed in terms of their ability to generalize to other similar tasks. In other words, PPO software can make performance predictions for a wide variety of tasks when appropriate and specific input is provided, but the predictions and estimations of learning and decay may not generalize to similar training content or tasks.

Predictive Performance Equation

The Predictive Performance Equation (PPE) is the underlying mathematical model and theory of PPO. Researchers designed the algorithm to capture effects of recency, frequency, and spacing of training as well as individual differences in learning and decay rates (Jastrzembski et al., 2006). The model is derived from the general performance equation (see Anderson & Schunn, 2000 for a description of the general

performance equation) but includes an additional mechanism to capture the effects of spacing on retention. The inclusion of a spacing effect term in the algorithm allows the model to account for massed or distributed training sessions, which affect learning and retention (Bahrick & Phelps, 1987; Jastrzembski et al., 2006; Walsh, Gluck, Gunzelmann, Jastrzembski, & Krusmark, *in preparation*).

PPE contains a total of four free parameters. These can be fit to the performance data for each participant, enabling individualized predictions and prescriptions. Equation 1 represents a general form of the PPE. The *Performance* term is the level of performance. The *Practice* term equals the amount of practice accumulated in training. *Time* is the period of nonuse, or time since practice occurred. These terms account for the training program and schedule.

$$Performance = Practice^{c} * Time^{-d}$$

(Equation 1)

In this equation, c is the learning rate (fixed to 0.1), and d is the decay rate. Decay is calculated based on the distribution of practice over time, or lags, between successive practice opportunities.

$$d = decay\ intercept + decay\ slope * (1/\log(lags))$$

(Equation 2)

Larger lags reduce the value of decay expressed in Equation 2, and thus lead to better retention. The two free parameters in Equation 2, *decay intercept* and *decay slope*, correspond to an individual's overall level of forgetting and their susceptibility to the

spacing effect, respectively. The model's output, *Performance* in Equation 1, is scaled to the actual performance scores recorded in a particular task using a logistic function. The logistic function includes two additional free parameters, a *logistic intercept* and a *logistic slope*. These affect how level of performance map onto the specific performance measure for a given task.

The model's four free parameters, *decay slope, decay intercept, logistic slope,* and *logistic intercept*, are estimated separately for each individual using their historical performance data. This approach to parameter estimation allows for model calibrations and predictions based on individual differences that drive the individualized training recommendations or prescriptions of PPO output.

PPO Process

PPO input requires objective, quantitative, and historical training data for an individual, group, or team. The tool then calibrates model parameters to the data, tracks learning and forgetting, predicts future task performance, and then prescribes optimal training schedules to sustain future performance (Jastrzembski et al., 2006). Figure 1 illustrates the process of PPO input and output. The first step is to calibrate model parameters by fitting a learner's training history. Next, the PPE extrapolates a learner's unique learning regularities to make precise, quantifiable predictions of performance at specific points in time. Finally, analyses of predictions are tailored based on defined training regimens, objectives, or optimization goals.

PPO output is displayed in a visual and intuitive manner to allow decision makers to track predictions and produce recommendations for training. Further, users can tailor output parameters to investigate future training requirements for an individual or to restructure training schedules to best meet training needs based on predictions. For instance, if the main goal of a training intervention is to increase skill proficiency retention, decision makers can use the outcome of PPO to recommend a distributed training schedule that best supports this requirement.

PPO outcomes enable decision makers to structure training based on individual need for refresher training as opposed to a one-size-fits-all training program. The intentions of using PPO software for training prescriptions are to avoid overtraining or undertraining a given individual, thus improving training efficiency and reducing costs associated with a long-term training strategy without compromising performance. Figure 2 displays an example of PPO output data.

Validation and Verification of PPE

Since initial development of PPE, researchers of the Human Performance Wing at the Air Force Research Laboratory have extensively validated the model across a variety of domains and contexts. To best capture precise and accurate predictions of future performance, a computational cognitive model must account for known effects of training variables on learning, forgetting and retention including amount of practice, number of re-learning sessions, spacing between practice or the scheduling structure of training, periods of nonuse, and individual differences. PPE model predictions have been tested

and validated using data from many published studies in the areas of education and training. In addition to evaluating the theoretical adequacy of the model, Walsh et al. (*in preparation*) assessed PPE's applied potential; that is, its suitability for application in real-world training and education.

Initial validation efforts involved simulation studies based on representative publications on learning and memory from the psychological literature. These simulations showed that PPE could account for a wide range of training variables including the role of spacing on acquisition and retention (Bregman, 1967), the interaction between spacing and the length of the retention intervals (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Young, 1971), the interaction between spacing and the amount of practice (Cepeda et al., 2008; Pavlik & Anderson, 2005), relearning (Rawson, Dunlosky, 2013), and overlearning due to repetition (Begg & Green, 1988; Benjamin & Tullis, 2010). PPE parameter estimates were largely consistent across experiments of varying content, amount of practice, spacing manipulations, and duration. The simulations demonstrated the basic theoretical adequacy of PPE (Walsh et al., *in preparation)*.

Researchers also evaluated whether PPE could be applied beyond simple laboratory experiments and to increasingly complex scenarios more representative of military training and education. Researchers demonstrated PPE's ability to operate on educationally relevant time schedules ranging from days to years, make precise predictions of future performance and valid prescriptions of refresher training, and to

capture performance in a variety of tasks, contexts, performance measures and schedules (Walsh et al., *in preparation*). Collectively, these simulations have demonstrated utility of PPE predictions in declarative, procedural, and hybrid contexts. However a direct comparison of performance measures and PPE parameter estimates across task has yet to be assessed.

Limitation of PPE

A considerable limitation of PPE is the inability to generalize predictions to tasks outside of the specified input criteria. The output is considered valid for the specific task, quantitative performance data and training methods used. Therefore, performance predictions for a specific task are appropriate only to future performance for that exact task and training regimen. This limitation inhibits the utility and generalizability of PPE because performance data is required for each learner, learning requirement, or other training variable of interest.

Assumptions of PPE are that the collection and assessment of training scores are consistent across training sessions, as well as in its predictions of future performance. PPE models predicted outcomes but does not necessarily translate to understanding of the cognitive ability or general learning and decay rates of an individual. Instead, predictions are specific to an individual's performance on a specified task and for a specified measure of interest within that task. A better understanding of the generalizability of PPE optimized parameters and individual differences in learning and retention performance as a function of task type is required to determine the potential

generalizability of PPE predictions across learning tasks. The focus of this study is to begin evaluating of PPE's ability to account for similarities and divergences in learning and retention as a function of task type (Arthur et al., 1998; Reber, 1989).

Transfer and Model Generalizability

Transfer of training refers to the transfer of knowledge and skills learned in the training environment to the work environment (Baldwin & Ford, 1988), or the generalization of training performance across various contexts (Schmidt & Bjork, 1992). Thorndike (1923) concluded that previously learned material only assists future learning to the extent that learning principles overlap, or contain elements identical to those involved in the learning acquisition. Albeit, the extent of similarity between learning principles that is required is not fully understood. Although it is impractical to say that an individual will be able to transfer learned material from one memory-recall task to another memory-recall task that they have not yet learned, it may be reasonable to test for similarities in learning trends at the level of the individual. To assess this question, I will test performance and model predictions to determine if the type of task (e.g., memory-recall task) accounts for a significant amount of variance individual learning and retention.

Analogical transfer studies involve training of one task and assessing subsequent performance on a novel analogical task (Barnett & Ceci, 2002). Results for stability and presence of learning transfer are mixed including evidence of transfer (e.g., Gick & Holyoak, 1980), lack of transfer (e.g., Reed, Ernst, & Banerji, 1974), negative transfer

(e.g., Woodworth & Schlosberg, 1954), overtransfer (e.g., Halpern, Hansen & Riefer, 1990) and uncertain results (Brown, Kane, & Long, 1989). Formal discipline transfer describes the hypothesis that transfer likelihood depends on the knowledge domain of the trained and tested skill, though validity in evaluations of this phenomenon are uncertain (Lehman, Lempert, & Nisbett, 1988).

Barnett and Ceci (2002) created a taxonomy of transferability to identify dimensions along which the likelihood of transfer varies: a) content or nature of the skill and performance change measures, and b) context, or distance between the trained skill and target skill based on the knowledge domain of the task, physical, temporal, functional or social context of the training, and modality of training. The content factor can be broken down into three dimensions: specificity or generality of the learned skill, performance change measures, and memory demands of the tasks (Barnett & Ceci, 2002). Critical conditions for transfer, such as stability of learning within a learning domain, remain uncertain.

Extensive research involving transfer of simple word-associative learning tasks and visuospatial learning tasks also demonstrates mixed reviews on stability of learning transfer (e.g., Conway, Kane, Bunting, Hambrick, Wilhelm, & Engle, 2005; Uttal, Meadow, Tipton, Hand, Alden, Warren, & Newcombe, 2013). A recent study by Jaeggi, Buschkuehl, Shah and Jonides (2014) linked task specific spatial task performance to a general composite score representing five visuospatial reasoning measures, indicating a visuospatial reasoning ability that is not task specific. The same procedures were used to

assess verbal-reasoning tasks, but reliability in performance measures were far less than that of the visuospatial tasks.

It is important to consider the literature in training transfer when determining a theoretical possibility of PPE model generalizability. Tying the principle of transfer to cognitive modeling, models gain weight when its principle predictions generalize to more complex, or different environments (Sanders, 1998). The value, generalizability, and applicability of a cognitive model is determined by maximum goodness of fit (to quantitative data measurements) with minimal model complexity (Pitt, Myung, & Zhang, 2002). Validity of model predictions across tasks of the same type implies a greater sense of generalizability and substantially increases the utility of the model. If the model, calibrated using performance data from one task, adequately predicted performance on another related task, it would suggest that the mental processes approximated by model parameter estimates were invariant within an individual and across related tasks. Such model generalization would not be expected across unrelated tasks that evoked different mental processes, however, and perhaps inform how variables within the context affect learning. In this study, I control variation of training context (all three tasks will be assessed in the same location, time, modality, etc.), to better understand the generalizations of performance based on task context and type to assess the extent to which type of task affects learning and retention, and model fits/prediction over time.

Task Characteristics

The literature demonstrates that different types of knowledge and skills decay at different rates (e.g., Arthur et. al., 1998). Understanding task type and the type of knowledge and/or skills that must be retained to competently perform a task to proficiency is necessary for training program and schedule design (MacLean & Cahillane, 2015). Farina & Wheaton (1973) developed a task characteristic approach to classify tasks and improve generalization of research results about human performance. They found several correlations between major components of a task, which were identified and treated as categories and performance measures. More recent findings demonstrated significant effects of task categorization on learning outcomes (Arthur et al., 1998; Arthur et al., 2003). Results suggest that task type may be a useful quality of training and predictor of performance outcomes.

Learning Domain. Learning domains are typically designated into three categories of learning: cognitive, affective, and psychomotor (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). The three domains of learning are divided into subsets and arranged hierarchically, ranked from simple to complex forms. The affective domain pertains to feelings and emotions. The psychomotor domain includes psychomotor behaviors, skills, or actions that include a physical-psychological interaction and can be measured in terms of quantitative values such as speed, precision, distance, or execution technique. The cognitive domain pertains to thought processing and consists of six

subsets: knowledge, comprehension, application, analysis, synthesis, and evaluation (Bloom et al., 1956).

Anderson and Krathwohl (2001) rearranged the cognitive domain taxonomy to illustrate the interaction of intellectual abilities and types of knowledge: remembering, understanding, applying, analyzing, evaluating, and creating. Skills and abilities within the cognitive domain have a direct interaction with knowledge. Types of knowledge were defined as factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge (Anderson & Krathwohl, 2001). The taxonomy approach to learning domains is indicative of the assumption that the abilities and skills develop and build upon each other through advancement from simple to complex forms of understanding (Bloom et al., 1956).

The domain of the task being trained is related to both acquisition of the task and skill decay. In a meta-analysis of the effectiveness of training in organizations researchers found that individuals learned psychomotor tasks (d = .80) more than they did cognitive tasks (d = .58) during training (Arthur et al., 2003). Arthur et al. (1998) also reported that the level of skill decay was less for physical tasks (d = -.76) than for cognitive tasks (d = -1.18) across all retention levels. In this study, I will focus on the cognitive domain and manipulate task type.

Type of Task. The type of task, within a similar learning domain, is determined by other features of the task. For example, declarative and procedural knowledge are associated with different learning algorithms, memory representations, and brain

networks (Ritter, Baxter, Kim, & Srinivasmurthy, 2013). In this study I focus on three tasks from the cognitive domain. Two are paired-associate, memory-recall tasks, and one is a cardinal direction, spatial learning task. The two memory-recall tasks are intended to capture the same memory capabilities.

A cognitive task analysis is used to model the cognitive process that a learner adopts when he/she performs a certain task (Jonassen, Tessmer, & Hannum, 1999). That is, the cognitive task analysis is an aid to identify and analyze cognitive processes that underlie performance of tasks in consideration of observable behavior (Carlisle, 1986). One categorical principle of a task analysis is a declarative knowledge versus procedural knowledge classification. Declarative knowledge includes facts, or information about a task (i.e., explicit knowledge). Procedural knowledge refers to knowing the actions required for the execution of a task and how to carry them out (i.e., tacit knowledge). Declarative knowledge is generally reliant on working memory, while procedural knowledge performance becomes increasingly automatic with time (Ritter et al., 2013). However, some learning nuances, such as the finding that additional practice increases performance at a diminishing rate over time, apply to learning acquisition of declarative knowledge and procedural skills (Ritter et al., 2013).

In the current study, two memory-recall tasks are used. These tasks align with declarative memory and require use of working memory and memory storage and processing over time. The cardinal direction task is a skill-based and procedural task that requires mental rotation cognitive processing. It is less important that a person

performing this task remembers the location of each trial, but that they learn a strategy to perform the task.

Current Study

In tying the review of transfer of training and model generalizability to cognitive modeling of learning, the question becomes: if two tasks exist in the same task domain and type, can PPO parameter estimates generalize across the two tasks? Determinations of stability of learning and retention based on type of task has implications for training predictions in education and training (e.g., Klausmeier, 1961), cognitive modeling (e.g., Singley & Anderson, 1989), and understanding of domain-specificity of expertise (e.g. Glaser, Chi, & Farr, 1988; Ericsson & Smith, 1991). Despite the implications, there are currently no published studies on potential generalizability of PPO predictions across similar tasks.

The underlying mathematical model of the system does not account for generalizability of an individual's learning and retention rates or cognitive processes associated with task type. The purpose of this study is to understand the relationship of task performance, PPE parameter estimates, and PPO predictions across learning tasks to diversify applications of PPO. Similarities or variance in PPO outcomes across tasks will inform the potential of PPO prediction generalizations across tasks. I will assess performance and PPO measures for two similar type tasks (from the same learning domain) using two memory-recall tasks and one different type task (from a different learning domain) using a cardinal direction task. I predict that all measures will be more

similar across the two similar type-tasks and more different when compared to different types of learning tasks.

Hypotheses

Hypothesis 1. Raw individual performance measures will be similar between the two memory-recall tasks.

Hypothesis 2a. PPO optimized parameters of one memory-recall task will be positively correlated to PPO optimized parameters of the other memory-recall task, but there will be a weaker relationship to the predictions of the other task.

Hypothesis 2b. Optimized parameters from one memory-recall task, when applied to the other memory-recall task, will produce a similar goodness of fit.

Hypothesis 3. PPO predicted retention intervals will significantly correlate for the two memory-recall tasks, but not for the spatial learning task.

Method

Participants and Design

Participants were required to have normal/corrected vision and normal cognitive function to participate in this study. Participants who expressed proficiency with the Japanese language were excluded from participation because Japanese-English word pairs were included in one of the memory-recall tasks. A power analysis was conducted to determine a required sample size of about 50 students to have a power level of .80 for this within-subject, repeated measures design assuming a small to moderate effect size (f = .175) and moderate correlations between repeated measures (r = .50). It should be noted that this required sample size decreases as the expected correlation between repeated measures is increased; .50 was set to provide a more conservative estimate.

I recruited 83 participants from the Wright State University Psychology Department for this study. There were 35 males (42.2%) and 48 females (57.8%) between the ages of 18 and 36 years (M = 20.45, SD = 3.75). A majority of the participants were recruited using the department's human-subjects management system, SONA. Introductory psychology students earn credits on SONA to complete course requirements or earn extra credit. Participants were required to sign up for all four, one hour sessions, to participate in this study. Any participants signing up through SONA received one credit for each 30 minutes of participation, totaling 8 credits. Participants not signed up through SONA were recruited via fliers within the department or via word of mouth. As

an extra incentive to attend all four sessions, participants with four session completions were placed in a drawing for one of twelve $50 gift cards.

There were several computer issues during data collection and the data of 23 participants was corrupted. Two individuals did not complete all four sessions. Of the recruited participants, 58 completed all four sessions with uncorrupted data files. These participants were used in data analysis to minimize error variance. This pool included 27 males (46.6%) and 31 females (53.4%) ages 18-32 (M = 20.07, SD = 2.74). Of these participants, 10.3% were left handed and 89.7% were right handed. Participant ethnicity identification was as follows: 63.8% Caucasian, 22.4% African American, 5.2% Asian, 6.9% Latin American, 1.7% Other.

Procedure

Participants were informed that the study required participation in four training sessions consisting of four learning tasks each: one session per day for three days and one session one week after their third training session. The first three sessions comprised the learning acquisition phase, and the fourth session comprised the retention phase (see Figure 3).

Participants were scheduled for all four sessions and were asked to schedule themselves for the same time in each session when possible. This would result in a 24 hour (+/- six hours) interval between the first three learning acquisition sessions and a one week, or 168 hour (+/- six hours), retention interval between sessions three and four. Prior to study enrollment, participants were asked to read the informed consent document

(ICD). If they agreed to the terms, they were instructed to sign the ICD. By agreeing to the ICD, participants agreed to the four-session time commitment required by the study. Once enrolled in the study, participants were asked to complete a demographics questionnaire and were provided with additional information and instructions for the four tasks.

Each session consisted of four successive learning tasks: two memory-recall tasks, one spatial recognition task, and one procedural learning task. The procedural learning task was not used for data analysis because of technical issues with the server for this task. In general, session one lasted about one hour and each subsequent session was completed quicker. Session four required approximately 30 minutes to complete. All four tasks were run in succession using a MATLAB program on a standard desktop computer. The study took place in a computer lab with eight computer work-stations consisting of a desk, computer tower, monitor, mouse and keyboard. Dividers were placed between each station to reduce distractions from other participants during the study.

Input for the PPE consisted of participant data from all four sessions. The PPE was able to account for close spacing intervals (in the acquisition phase) as well as a longer interval (between session 3 and 4, the retention interval) to better estimate individualized, optimized parameters of the model based on fluctuation in performance over time, and predict future performance on task. I will use these parameter estimates to compare tasks. Overall performance on task, PPE optimized parameters and model fit,

and PPO prescribed retention intervals were evaluated and compared across the two memory-recall tasks and the spatial learning task.

Session procedures. Participants were asked to sign in and provided with instructions for the session (instructions remained the same for all four sessions). Each session consisted of four learning tasks that appeared in a randomized order. The next task began as soon as the previous task was complete. Participants were instructed to sit at their assigned station, read the instructions on the screen and press the spacebar to start the task. Figure 4 shows an example of an instruction screen that appeared on each workstation monitor. Formatting of instructions was similar for all four tasks.

After all four tasks were completed, the session ended and the participant was escorted out of the laboratory. Participants were instructed not to practice the tasks during the time between sessions, though the novel nature of the tasks would make this difficult and highly improbable. Participants received an appointment reminder email each evening before his or her scheduled return session.

Session variation. All four sessions were nearly identical although subsequent sessions were generally quicker as a result of learning through repetition of the task (practice). For each of the four sessions, participants were required to complete four learning tasks at their computer. During session one, the first trial for both memory-recall tasks displayed the stimulus and response pair. In subsequent trials, only the stimulus appeared, prompting the participant to recall and input the correct paired response. The stimulus and response pair did not appear together on the screen in the

other sessions, except in the feedback following the response input. In all sessions, the participant received feedback to their response as well as the correct response. PPO data analysis occurred once participants completed all four sessions.

Task Description

Task 1. Task 1 was a memory-recall task based on learning English-Japanese paired associates used by Pavlik & Anderson (2005). Japanese was chosen to minimize the prior learning participants could bring to the task. Translations are in English characters rather than Japanese symbols. In the original use of the task, stimuli and responses were 104 Japanese–English associate word-pairs. Only four-letter English words, and four-letter to seven-letter Japanese translations, were used for this task (Pavlik & Anderson, 2005). The modified version of the task used for this study consisted of 10 Japanese–English associate word-pairs each shown ten times throughout the session.

For the first trial in session one, both stimulus and response of the word-pair were displayed centered on the screen, with the Japanese recognition word on top and the English recall word on bottom. Each pairing remained onscreen for two seconds. In all subsequent trials, the Japanese recognition word appeared and the participant was asked to respond with the English recall word as quickly and accurately as possible. The typed answers were displayed on the screen directly below the stimulus word. The stimulus word was shown for two seconds and the participant was allotted five seconds to respond. If the participant responded correctly, a green smiley face appeared on the screen, along

with the correct response. An incorrect or failed response elicited a red face with a frown, along with the correct response displayed below to facilitate relearning. The feedback and correct response remained on the screen for two seconds regardless of whether the participant responded correctly or incorrectly. The participant was instructed to consider this feedback to aid future responses. Participants who completed 126 trials of the Japanese-English task in the first session and 112 trials in each of the three subsequent sessions were included in the data analysis. This discrepancy in trial numbers from Session 1 to the other three sessions was due to the simultaneous appearance of the Japanese stimulus and English word pair in the first session. Performance and model fits were assessed based on accuracy scores.

Task 2. Task 2 was a memory-recall task, called the Digit-Droodle task, that used picture to number paired-associates. The pictures used were a normative set of nonsensical images also known as "droodles" developed by Nishimoto, Ueda and Miyawaki (2010). The droodle images were paired with a two-digit number for memorization and recall. Digits and symbol pairings were chosen to eliminate prior learning confounds participants could bring to the task. The original task consisted of 196 droodles of similar complexity and distinction. The version used for this study included 10 digit-droodle pairs. When a pair was presented for the first time, the droodle (recognition stimulus) and corresponding number (recall response) appeared simultaneously on the computer monitor, with the droodle on top and the digit on the bottom. Participants typed the number using a standard keyboard, and they received

positive feedback (Figure 4). During each subsequent presentation, only the droodle appeared. Participants were instructed to try to recall the corresponding digit and to type it. After they responded (or after seven seconds passed), positive or negative feedback in the form of a smiling or frowning face appeared (same as the first task). The correct response appeared below the droodle. Each pairing remained on screen for two seconds and the participant had five seconds to respond. The feedback and correct response remained on the screen for two seconds regardless of whether the participant responded correctly or incorrectly. Participants who completed 126 trials of the Digit-Droodle task in the first session and 112 trials in each of the three subsequent sessions were included in the data analysis, same as in the Japanese-English task. Performance and model fits were assessed based on accuracy scores.

Task 3. Task 3 was a cardinal direction (spatial) learning task (Gunzelmann, Anderson, & Douglass, 2004). In this task, participants were shown a sequence of static image pairs that depicted a spatial relationship between a viewer and a target in two different frames of reference. The left side of the screen represented a target field as viewed from an ego-oriented perspective with red circle representing the target. The right side of the screen showed the target field from an allocentric orientation. For each pair of images (a trial) participants were asked to determine the location of the target in the allocentrically oriented perpective based on the perspective depicted in the visual scene on the left. Participants responded by pressing the corresponding key on a numeric

keypad. After a response, participants received visual feedback. The task ended once the participant had responded correctly to each of the 64 image pairs.

Participants who successfully completed all four sessions were included in the data analysis for the Cardinal Direction, spatial learning task. This task consisted of a drop-out design in which each participant was required to complete all 64 configurations correctly before completing the task. All 64 configurations were presented sequentially and then false responses were repeated until all configurations were answered correctly. This process allowed me to warn against and check for careless responding. Performance was assessed based on accuracy for the first 64 responses in order to achieve a comparable measure of performance to the Japanese-English and Digit-Droodle tasks. The instructions for this task are shown in Figure 5 below.

Measures

Task performance. Task performance was measured in terms of accuracy (percentage correct). In the Cardinal Direction task, a correct response for each of the 64 configurations was required in order for the participant to move on. Because of this, accuracy for only the first 64 trials was evaluated. The accuracy measures for each task act as an indicator of learning and enhanced performance over time and can be compared across all three tasks.

PPE parameter estimates. To calculate performance trajectories and predicted decay rates, the PPE produces four optimized parameters for each participant. These are the logistic function intercept, logistic function slope, decay intercept, and decay slope.

Parameters were compared across tasks. Additionally, optimized parameters from one task were applied to each of the other tasks (per individual) to determine generalizability of parameters across tasks.

PPE fit. I used mean standard error (MSE) and correlation (*r*) to evaluate model fit. The goodness of fit measures PPE's parameter estimates in comparison to observed performance scores for each individual per task and were averaged across participants to provide an overall estimate of model fit. An r value of 1 and a MSE value of 0 would indicate that the model fit perfectly.

PPO predictions. For validation, PPO predictions for the fourth session were made from the first three sessions. PPO predictions of future performance were estimated per day for 365 days. I compared performance across tasks at 14, 30, 60, and 120 days post-training.

Results

Descriptive Statistics

Because PPO predictions require three inputs and I needed to assess the correct retention predictions for PPO against actual performance data, I removed data from the analysis for the 2 participants who failed to complete all four training sessions on time (+/- six hours). I also removed data for the 23 participants who experienced technical problems with their computer as to ensure accurate performance measures. Data was analyzed from 58 participants who successfully completed all trials and sessions for all

three tasks. Means, standard deviations, and overall performance scores for all three tasks across all four sessions can be found in Table 1.

Initial PPE Fit and Validation

PPE model fit to performance data was validated for both same-type tasks and then assessed for fit within different-type task. I assessed the adequacy of the model's fit to actual performance data from the first three sessions (the knowledge acquisition phase), and the adequacy of its predictions during the fourth session (retention session). For both memory-recall tasks, high correlation and low MSE measures are seen in the acquisition phase (Sessions 1-3). For the Japanese-English task, the metrics of fit for the knowledge acquisition phase were: r = .998, MSE = 4.85e-04; r = .986, MSE = 2.52e-04; r = .971, MSE = 1.27e-04. For the Digit-Droodle task, the metrics of fit for the knowledge acquisition phase were: r = .997, MSE = 3.46e-04; r = .978, MSE = 4.56e-04; r = .965, MSE = 1.32e-04. Model predictions for the retention phase, which data were not included during model calibration, were consistent with actual performance one week later for both tasks: r = .981, MSE = 6.30e-05 for the Japanese-English task and r = .995, MSE = 5.25e-04 for the Digit-Droodle task.

For the Cardinal Direction task, the metrics of fit for the knowledge acquisition phase were: r = .66, MSE = .011; r = .16, MSE = .0016; r = .00, MSE = .0019. The model seemed to do better for Japanese-English and Digit-Droodle tasks. This is likely because the PPE must predict some forgetting and there was little variance in accuracy scores at the individual level. Although this negatively affected the correlations for this

task, the mean standard error between model fit and actual performance was low. Regardless of this trend, the model predicted performance for the retention phase with very little error, $r = .05$, MSE = .0016.

It is important to note that model fit for the retention phase is computed within person across the session and averaged across participants. Another method of validating the PPE predictions is to correlate the predicted overall accuracy for the retention phase with participants actual overall accuracy in the retention phase. This resulted in correlations of .77 ($p < .01$), .59 ($p < .01$), and .66 ($p < .01$) for the Japanese-English, Digit-Droodle, and Cardinal Direction tasks, respectively. This provides validation data of the PPE predictions within the current study.

Task Performance and Overall Model Fit

Japanese-English Task. Figure 6 illustrates average accuracy for each session (see also Table 1). Correcting for violations in sphericity with the Greenhouse-Geisser adjustment to degrees of freedom, results demonstrated learning over time, F (1.43,81.63) = 194.02, $p < .01$. Accuracy increased in session 2, $F(1,57) = 234.68$, $p < .01$, and session 3, $F(1,57) = 41.42$, $p < .01$), from the previous session but plateaued between sessions 3 and 4 (performance ceiling), which demonstrated continued learning and potential overlearning.

I also assessed learning within sessions to determine learning trends between trials (or pair occurrences) within each session. Figure 7 shows the average accuracy per trial for all four sessions in black and the PPE model fit to the observed performance data

in red. In this case, model parameter values were estimated using data from all four sessions. The first point in session one represents the average accuracy of responses (for the group of participants) when the stimulus word is presented for the first time and the participant is asked to recall its pair. The second point within each curve represents the second time that each word pair appeared on the screen. Subsequent points represent subsequent trials.

The curves illustrate the average learning curve for the Japanese-English task. The curve demonstrates distinct learning rates within each session, the decay seen between sessions (e.g., lower intercept for trial 1 within session 2 than trial 10 at session 1), and faster relearning in later sessions, all of which are accounted for in PPE model parameters and fit. Overall, the model fits the data extremely well (r = .928, MSE = .006). Four model parameters were optimized for each participant based on model fit. These parameters will be used when comparing participant performance and model generalizability below.

Digit-Droodle task. Figure 8 illustrates average accuracy for each session (see also Table 1). Correcting for violations in sphericity with the Greenhouse-Geisser adjustment to degrees of freedom, results demonstrated learning over time, $F(1.36,77.34) = 131.74$, $p < .01$. Accuracy increased in session 2, $F(1,57) = 130.39$, $p < .01$, and session 3, $F(1,57) = 68.16$, $p < .01$), from the previous session but plateaued between sessions 3 and 4 (performance ceiling), which demonstrated continued learning and potential overlearning.

As above**,** I assessed learning within sessions to determine learning rates between trials within each session. In this task, there were ten stimulus pairs shown ten times per session. Figure 9 shows the average accuracy per trial for all four sessions in black and the PPE model fit to the observed performance data in red. As in the curves for the Japanese-English task, the curves within each session demonstrate the average learning curve for this task. The model fits the data extremely well (r = .907, MSE = .005).

Cardinal Direction task. Figure 10 illustrates average accuracy for each session (see also Table 1). Correcting for violations in sphericity with the Greenhouse-Geisser adjustment to degrees of freedom, results demonstrated learning over time, $F(1.21,69.06) = 131.74$, $p < .01$. Accuracy increased in session 2, $F(1,57) = 130.39$, $p < .01$, and session 3, $F(1,57) = 68.16$, $p < .01$, from the previous session but plateaued between sessions 3 and 4 (performance ceiling), which demonstrated continued learning and potential overlearning.

As above, I assessed learning within sessions to determine learning rates between each of the 64 configurations, or trials within each session. I assessed only the first 64 configuration occurrences to determine a measure of accuracy for the first time each configuration was answered. Figure 11 shows the average accuracy per configuration for all four sessions in black and the PPE model fit to the observed performance data in red. Notice that the observable accuracy scores (in black) appear substantially different than the trial scores for the Japanese-English or Digit-Droodle task. This appearance of difference is likely superficial due to the difference in "trial" designation, but does not

necessarily represent difference in performance, which will be assessed in the task comparisons below. In the two memory-recall tasks there are 10 trials, each consisting of 10 word pairs. In the spatial learning task one trial is one configuration, meaning that the observable data includes 64 trials instead of 112. Regardless, the first point in session one represents the average accuracy of responses (for the group of participants) when the first configuration is shown, and the second point represents the second configuration.

Although the black curves are less distinct or smooth in this task, the PPE is fit based on optimized parameters. The model fits the data extremely well considering the distribution shown (r = .231, MSE = .067). Next, I will discuss performance and PPO optimized parameter and retention predictions as compared across tasks.

Comparing Across Tasks

Raw Performance data. Hypothesis 1, that raw individual performance measures will be similar between the two memory-recall tasks, was tested by computing the correlations between all 12 performance measures (three tasks with four sessions; see Table 1). For Session 1, the two memory-recall tasks were significantly correlated (r = .42, $p < .01$), but the Cardinal Direction task was not significantly correlated with either the Japanese-English (r = .06, p = .66) or the Digit-Droodle (r = .14, p = .30) task. For Session 2, the two memory-recall tasks were significantly correlated (r = .51, $p < .01$) but the Cardinal Direction task was not significantly correlated with either the Japanese-English (r = .09, p = .50) or the Digit-Droodle (r = -.06, p = .65) task. For Session 3 the two memory-recall tasks were significantly correlated (r = .40, $p < .01$) but the Cardinal

Direction task was not correlated with either the Japanese-English (r = .149, p = .27) or the Digit-Droodle (r = .06, p = .68) task. For Session 4 the two memory-recall tasks were significantly correlated (r = .48, p < .01) but the Cardinal Direction task was not correlated with either the Japanese-English (r = .21, p = .11) or the Digit-Droodle (r = -.06, p = .67) task. This pattern provides clear support for Hypothesis 1.

I also analyzed performance scores using 3 x 4 repeated measures ANOVA to evaluate increases in performance across time and to determine if these increases were task specific. Correcting for violations in sphericity with the Greenhouse-Geisser adjustment to degrees of freedom, there were main effects for Task, F(1.88, 107.12) = 6.15, p < .01, and for Session number, F(1.27, 72.49) = 223.65, p < .01. These effects are aligned with expectations that learning increased by Session (over time) and was task dependent. Although Hypothesis 1 only focused on the level of performance, the effect of Task x Session interaction is most valuable to determining if changes in performance were similar across tasks. This Task x Session interaction was significant, F(2.58, 146.97) = 18.98, p < .01. Within-subject contrasts indicated that both the Digit-Droodle task, F(1,57) = 8.13, p < .01, and the Cardinal Direction task, F(1,57) = 24.43, p < .01, showed significantly different increases from Session 1 to Session 2 than the Japanese-English task. Increases from Session 2 to Session 3 were not different between the Japanese-English task and the Digit-Droodle task, F(1,57) = .65, p = .42, but the Cardinal Direction task was significantly different than the Japanese-English task, F(1,57) = 17.13,

$p < .01$. There were no significant differences in changes from Session 3 to Session 4 between tasks. Figure 12 shows this trend over time.

Consistency of parameter estimates. Two methods were used to test the consistency of PPO optimized parameters. For the first method, Hypothesis 2a examined whether the PPO optimized parameters were correlated across tasks. This was tested by correlating the PPO optimized parameters from one task to the PPO optimized parameters to a different task for each individual. Results are presented in Tables 2 through 5.

Examining Tables 2 through 5, Hypothesis 2a was generally not supported. The parameters for logistic function intercept were significantly correlated for the similar tasks ($r = .30$, $p < .05$) but the logistic function intercepts for the Cardinal Direction task were not correlated with the logistic function intercepts for either the Japanese-English ($r = -.06$, $p > .05$) or the Digit-Droodle ($r = .14$, $p > .05$). All others correlations were non-significant. It should be noted that Hypothesis 2a is a potentially insensitive test of the PPO predictions. That is, parameters might show little consistency across individuals but still result in similar predictions as represented in Hypothesis 2b and Hypothesis 3.

PPO fit with generalized parameters. The second method of testing the consistency of PPO optimized parameters was reflected in Hypothesis 2b, that applying the optimized parameters from one task to another would result in similar fit indices. To test Hypothesis 2b I analyzed PPE fit when optimized parameters from one task are applied to the other two tasks. This was tested by running the PPO with the data for one

task but specifying the optimized parameters from another task. Evaluations of model fit are presented in Tables 6 and 7.

Figure 13 shows PPE fit when the parameters optimized for Japanese-English are applied to the Digit-Droodle and Cardinal Direction tasks. Figures 14 and 15 illustrate the same relationship when PPE optimized parameters from the Digit-Droodle and Cardinal Direction tasks are applied to the other tasks (respectively). There is an apparent overlap in PPE fit when the optimized parameters for the Japanese-English or the Digit-Droodle task are applied to each other. There are also apparent differences in PPE curves when those parameters are applied to the Cardinal Direction task or the parameters optimized for the Cardinal Direction task are applied to either of the two memory-recall tasks.

As reported above and displayed in Tables 6 and 7, the optimized parameters fit best to the task that they were estimated from. For example, when looking across the first row of both Tables 6 and 7, fit for the Japanese-English task decreased from the base model more when using the Cardinal Direction task parameters, r =.778, MSE = .087, than the Digit-Droodle parameters, r = .837, MSE = .029. Row two of both tables similarly shows that fit for the Digit-Droodle performance was better when estimated with the Japanese-English parameters than with the Cardinal Direction parameters. Although the pattern of these results appear to support Hypothesis 2b, they do not provide a statistical test of the hypothesis; this was evaluated by testing the difference between the r values with a Steiger (1980) Z test.

Results indicated that fit for the Japanese-English task was not significantly different when estimated with the Digit-Droodle parameters than with the Japanese-English parameters, $Z = 1.86$, $p = .06$, but was significantly different when estimated with the Cardinal Direction parameters, $Z = 3.43$, $p < .01$. Fit for the Digit-Droodle task was not significantly worse when estimated with the Japanese-English parameters than with the Digit-Droodle parameters, $Z = 1.44$, $p = .15$, but was significantly different when estimated with the Cardinal Direction parameters, $Z = 2.95$, $p < .01$. These results support Hypothesis 2b and show that using parameters from the different memory recall tasks did not result in a significant change in r model values but that using parameters from the Cardinal Direction task resulted in significantly different r values for the memory recall tasks.

PPO predicted performance

Hypothesis 3, that PPO predicted retention performance will significantly correlate for the two memory-recall tasks, but not for the spatial learning task, was tested using correlations of the predicted level of retention per individual and across tasks at fixed times in the future. Figures 16, 17 and 18 show model calibration on the left, and PPO predicted performance of each participant ($N = 58$) for 365 days following training for the Japanese-English, Digit-Droodle, and Cardinal Direction task, respectively, on the right.

Because retention performance at the one-week retention interval was high (low forgetting rate associated with any of the three tasks), I looked at predicted performance

for 14, 30, 60, and 120 days post-training. Predicted performance levels and correlations across tasks are presented in Table 8. As hypothesized, predicted performance for the Japanese-English task was correlated with predicted performance for the Digit-Droodle task at 14 ($r = .50$, $p < .05$), 30 ($r = .48$, $p < .05$), 60 ($r = .44$, $p < .05$), and 120 ($r = .40$, $p < .05$) days post-training; however, predicted performance for the Japanese-English task was also correlated with predicted performance for the Cardinal Direction task at 14 ($r = .30$, $p < .05$), 30 ($r = .32$, $p < .05$), 60 ($r = .35$, $p < .05$), and 120 ($r = .37$, $p < .05$) days post-training. Predicted performance on the Digit-Droodle task was not correlated with predicted performance on the Cardinal Direction task at any post-training time point (*r's* = .03, .04, .05, and .06 for 14, 30, 60, and 120 days post-training). These results partially support Hypothesis 3.

Similar to the analysis conducted for Hypothesis 1, predicted retention data was analyzed with a 3 x 4 repeated measures ANOVA. Correcting for violations in sphericity with the Greenhouse-Geisser adjustment to degrees of freedom, there were main effects for Task, $F(1.524, 86.88) = 6.30$, $p < .01$, and for days post-training, $F(1.007, 57.38) = 151.35$, $p < .01$. These effects are aligned with expectations that retention is expected to decrease over time and that performance was task dependent. Again, the Task x Days Post-Training interaction is the most valuable for determining if decreases in retention were similar across tasks. This interaction was significant, $F(1.725, 98.311) = 17.82$, $p < .01$. Within-subject contrasts indicated that both the Japanese-English task and Cardinal Direction task decreased at similar rates from 14 days post-training to 30 days post-

training, $F(1, 57) = .28$, $p = .59$, from 30 days post-training to 60 days post-training, $F(1, 57) = .38$, $p = .54$, and from 60 days post-training to 120 days post-training, $F(1, 57) = .48$, $p = .49$. However, retention on the Digit-Droodle task decreased significantly faster than on the Japanese-English task at each point: $F(1, 57) = 17.61$, $p < .01$; $F(1, 57) = 18.71$, $p < .01$; $F(1, 57) = 22.03$, $p < .01$. Figure 19 shows this trend over time.

Discussion

The purpose of my study was to understand the relationship of task type designation, task performance, PPE optimized parameters and PPO predictions to better understand potential generalizability of PPO across tasks. The underlying cognitive model of PPO tracks and predicts future performance for skills and knowledge retention based on historical performance data for a specific task and training structure. In my study, I compared performance outcomes, optimized PPE parameters and PPO prescribed retention intervals across three tasks -- two of the same type (memory-recall) and one of a different type (spatial) -- to evaluate model generalizability. Generally, the results demonstrated significant similarities in raw individual performance, or accuracy scores, for tasks of the same type (the two memory-recall tasks), but not for the different type task (spatial task). This trend was also consistent between all four training sessions, demonstrating that learning increased over time and was task type dependent. Although model parameters did not directly correlate across similar tasks, the fit of the model from one memory-recall task to the other was far superior to the fit seen when parameters from one memory-recall task were applied to the spatial task and vice versa. Finally, predicted

performance measures across all three tasks demonstrated mixed assessments of similarity based on task type. Similarities of these measures within the same type task suggest that type of task may be a viable indicator of transfer and generalizability of PPO outcomes. Specifically, these results raise issues relating to the effect of task type on learning acquisition and retention and the potential of modeling this type of learning phenomenon. It is apparent that task type designation accounts for some, but not all, of the variance in learning outcomes. This means that although this may be a valuable parameter for the PPO, and qualifier of task content, other variables of the task affect learning outcomes.

Hypothesis Support

The results and analyses support my predictions for Hypothesis 1, Hypothesis 2b, and partially support Hypothesis 3. In general, performance over time, optimized parameter fit across learning curves, and predicted future performance was similar between the two same-type tasks, Japanese-English and Digit-Droodle, and different for the different-type task, Cardinal Direction. Raw performance data and changes in performance over time were significantly correlated within tasks of the same-type and different for the different-type task. This finding suggests an underlying learning tendency based on task-type designation.

Hypothesis 2a was likely not supported due to the nature of parameter variance. For each task, PPE optimizes fit with four free varying parameters based on observable and quantifiable performance history. Maximum likelihood estimation is used to find the

very best fitting parameter values per individual and task. Low correlations in this analysis may be a consequence of noise in parameter estimation, rather than lack of parameter generalizability. It is possible for the variables to cancel each other out or vary in different ways. For example, change in one parameter value (decay intercept) can be offset by a change in another parameter value (decay slope). Hypothesis 2a tested the extent to which parameters from one model were consistent with another model, but Hypothesis 2b tested model fit when the optimized parameters of one test were applied to actual performance data of the other. This may actually be a more reasonable measure of model generalizability. Results for Hypothesis 2b showed that model fit with optimized parameters of similar-type task were significantly better than model fits parameters from the different-type task. This finding implies potential generalizability of PPE fit as a function of task type designation.

Hypothesis 3 was only partially supported because predicted performance for the Japanese-English task was correlated with predicted performance in both the similar type task and the different type task post-training. However, the Digit-Droodle task predicted performance was not correlated with predicted performance for the different type task at any level of retention analysis. Results showed that PPE, when calibrated using performance data from one memory-recall task, adequately predicted performance on another related task. This finding suggests that the mental processes approximated by model parameter estimates were invariant within an individual and across related tasks. However, this suggestion would also imply that similar model generalization would not

be expected across different tasks that evoked different mental processes. Results for Hypothesis 3 were mixed, thus additional research is needed to determine how variables within the context of the task affect learning.

Theoretical and Practical Implications

These results are consistent with meta-analytic results that show that task type designation accounts for some of the variance in learning and knowledge retention over time (Arthur et al., 2003; Arthur et al., 1998). Adding to this past research, current results suggests the plausibility of generalizing performance predictions across similar tasks. Further research is required to determine the importance of task type and task content in learning outcomes. For example, although predictions between the two memory recall tasks were similar, examination of Figures 7 and 9 shows that performance on the Digit-Droodle task tended to drop more between session than the Japanese-English task but that relearning for the Digit-Droodle task occurred at a faster rate than the Japanese-English task. This is also observed in Figures 16 and 17 as predicted retention in the Digit-Droodle declines at a faster rate than in the Japanese-English task. It is possible that a more fined grained distinction of task type, such as sensical or nonsensical pairs in memory recall tasks, may assist researchers and practitioners understand similarities in learning.

The data from this study reveals several practical applications worthy of future study. That task type designation accounts for a significant amount of the variance in performance and predicted performance on various tasks implies that task type is a

valuable component of training transfer. This could lead to a better understanding of how people learn and how learning transfers across tasks or environments. This type of pattern has implications for training managers and educators, suggesting that type of task should be considered when creating a training program.

Findings from this study imply a potential for PPO generalizability and ability to make training and retention predictions and prescriptions for tasks outside of the specific input criterion. Expanding this knowledge potentially increases the utility of PPO substantially. Substantial research and validation studies are required to support this claim, however, this study provides a proof of concept that type of task, and potentially other task qualifiers, play a role in training transfer, and thus training prescription generalizability.

With respect to the cognitive model, an increased ability to generalize results of PPO predictions can save money and time in organizational training contexts. Rather than requiring three instances of historical performance on a given task, PPO could use performance from three similar tasks and output a general prediction for that type of task. PPO utility as a tool to aid scheduling of training for a wide variety of task types or cohorts increases substantially if optimized schedules and training plans/predictions can be generalized to a wider selection of inputs.

Model generalizability also has potential utility in the realm of selection because it captures learning and forgetting nuances at the individual level. A better understanding of learning and retention rates for an individuals based on type of task may help

employers match employees to a certain position based on fit of the individual's learning patterns and task requirements of the job. This process could help aid training goals by capturing individualized patterns of learning and extending retention based on learning capabilities of an individual. This would allow organizations to use a sample of behavior to select individuals who are likely to perform well in the job setting required by a certain position. To apply PPO in this manner would require additional testing, validation, and understanding of the outcomes. Further analysis of PPO utility and accuracy in this realm is necessary to avoid legal issues associated with selection tool validation.

Limitations

As with all studies, there are several limitations that should be considered in the current study. A limitation to the confidence in which I draw my results is the performance measures for each of the tasks. The two memory recall tasks are the same type of task, but also vary based on content. The Japanese-English task contains sensical word pairs, while the Digit-Droodle task contains nonsensical droodle-number pairs. This difference could confound the findings and explain some of the variance in the results. Additionally, as seen in Figures 10 and 11, ceiling effects were observed for the Cardinal Direction task. That is, once participants mastered the spatial orientation procedure, their overall level of accuracy was quite high, limiting variability in performance. Reaction time is a more sensitive measure of performance for this task and could be used in future comparisons.

While a strength of the current study was that task performance was evaluated on four different days, the requirement for participants to return four times throughout the study also introduces limitations. First, this requirement increased the potential for participant attrition. I conducted a power analysis to ensure that a large enough sample size was achieved regardless of data errors or participant attrition. I accounted for attrition by using a large sample size and collecting data until the correct number of data points were achieved. However, this could have resulted in the potential that participants who completed the study were different from participants who dropped out of the study; however, this is unlikely because most of the attrition occurred because of random computer problems (only 2 of the 25 participant attrition points were due to failure to show and attempt all four sessions). Second, ideally participants would have returned at the same time each day (session), but scheduling conflicts resulted in some variance in interval times (24 hours +/- 6 hours). These potential differences in session time could account for some performance variability, but this is generally considered a minor source of variation.

There were several computer issues throughout this study. Although data affected by these computer issues were removed from analysis, it is possible that some of the remaining data were impacted by computer errors such as slow processing, delays between trails, and frozen screens. Finally, caution should be used in interpreting the Steiger Z tests used to provide an empirical test to Hypothesis 2b. Because the model fit is averaged across participants, the *r* values found in Table 6 are the averages across

participants. Steiger (1980) developed his test to compare two correlations from a single sample whereas my procedure is more analogous to comparing correlations from a meta-analysis. Additionally, because of task differences, sample sizes for these *r* values varied between the two types of tasks. Within each participant there were only 40 data points for the memory recall tasks but 256 data points for the Cardinal Direction task. Because of this MSE values might be a better indicator of model fit.

Future Research

Given these results, more research is required to determine plausible contexts for PPO generalizability. Task type may be a valuable designator, but likely in conjunction with other contextual factors such as task complexity and sensical or nonsensical pairs. Also, future studies should focus on other task types (e.g., declarative, cognitive, or psychomotor tasks). Further analysis and validation of PPO is necessary to determine the utility of task type variation and the confidence with which PPO predictions generalize to other tasks. PPO currently accounts for individual differences in learning by fitting free parameters within the model to each individual. It would be interesting to see if an added parameter, reflecting task type designation, facilitated individualized training predictions and prescriptions or if task qualification could aid a theory of a priori parameter estimates that accounted for individual differences. A measure of task type as a parameter of the PPE could enhance the utility of PPO across contexts. Finally, it would be beneficial to perform a study in which scheduling of training is manipulated to determine if training

schedules can be optimized to improve learning acquisition and retention rates across various types of tasks.

Conclusion

The purpose of my study was to understand the relationship of task type designation and PPO prediction generalizability. This study demonstrates potential generalizability of PPO across memory recall tasks. Task type designation appears to account for some of the variance in individual learning rates and retention over time, but other task factors may also play an important role in this phenomenon. Additional research is required to determine the appropriate weight of task type designation when making individualized predictions of performance.

References

Adner, R., Polos, L., Ryall, M. & Sorenson, O. (2009). The case for formal theory. *The Academy of Management Review, 34*, 201-208.

Anderson, J. R. (1995). *Learning and Memory*. New York: Wiley.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: a unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1120.

Anderson, L. W., & Krathwohl, D. R., (2001). *A Taxonomy of Learning, Teaching and Assessing: A Revision to Bloom's Taxonomy of Educational Objectives.* Allyn & Bacon. Boston, MA. Pearson Education Group.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological science*, *2*(6), 396-408.

Anderson, J. R., & Schunn, C. (2000). Implications of the ACT-R learning theory: No magic bullets. *Advances in instructional psychology, Educational design and cognitive science*, 1-33.

Arthur, W. Jr., Bennet, W. Jr., Stanush, P.L. & McNelly, T.L. (1998). Factors that influence skill decay and retention: a quantitative review and analysis. *Human Performance, 11*, 57-101.

Arthur, W. J., Bennett, W. J., Edens, P.S., & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology, 88*(2), 234-245. doi:10.1037/0021-9010.88.2.234.

Arthur Jr, W., Day, E. A., Bennett Jr, W., & Portrey, A. M. (Eds.). (2013). *Individual and team skill decay: The science and implications for practice*. Routledge.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects

when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*(3), 144-151.

Bahrick, H. P. (1967). Relearning and the measurement of retention. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 89-94.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308.

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology: General, 113,* 1–29.

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

Baldwin, T. T., & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41, 63-105.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612-637.

Bayliss, D. M., & Jarrold, C. (2015). How quickly they forget: The relationship between forgetting and working memory performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 163-177. doi:10.1037/a0037429.

Begg, I., & Green, C. (1988). Repetition and trace interaction: Superadditivity. *Memory and Cognition, 16*, 232–242.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*, 228-247.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A. (2013). Desirable difficulties perspective on learning. In H. Pashler (Ed.), *Encyclopedia of the mind.* Thousand Oaks: Sage Reference.

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

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). Taxonomy of educational objectives : The classification of educational goals. *Handbook I: Cognitive domain.* New York, Toronto: Longmans, Green.

Boden, Margaret (2008). *Mind as Machine: A History of Cognitive Science*. OUP Oxford.

Bregman, A. S. (1967). Distribution of practice and between-trials interference. *Canadian Journal of Psychology*, *21*, 1-14.

Brown, A. L., Kane, M. J., & Long, C. (1989). Analogical transfer in young children: Analogies as tools for communication and exposition. *Applied Cognitive Psychology, 3,* 275-293.

Carlisle, K. E. (1986). *Analyzing jobs and tasks*. Educational Technology.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. *Psychological Bulletin, 132*(3), 354-380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, *12*(5), 769-786.

de Jonge, M., Tabbers, H. K., & Rikers, R. P. (2014). Retention beyond the threshold: Test-enhanced relearning of forgotten information. *Journal of Cognitive Psychology, 26*(1), 58-64.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of learning and motivation*, *53*, 63-147.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, 84, 795-805.

Dorfberger, S., Adi-Japha, E., & Karni, A. (2007). Reduced susceptibility to interference in the consolidation of motor memory before adolescence. PLoS ONE, 2, e240.

Doyon, J., & Benali, H. (2005). Reorganization and plasticity in the adult brain during learning of motor skills. *Current opinion in neurobiology*, *15*(2), 161-167.

Driskell, J. E., Willis, R. P, & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology*, 77, 615-622.

Ebbinghaus, H. (1885). *Über das Gedachtnis: Untersuchungen zur Experimentellen Psychologie.* Leipzig, Germany: Duncker & Humblot.

Ericsson, K. A, Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.

Ericsson, K. A., & Smith, J. (1991). *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press.

Farr, M. J. (1987). *The long term retention of knowledge and skill: A cognitive and instructional perspective.* New York: Springer-Verlag.

Farina, A. J., Jr., & Wheaton, G. R. (1973). Development of a taxonomy of human performance: The task-characteristics approach to performance prediction. JSAS Catalog of Selected Documents in Psychology, 3, 26–27 (Manuscript No. 323).

Glaser, R., Chi, M. T., & Farr, M. J. (Eds.). (1988). *The nature of expertise*. Lawrence Erlbaum Associates.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306-355.

Goldstein, I. L., & Ford, J. K. (2002). Training in organizations: Needs assessment, development, and evaluation (4th ed.). Belmont, CA: Wadsworth.

Gunzelmann, G., Anderson, J. R., & Douglass, S. (2004). Orientation tasks with multiple views of space: Strategies and performance. *Spatial Cognition and Computation*, *4*(3), 207-253.

Guthrie, E. R. (1952). The psychology of learning (rev).

Halpern, D. F., Hansen, C., & Riefer, D. (1990). Analogies as an aid to understanding and memory. *Journal of Educational psychology, 82,* 298-305.

Hull, C. (1943). Principles of behavior.

Hurlock, R. E., & Montague, W. E. (1982). *Skill retention and its implications for navy tasks: an analytical review* (No. NPRDC-SR-82-21). Navy Personnel Research and Development Center: San Diego, CA.

Jaeggi,S. M., Buschkuehl, M., Shah, P., and Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition, 42(3),* 464-480.

Jastrzembski T.S., Gluck K.A., Gunzelmann G. (2006). Knowledge tracing and prediction of future trainee performance. *I/ITSEC Annual Meetings*, Orlando, December 4–7.

Jastrzembski, T. S., Gluck, K. A., Rodgers, S. (2009) Improving Military Readiness: A State-of- the-Art Cognitive Tool to Predict Performance and Optimize Training Effectiveness. *I/ITSEC Annual Meetings.*

Jastrzembski, T. S., Rogers, S. M., Gluck, K. A., Krusmark, M. A. (2013). *U.S.Patent US8568145 B2.* Washington, DC: The United States of America as Represented by the Secretary of the Air Force.

Jonassen, D. H., & Grabowski, B. L. (1993). *Handbook of Individual Difference, Learning, and Instruction*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Jonassen, D., Tessmer, M., & Hannum, W. H. (1999). Task analysis methods for instructional design. Mahwah, NJ: Erlbaum Associates.

Klausmeier, H. J. (1961). *Learning and human abilities: Educational psychology*. Harper & Row.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85-97.

Krinsky, R., & Nelson, T. O. (1985). The feeling of knowing for different types of retrieval failure. *Acta psychologica*, *58*(2), 141-158.

Kurtz, T., Mogle, J., Sliwinski, M., & Hofer, S (2012). Individual Differences in Task-Specific Paired Associates Learning in Older Adults: The Role of Processing Speed and Working Memory. *Experimental Aging Research, 39*(5), 493-514.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605-621.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1998). The Effects of Graduate Training on reasoning: Formal Discipline and Thinking About Everyday-life Events. *American Psychologist, 43,* 431-442.

MacLean, P., & Cahillane, M. (2015). The human factor in learning design, research, policy, and practice. *The International Journal of Information and Learning Technology*, *32*(3), 182-196.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 126.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. Psychometrika, 55, 107–122.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308-313.

Nelson, T. O. (1978). Detecting small amounts of information in memory: Savings for nonrecognized items. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(5), 453.

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

Nishimoto, T., Ueda, T., Miyawaki, K., & Yne, Y. (2010). A normative set of 98 pairs of nonsensical pictures (droodles). *Behavior Research Methods, 42(3),* 685-691.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*,*115*(1), 39.

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

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology*, *66*(2), 206.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological review*, *109*(3), 472.

Rast, P. (2011). Verbal knowledge working memory and processing speed as predictors of verbal learning in older adults. Developmental Psychology, 47, 1490–1498.

Rast, P., & Zimprich, D. (2009). Individual differences and reliability of paired associates learning in younger and older adults. Psychology and Aging, 24, 1001–1006.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? Journal of Experimental Psychology. General, 140, 283–302.

Rawson, K. A., & Dunlosky, J. (2012). Relearning attenuates the benefits and costs of spacing. Journal of Experimental Psychology: General. doi:10.1037/a0030498.

Rawson, K. A., & Dunlosky, J. (2013). Relearning attenuates the benefits and costs of spacing. *Journal of Experimental Psychology: General*, *142*, 1113-1129.

Rawson, K. S., Dunlosky, J., Sciarteli, S. M. (2013) The Power of Successive Relearning: Improving Performance on Course Exams and Long-Term Retention. *Educational Psychology Review, 24*(4), 523-548.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*, 219-235.

Reed, S. K., Ernst, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. *Cognitive Psychology, 6,* 436-450.

Ritter, F. E., Baxter, G., Kim, J. W., Srinivasmurthy, S. (2013). Learning and retention. In J. D. Lee and A. Kirlik (eds.). *The Oxford handbook of cognitive engineering* (pp. 125-142). New York, NY: Oxford.

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

Sanders, A. F. (1998). Elements of human performance. *Reaction processes and attention in humans. Mahwah,* New Jersey: Lawrence Earlbaum Associates.

Savion-Lemieux, T., Bailey, J. A., & Penhune, V. (2009). Developmental contributions to motor sequence learning. Experimental Brain Research, 195, 293–306.

Savion-Lemieux, T., & Penhune, V. (2005). The effects of practice and delay on motor skill learning and retention. Experimental Brain Research, 161, 423–431.

Schmidt, R. A., and Bjork, R. A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts in training. *Psychological Science, 3* 207-217.

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill* (No. 9). Harvard University Press.

Skinner, B. F. (1938). The behavior of organisms: an experimental analysis.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science 10*(2), 176-199.

Storm, B. C., Bjork, E. L., & Bjork, R. A, (2005). Social metacognitive judgments: The role of retrieval-induced forgetting in person memory and impressions. *Journal of Memory and Language*, *52*, 535-550.

Thorndike, E. L. (1923). The influence of first year Latin upon the ability to read English. *School Sociology*. 17: 165-168.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 381-391.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139,* 352-402. Doi:10.1037/a0028446.

Vancouver, J., & Weinhardt, J. (2012). Modeling the Mind and the Milieu: Computational Modeling for Micro-Level Organizational Researchers. *Organizational Research Methods*, *15*(4), 602-623.

Wallace, S. K., Abella, B. S., & Becker, L. B. (2013). Quantifying the effect of cardiopulmonary resuscitation quality on cardiac arrest outcome a systematic review and meta-analysis. *Circulation: Cardiovascular Quality and Outcomes*, *6*(2), 148-156.

Walsh, M., Gluck, K., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (*in preparation*). Evaluating the Theoretical Adequacy and Applied Potential of Computational Models of the Spacing Effect.

Wang, X., Day, E. A., Kowollik, V., Schuelke, M. J., & Hughes, M. G. (2013). Factors influencing knowledge and skill decay after training: A meta-analysis. In W. Arthur, E. A. Day, W. Bennett, & A. M. Potrtrey (*Eds.*), *Individual and Team Skill Decay: The science and implications for practice*. New York: Routledge.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & cognition*, *25*(5), 731-739.

Woodworth, R. S., & Schlosberg, H. (1954). Experimental psychology (rev).

Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology, 8*, 58–81.

Tables

Table 1

Correlations Between Task per Session

Tasks	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. JE S1	62.34	20.12	--											
2. JE S2	87.71	14.98	0.78*	--										
3. JE S3	95.26	7.55	0.63*	0.89*	--									
4. JE S4	95.76	6.83	0.63*	0.88*	0.92*	--								
5. DD S1	69.45	19.62	**0.42***	0.43*	0.31*	0.27*	--							
6. DD S2	88.81	13.48	0.40*	**0.51***	0.40*	0.38*	0.76*	--						
7. DD S3	95.40	8.92	0.40*	0.54*	**0.40***	0.41*	0.66*	0.93*	--					
8. DD S5	95.31	7.14	0.40*	0.55*	0.41*	**0.48***	0.58*	0.85*	0.91*	--				
9. CD S1	81.03	20.76	**0.06**	0.02	0.07	-0.03	**0.14**	0.05	0.00	-0.08	--			
10. CD S2	93.56	9.02	-0.03	**0.09**	0.06	0.09	-0.07	**-0.06**	-0.04	-0.08	0.56*	--		
11. CD S3	95.07	5.30	0.09	0.24	**0.15**	0.21	-0.02	-0.01	**0.05**	0.03	0.42*	0.80*	--	
12. CD S4	94.80	5.85	0.10	0.26*	0.23	**0.21**	0.04	0.05	0.04	**-0.06**	0.42*	0.75*	.73*	--

Note. $n = 58$. JE represents the Japanese-English task, DD represents the Digit-Droodle task, and CD represents the Cardinal-Direction task. S1 represents Session 1, S2 represents Session 2, S3 represents Session 3, and S4 represents Session 1. Correlations between tasks in the same session are in bold to facilitate comparison.

* $p < .05$ (two-tailed).

Table 2
Correlations between PPO Logistic Function Intercept Parameter

	1	2	3
1. Japanese-English	--		
2. Digit-Droodle	.30*	--	
3. Cardinal Direction	-.06	.14	--

Note. $n = 58$. *$p < .05$.

Table 3
Correlations between PPO Decay Intercept Parameter

	1	2	3
1. Japanese-English	--		
2. Digit-Droodle	.16	--	
3. Cardinal Direction	-.011	-.003	--

Note. $n = 58$. $^{*}p < .05$.

Table 4
Correlations between PPO Decay Scalar Parameter

	1	2	3
1. Japanese-English	--		
2. Digit-Droodle	.20	--	
3. Cardinal Direction	-.11	-.24	--

Note. $n = 58$. $^{*}p < .05$.

Table 5
Correlations between PPO Logistic Function Scalar Parameter

	1	2	3
1. Japanese-English	--		
2. Digit-Droodle	.19	--	
3. Cardinal Direction	.16	.22	--

Note. $n = 58$. $^{*}p < .05$.

Table 6

Correlations Between Task Performance and Model Fit Using Optimized Parameters from Other Tasks

	Parameters		
Performance	1	2	3
1. Japanese-English	**.928**	.837	.778
2. Digit-Droodle	.826	**.907**	.749
3. Cardinal Direction	.152	.22	**.234**

Note. $n = 58$. The bold represents when the model fits performance data using its own parameters.

Table 7

Mean Squared Error Between Task Performance and Model Fit Using Optimized Parameters from Other Tasks

	Parameters		
Performance	1	2	3
1. Japanese-English	**.006**	.029	.087
2. Digit-Droodle	.028	**.005**	.083
3. Cardinal Direction	.089	.089	**.067**

Note. $n = 58$. The bold represents when the model fits performance data using its own parameters.

Table 8

Correlations Between Predicted Retention over Time

Tasks	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. JE 14	90.57	12.63	--											
2. JE 30	88.97	13.85	.99*	--										
3. JE 60	87.07	15.23	.99*	.99*	--									
4. JE 120	84.84	16.80	.97*	.99*	.99*	--								
5. DD 14	86.75	13.94	**.50***	.50*	.49*	.47*	--							
6. DD 30	83.63	15.58	.48*	**.48***	.47*	.45*	.99*	--						
7. DD 60	79.91	17.52	.45*	.45*	**.44***	.43*	.96*	.99*	--					
8. DD 120	75.40	19.76	.41*	.41*	.41*	**.40***	.92*	.97*	.99*	--				
9. CD 14	79.87	21.01	**.30***	.32*	.34*	.37*	**.03**	.03	.04	.04	--			
10. CD 30	78.40	22.11	.30*	**.32***	.35*	.37*	.03	**.04**	.04	.05	.99*	--		
11. CD 60	76.69	23.41	.30*	.33*	**.35***	.37*	.04	.04	**.05**	.05	.99*	.99*	--	
12. CD 120	74.68	24.94	.30*	.33*	.35*	**.37***	.04	.04	.05	**.06**	.99*	.99*	.99*	--

Note. $n = 58$. JE represents the Japanese-English task, DD represents the Digit-Droodle task, and CD represents the Cardinal Direction task. 14, 30, 60, and 120 represents predicted retention that many days post-training. Correlations between tasks at the same time period post-training are in bold to facilitate comparison.

* $p < .05$ (two-tailed).

Figures

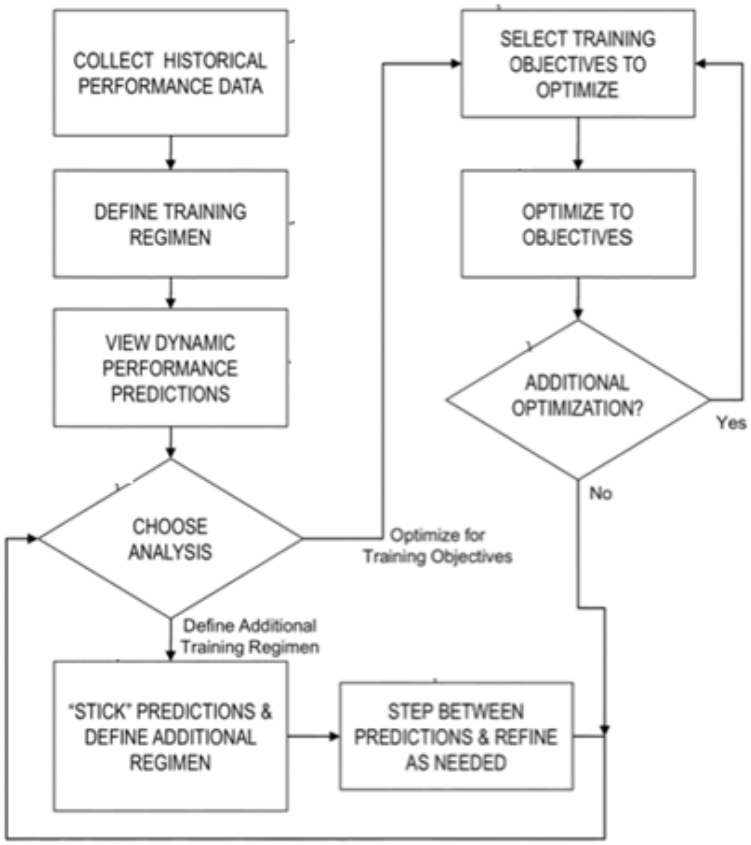

Figure 1. General Process Outline for PPO (Jastrzembski et al., 2013)

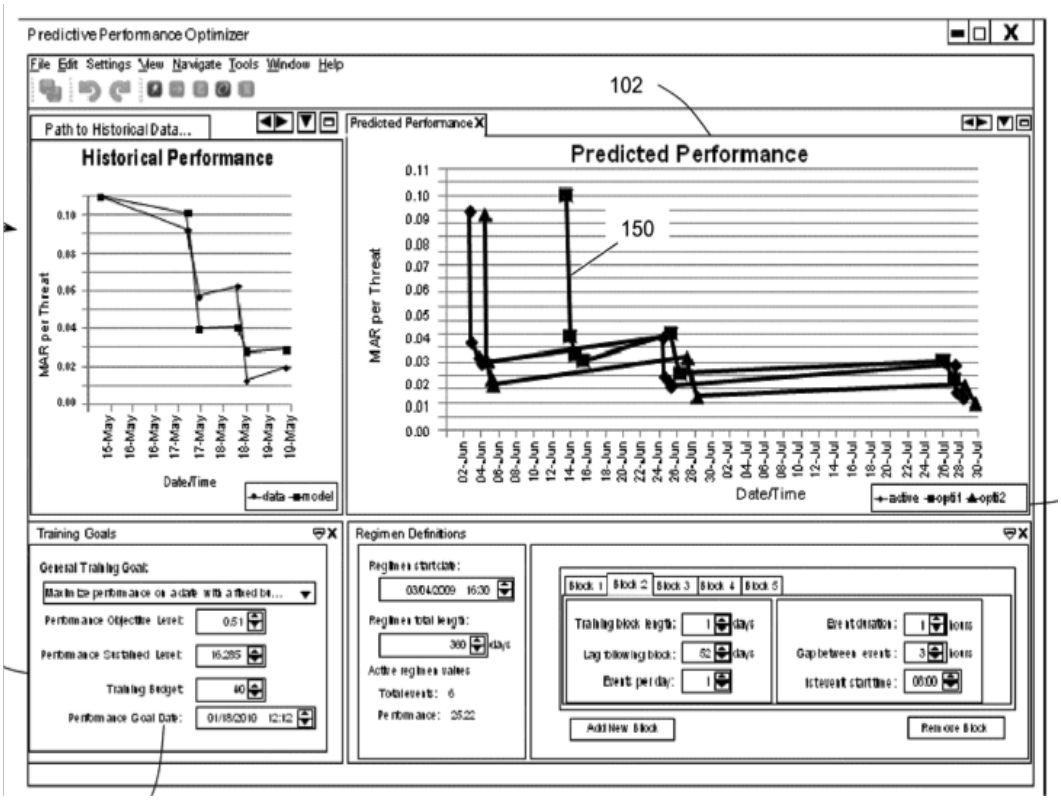

Figure 2. PPO Output Display Example (Jastrzembski et al., 2013)

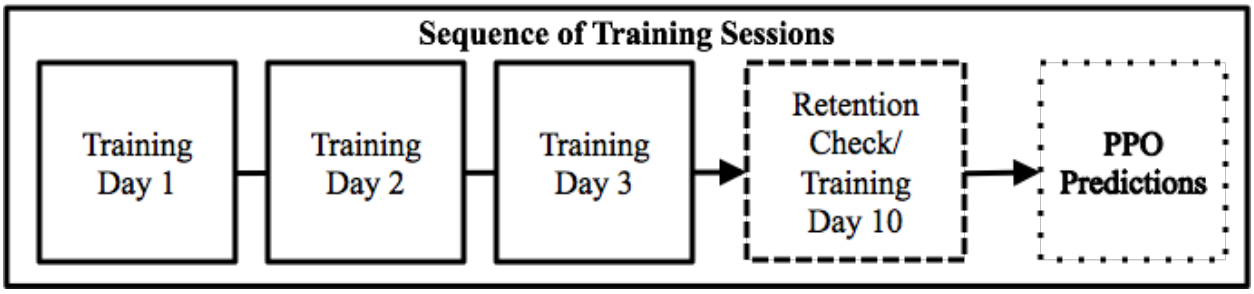

Figure 3. Demonstration of the experimental design structure

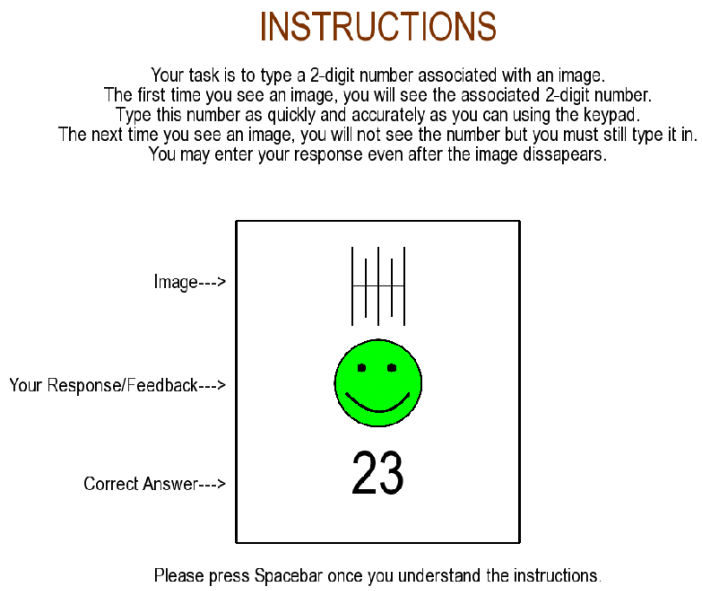

Figure 4. Instruction Screen for Task 2

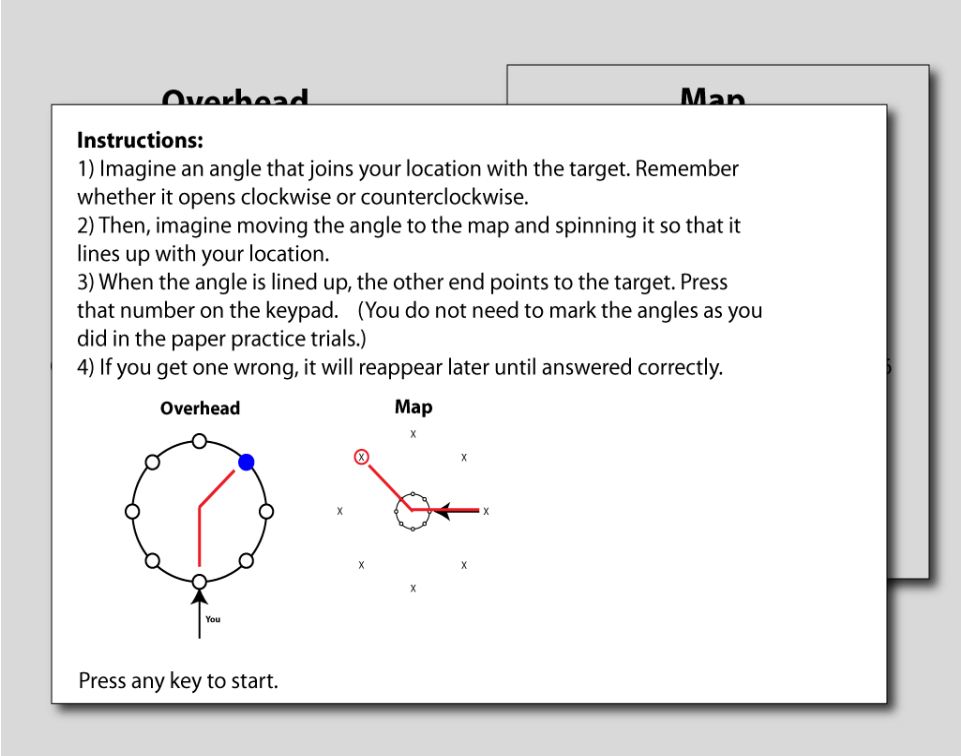

Figure 5. Spatial orientation task directions

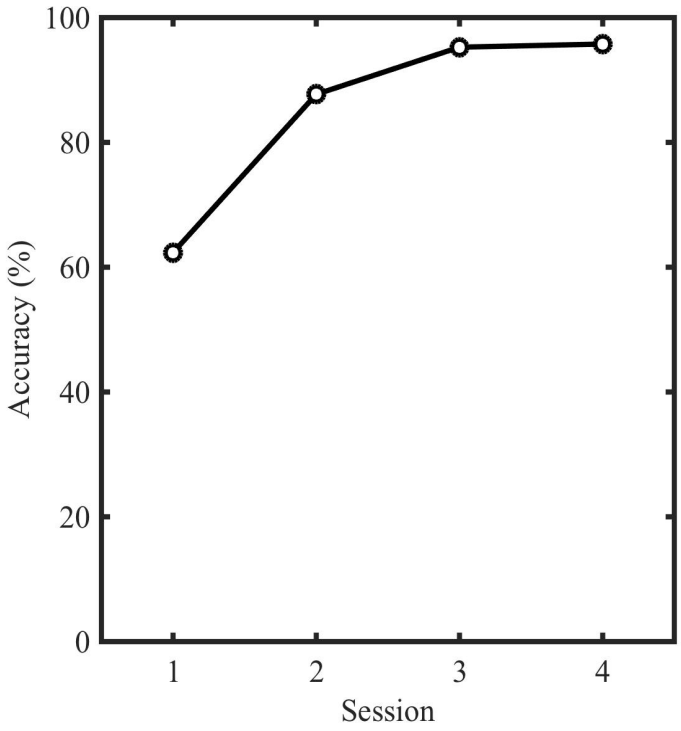

Figure 6. Japanese-English task: across session performance. Average performance (N = 58) in terms of accuracy for the Japanese-English paired associates task across all four sessions.

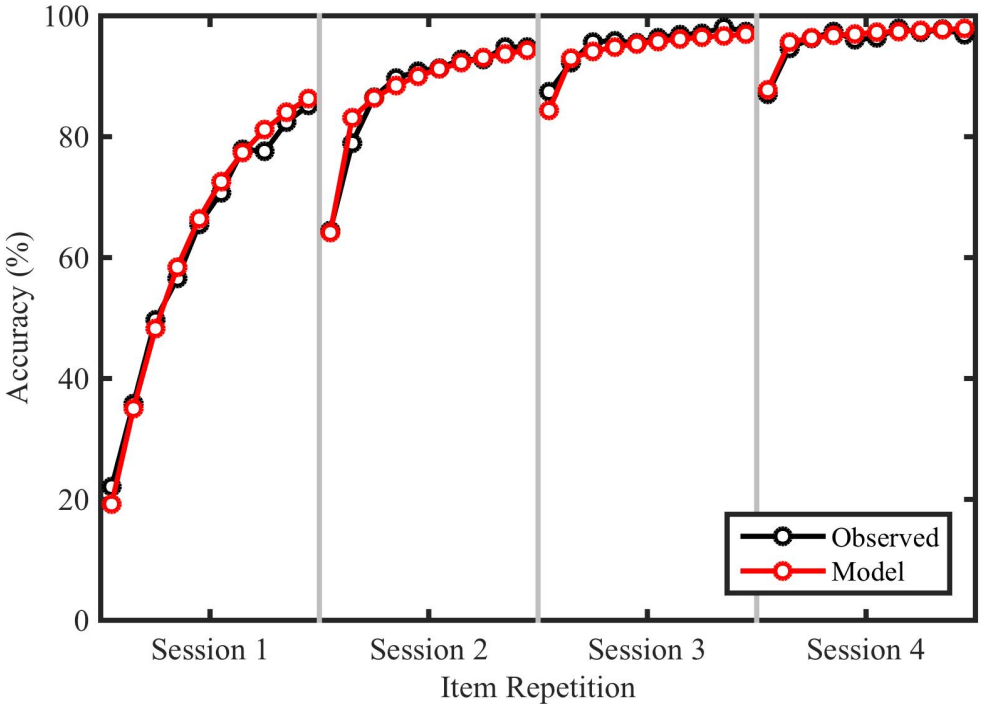

Figure 7. Japanese-English task: trial performance. PPE model fit (red) compared to observed performance accuracy per trial (within session) (black) for the four training sessions.

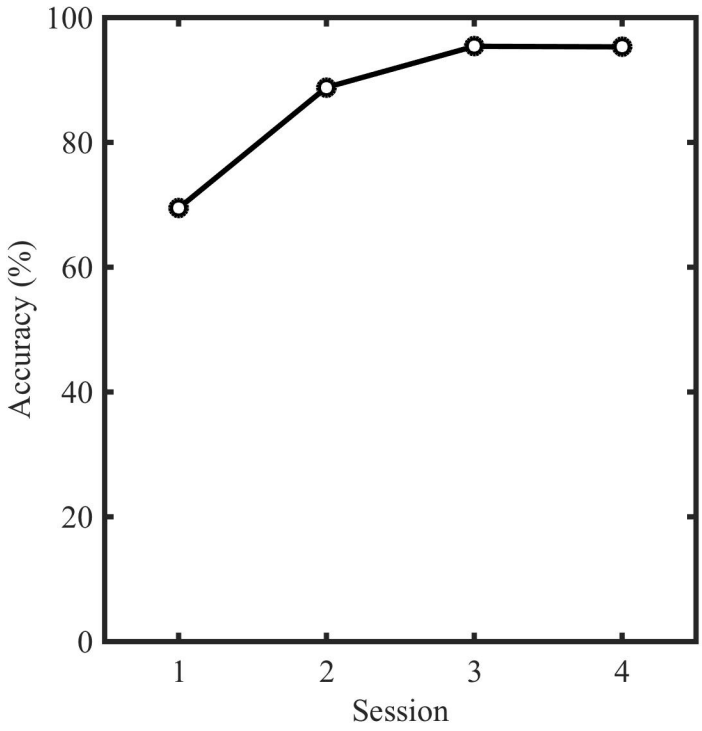

Figure 8. Digit-Droodle task: across session performance. Average performance (N = 58) in terms of accuracy for the Digit-Droodle paired associate task across all four sessions.

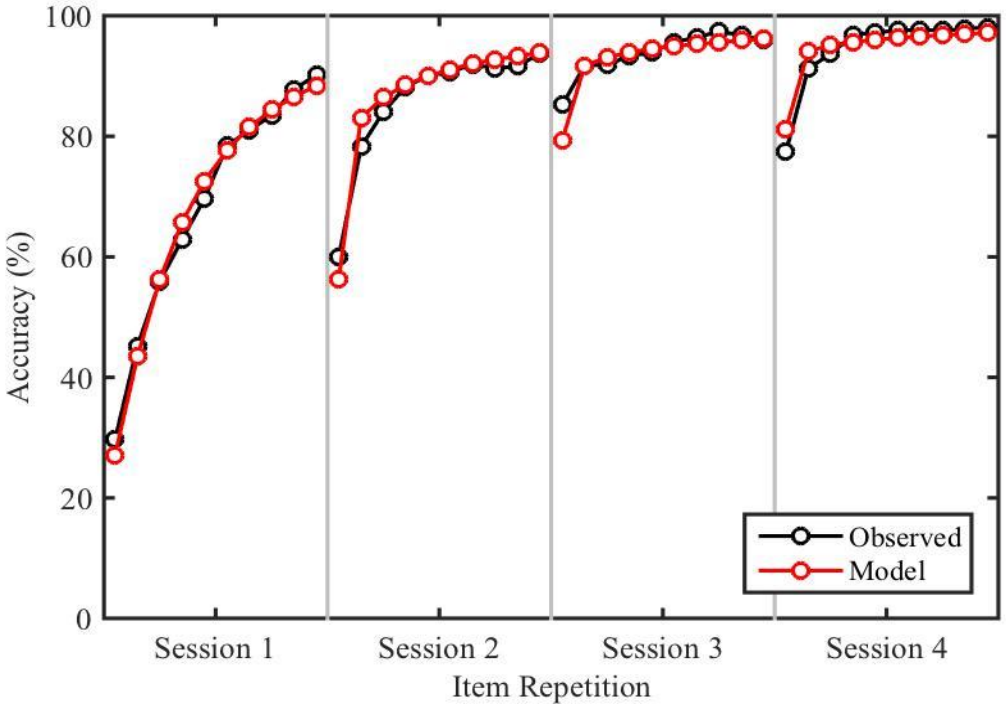

Figure 9. Digit-Droodle task: trial performance. PPE model fit (red) compared to observed performance accuracy per trial (within session) (black) for the four training sessions.

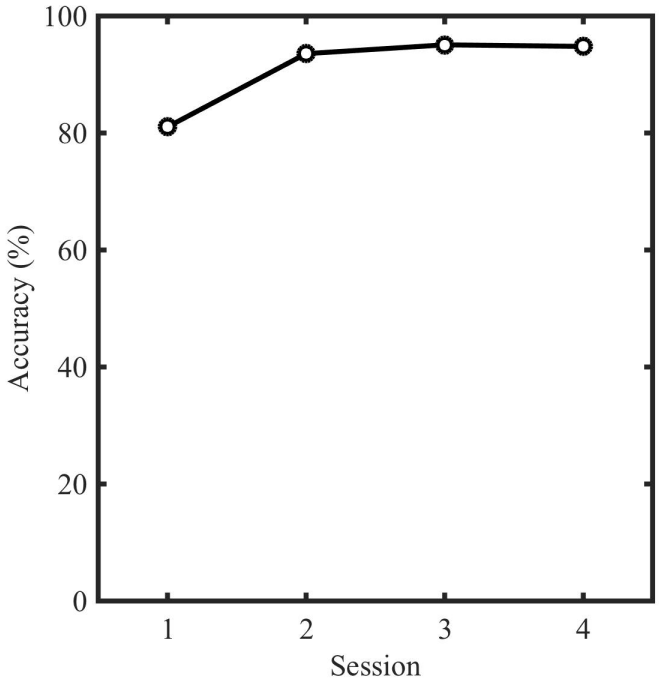

Figure 10. Cardinal Direction task: across session performance.Average performance (*N* = 58) in terms of accuracy for the Cardinal Direction, spatial learning task across all four sessions.

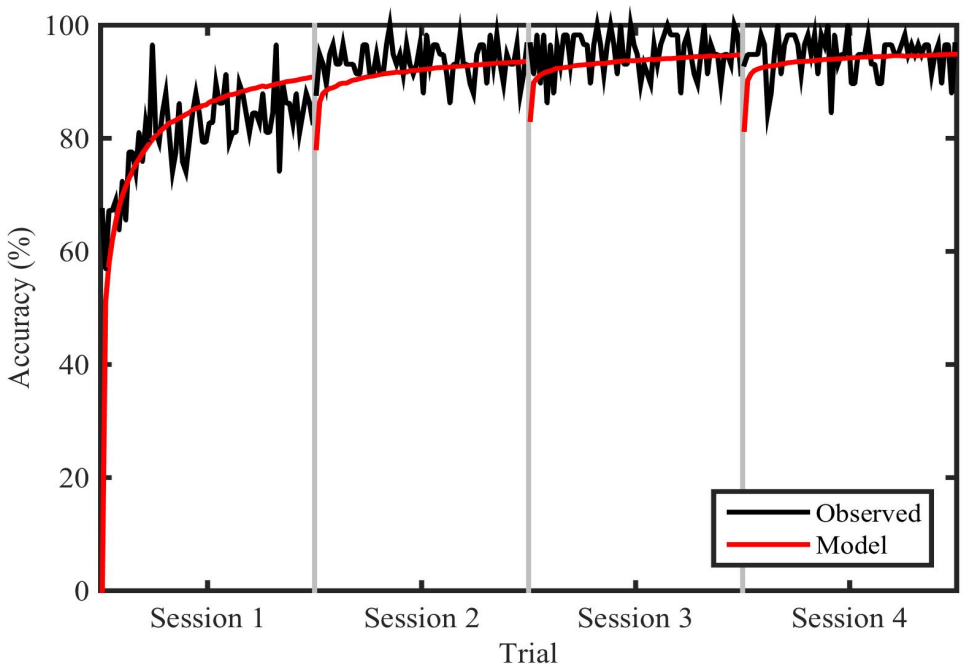

Figure 11. Cardinal Direction task: trial performance. PPE model fit (red) compared to observed performance accuracy per trial (within session) (black) for the four training sessions.

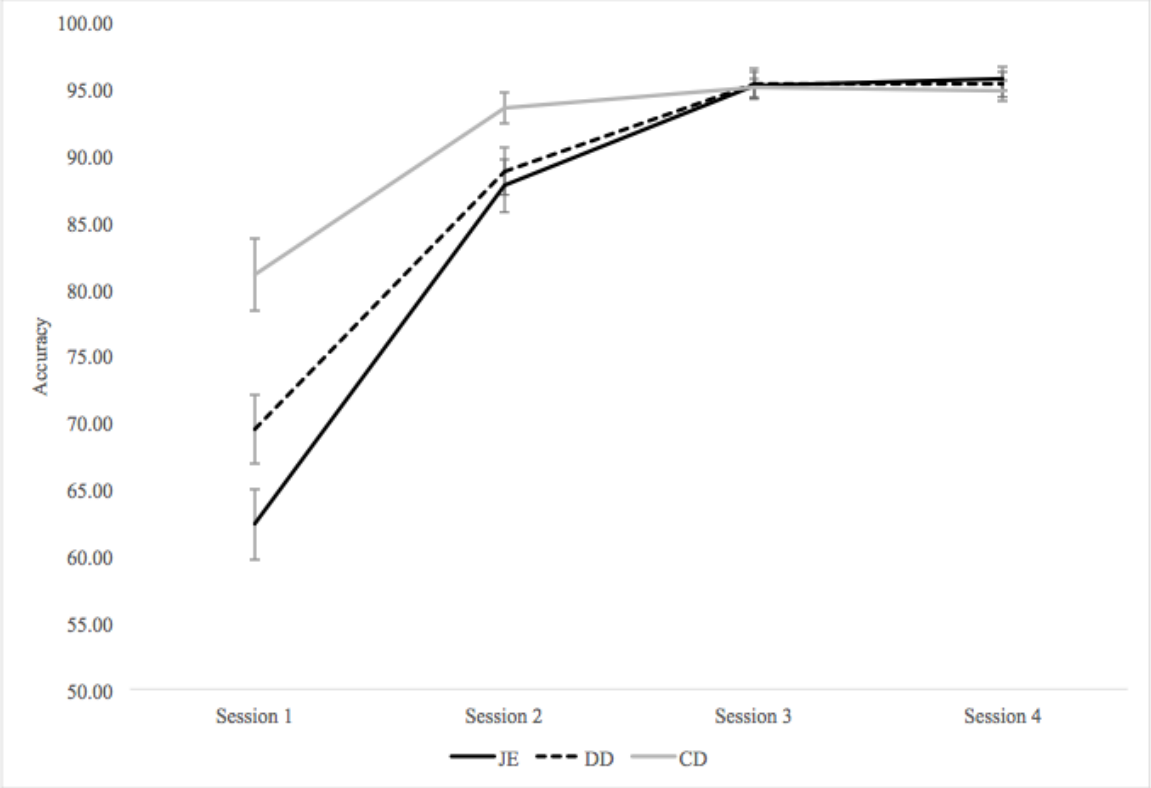

Figure 12. Mean differenes in increased performance per session per task

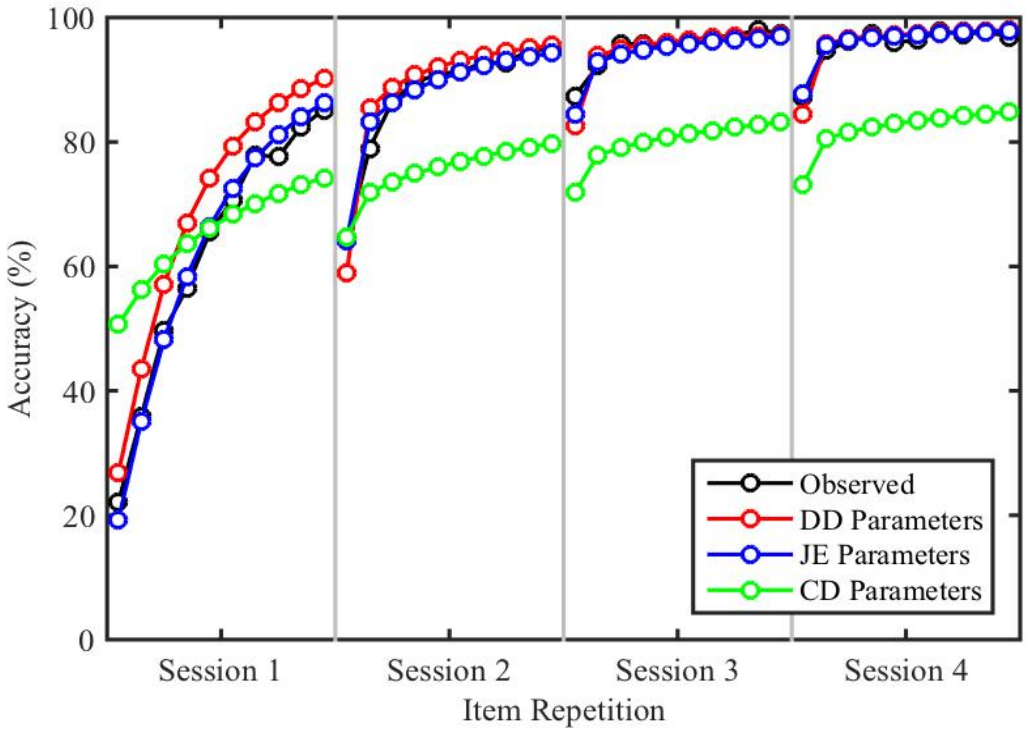

Figure 13. PPE fit when parameters optimized for the Japanese-English task are applied to the Digit-Droodle and Cardinal Direction tasks.

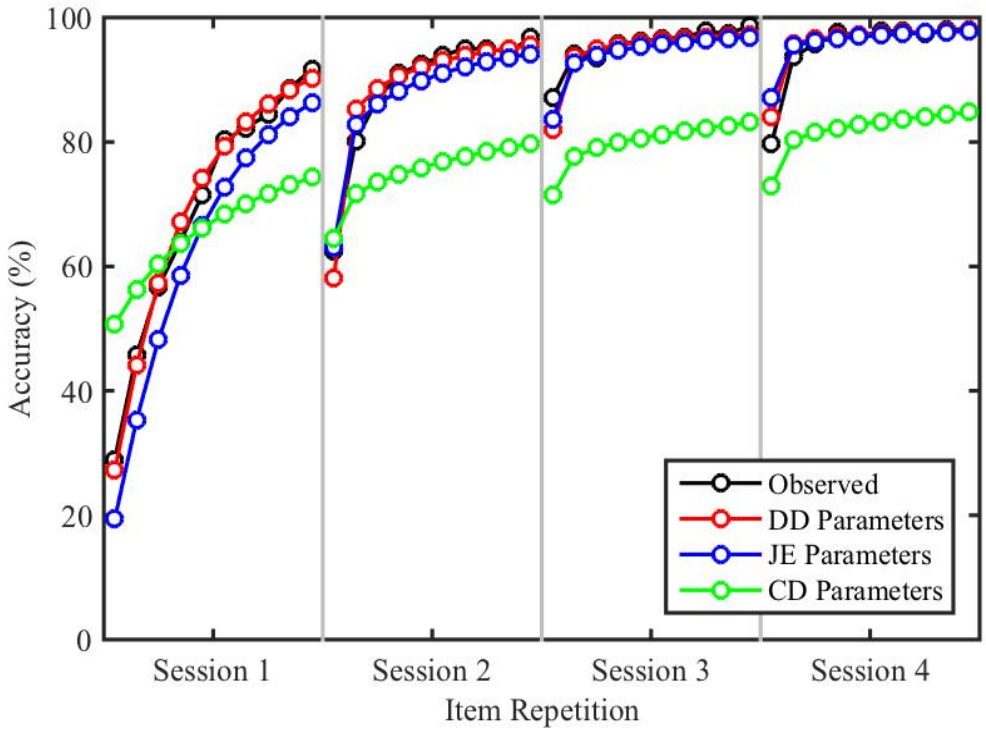

Figure 14. PPE fit when parameters optimized for the Digit-Droodle task are applied to the Japanese-English and Cardinal Direction tasks.

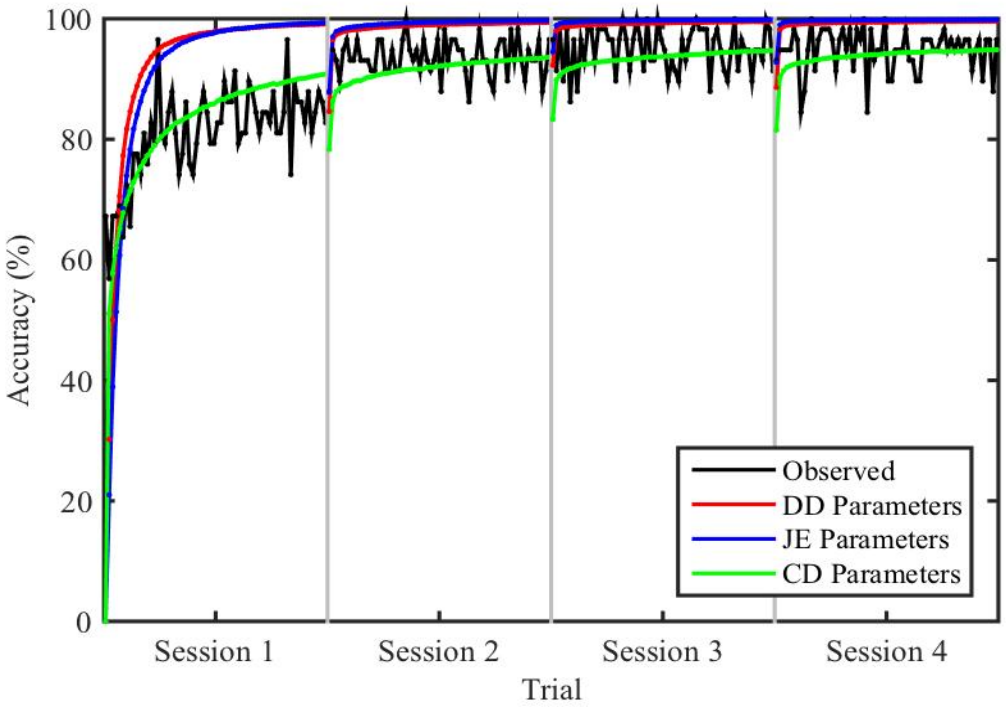

Figure 15. PPE fit when parameters optimized for the Cardinal Direction task are applied to the Japanese-English and Digit-Droodle tasks.

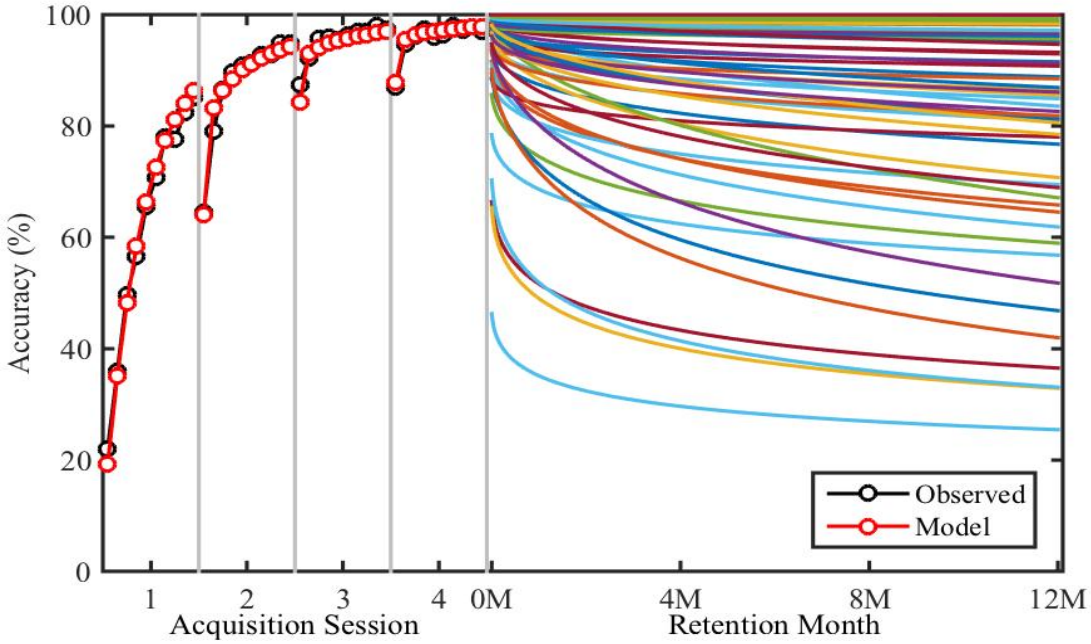

Figure 16. PPO predicted performance for each participant ($N = 58$) from one to 365 days after training completion on the Japanese-English task.

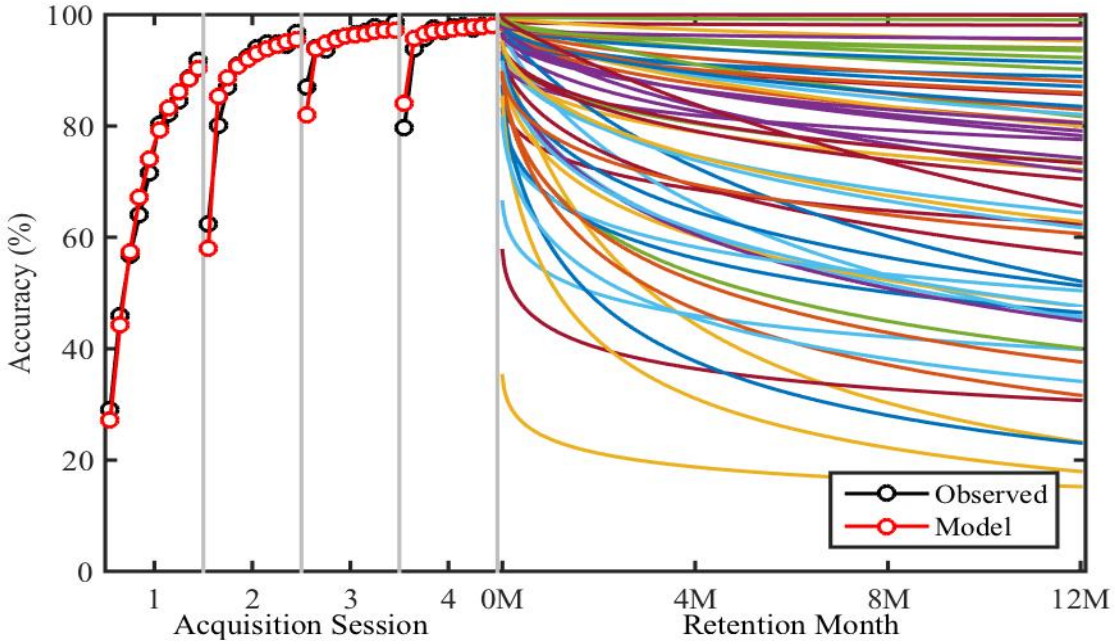

Figure 17. PPO predicted performance for each participant ($N = 58$) from one to 365 days after training completion on the Digit-Droodle task.

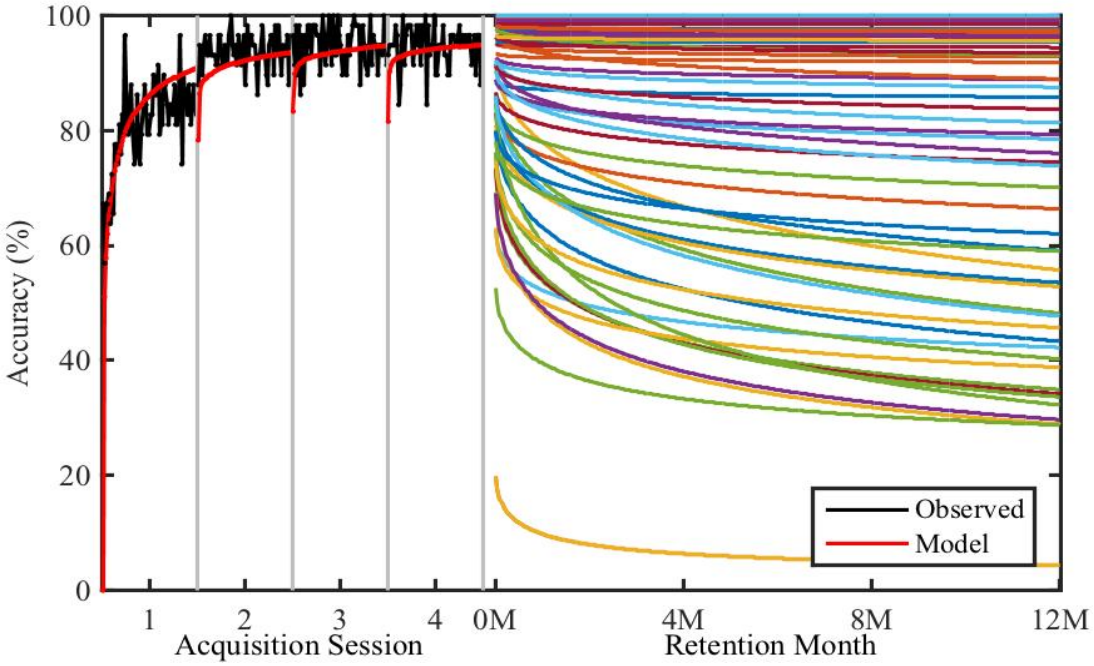

Figure 18. PPO predicted performance for each participant ($N = 58$) from one to 365 days after training completion on the Cardinal Direction task.

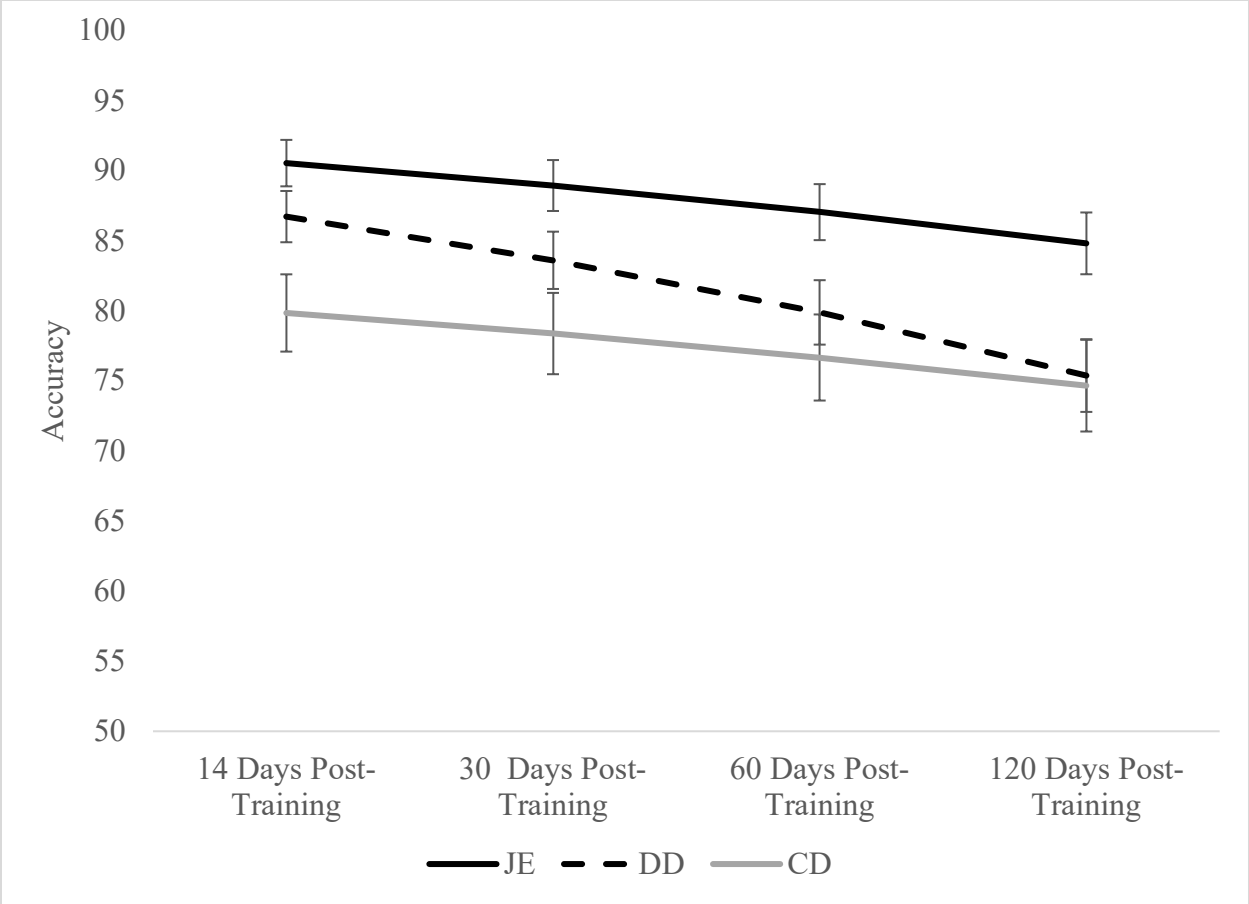

Figure 19. Predicted retention trends for each of the three learning tasks.