# A diffusion model account of the transfer-of-training effect

**Colin Kupitz[1] (ckupitz@uci.edu), Martin Buschkuehl[2] (mbuschkuehl@mindresearch.org),**
**Susanne Jaeggi[1,3] (smjaeggi@uci.edu), John Jonides[4] (jjonides@umich.edu),**
**Priti Shah[4] (priti@umich.edu), and Joachim Vandekerckhove[1,5] (joachim@uci.edu)**

[1]Department of Cognitive Sciences, University of California, Irvine; [2]MIND Research Institute, Irvine, CA;
[3]Department of Education, University of California, Irvine; [4]Department of Cognition and Cognitive
Neuroscience, University of Michigan; [5]Institute for Mathematical Behavioral Sciences, Irvine, CA

## Abstract

We revisit a transfer-of-training study and analyze its data using a cognitive modeling approach. Fitting a diffusion model to participant behavior over sessions allows conclusions as to the underlying causes of behavioral changes—be they changes in cognitive strategies, adaptation to the paradigm, increasing familiarity with the stimuli, or speed of information processing. Our diffusion model analysis revealed that participants simultaneously adapt speed-accuracy trade-off, increase their non-decisional response speed, and increase their speed of information processing. All three of these adaptations transferred to a similar, non-trained outcome task.

**Keywords:** transfer of training; diffusion model; cognitive psychometrics

## Introduction

As a research topic, working memory (WM) training has grown in both interest and controversy in recent years (e.g., Jaeggi, Buschkuehl, Shah, & Jonides, 2014; Morrison & Chein, 2011; Oberauer, Süß, Wilhelm, & Wittman, 2003; Rode, Robson, Purviance, Geary, & Mayr, 2014). The ideal goal of WM training is to improve the underlying cognitive process(es) that is (are) shared across other non-trained tasks. It is assumed that, if these basic underlying processes can be improved, the improvement will not only be observed in the trained task but will generalize to non-trained tasks that rely at least partially on the trained cognitive ability (e.g., WM).

In the current study, we focus on the *change-detection paradigm* (e.g., Luck & Vogel, 1997)—a WM task that has been used for more than a century. In a typical example of this paradigm, the participant is briefly presented with an array and, following a short delay, is asked to judge if a second presented stimulus array is identical to the first or not. Despite the prevalence of the change-detection paradigm in WM literature, the effect of training on task performance—and especially on transfer task performance—has not been investigated thoroughly (for a list of various possible reasons, see Buschkuehl, Jaeggi, Mueller, Shah, & Jonides, under review). Moreover, it has been argued that performance in the change detection paradigm is relatively fixed (Rouder et al., 2008; Zhang & Luck, 2011).

While measurement in the WM literature has traditionally focused on measures of accuracy, speed, and/or

capacity, some researchers have successfully applied cognitive models to WM tasks (e.g., van Vugt & Jha, 2011). We favor such a modeling approach because, while traditional analyses can sometimes provide interesting conclusions, they lack the ability to distinguish between qualitatively different sources of variability in the way that cognitive process models do. For example, if in a training paradigm participants respond more quickly in the last session than the first, this may be because they became more adept at processing the information needed for the task, but they might also have become more efficient at the perceptual or motor component of the response process, or they may have cognitively adapted to the task and act with less caution (either by shifting criterion or a change in speed-accuracy tradeoff). This lack of interpretability of simple summary statistics is an issue in and of itself, and further, averaging artefacts can produce inferential errors and/or biased estimates (Heathcote, Brown, & Mewhort, 2000; see also Clark, 1973). Thus, we believe generating a model to describe the underlying processes of WM tasks is especially important: not only does it provide a novel way of interpreting WM training and transfer, but it will additionally allow us to make stronger and more concrete claims as to the effect and efficacy of WM training tasks *on cognitive processes*, which might allow us to make predictions about near and far transfer depending on which cognitive process(es) improved during training. In this paper, we present a reinterpretation of WM training and transfer data in the context of a cognitive model, as a proof of concept that cognitive modeling is a useful tool in the study of WM tasks, especially in relation to training and transfer.

## Data

We will revisit data by Buschkuehl, Jaeggi, Mueller, Shah, and Jonides (2014). Here we describe only the subset of data that we will use. Other measures are described in Buschkuehl et al. (2014).

### Participants

A total of 40 participants were recruited for the study from two university communities, and were randomly assigned to one of two interventions. Six participants were excluded from the analyses due to either irregularites in

their training schedules (if they failed to complete all 10 training sessions within the 14 day period), or for failing to complete all of the pre- and post-test tasks, leaving a total of 17 participants in each of the two training groups.

## Procedure and tasks

Participants were tested on the two criterion tasks ("easy" and "hard") and then randomly assigned to either the easy or hard training group (test and training tasks are described below). The first session of training was completed in the laboratory in order to give participants the opportunity to ask any questions they might have about the training task or the procedure. The training program was then installed on the personal computers of the participants, and the remainder of the training took place on those computers. In order to ensure compliance, participants were required to send the training data generated after each session via email to the laboratory. Participants were asked to complete ten training sessions (no more than one per day) within 14 days. Following the training period, participants were tested again in the laboratory on the criterion tasks in order to evaluate the impact of the intervention.

*Easy Criterion Task.* Each trial of the easy criterion task began with a fixation cross presented in the center of the screen for 1,000ms. Then, an array of colored squares (possible colors: blue, red, yellow, purple, green, black, white) was presented on a screen with a dark grey background for 250ms, immediately followed by a 200ms blank screen. Next, a set of masks was displayed for 700ms, directly covering the colored square display locations. Each mask consisted of a colored striped square, with each mask being newly generated at each trial from the colors used within the colored squares of that trial. Subsequently a 100ms blank screen was presented, and then one of the squares from the initial array was presented again until a change or no-change judgement was made by the participant. A new trial began 1,000ms after the previous trial ended.

Participants were given task instructions through the computer program and went through ten practice trials. During the practice phase, the stimulus set size (i.e., the number of colored squares) was either two, four, or six, and accuracy feedback was given. After the practice trials, there were 150 test trials: 15 change trials and 15 no-change trials for each of the possible set sizes, 2, 4, 6, 8, and 10. The order of test trials was randomly determined by the computer, and no feedback was given on test trials.

*Hard Criterion Task.* The hard criterion task was similar to the easy criterion task described above with small alterations. Instead of colored squares, random black shapes were used (identical to those in Jaeggi et al., 2003, but black in color and smaller in size). The

stimulus array was presented for 500ms and followed by a 1,000ms blank screen. The entire array was shown again on the test portion of the trial, with the shape to be judged indicated by a black circle. Participants were asked to indicate if the encircled shape was the same as it was in the initial array presentation. The next trial began immediately after the participant made a judgement.

*Easy Training Task.* The easy training task was similar to the easy criterion task described above with three main differences. First, no mask was presented. Second, rather than only displaying the square to be judged, the entire array of squares was redisplayed with the square to be judged encircled. Third, feedback was provided at the end of each trial. The additional smaller alterations made included that the initial array was presented for 250ms followed by a 1,000ms blank screen, which was followed by the test display lasting until the participant responded.

Each training session consisted of 15 blocks of 20 trials. Participants started with a set size of two in their first training session. After each block, performance was evaluated and if accuracy was higher than 85%, the set size
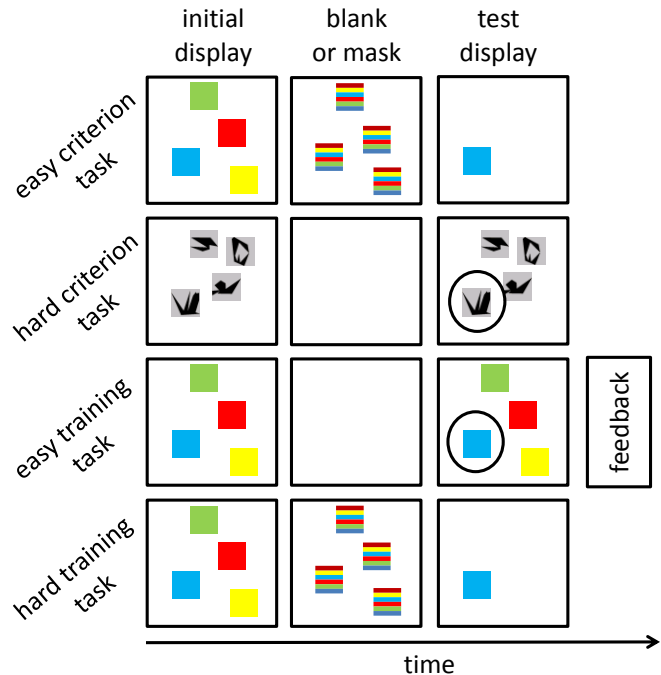


Figure 1: Example trials for each of the four tasks. The easy and the hard criterion tasks differ in the type of stimulus (color squares vs. shapes), the presence of masks, and the number of items remaining in the test display. The easy and hard training tasks differ only in the the presence of masks, the number of items remaining, and the presence of feedback. Note that the hard training task and easy criterion task are the same.
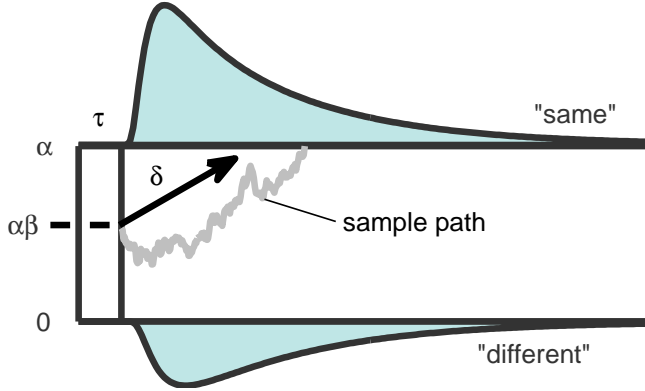
Figure 2: A graphical representation of the Wiener diffusion model. The accumulation process begins at evidence value $\alpha\beta$ and unfolds with an average increase of $\delta$ per second until a boundary at $\alpha$ or 0 is reached. $\tau$ is an additive time constant for nondecisional processes. The shaded area is the model-predicted probability density function over response and response time, $W(\alpha, \beta, \tau, \delta)$.
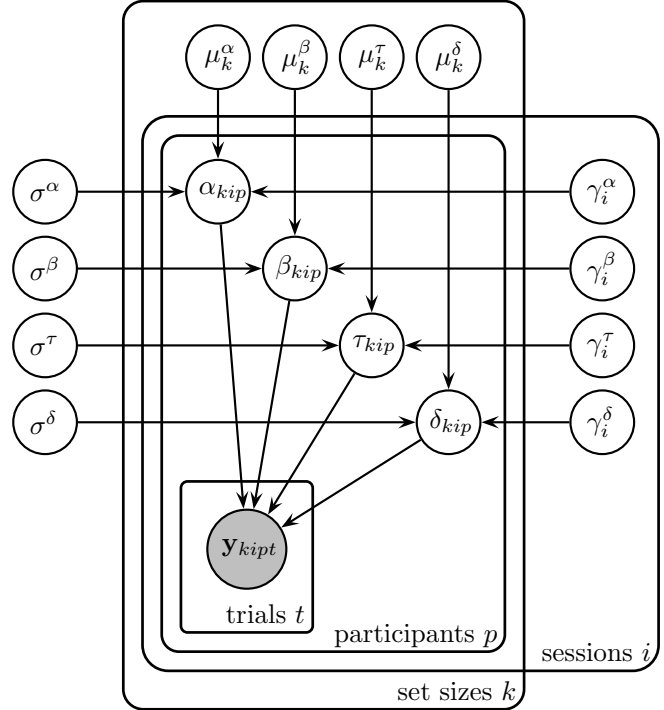


Figure 3: A graphical representation of our exploratory hierarchical diffusion model. Parameters $\mu$ indicate the set-size-specific population mean of each parameter; parameters $\gamma$ indicate the effect of session on each parameter; and parameters $\sigma$ indicate the between-person variability in each parameter. Node $\mathbf{y}_{kipt}$ is the $t^{\text{th}}$ choice response time data point for participant $p$ in session $i$ with set size $k$. For example, the supposed distribution of $\delta_{kip}$ is normal with mean $\mu_k^\delta + \gamma_i^\delta$ and standard deviation $\sigma^\delta$, and the distribution of $\mathbf{y}_{kipt}$ is the Wiener distribution with unit diffusion coefficient. The figure displays only part of the model, which was fit to the training and criterion behavior simultaneously, with the same set-size parameters but freely estimated session offsets.

was increased by one; similarly, if the accuracy dropped below 70%, set size was reduced by one. Otherwise set size remained unchanged. The set size of the first block of subsequent training sessions was determined by subtracting two from the set size of the last block in the previous training session (as 'warm-up time'). The program had a maximum set size of 20, but no participants reached a set size higher than 16.

**Hard Training Task.** The hard training task was identical to the easy criterion task described above. Thus it differed from the easy training task in that there was no feedback provided, there was a mask presented, and only one of the squares was shown in the test display (to preclude any context or configuration effects).

**Data preprocessing.** We did minimal data preprocessing. Beyond the data from excluded participants, we discarded only data from trials in which the response time was clearly too fast (less than 200ms) or too slow to be a one-shot response process (more than 2000ms).

## Diffusion model

Our modeling analysis uses an hierarchical diffusion model for two-choice reaction times introduced by Vandekerckhove, Tuerlinckx, and Lee (2011), which is an extension of a model first described by Stone (1960).

In the diffusion model, it is assumed that participants make task decisions through a process of sequential accumulation of information, executing a response when sufficient information is garnered. Figure 2 illustrates the process. The parameters of interest are $\alpha$, the amount of information required before a decision is made (which captures the speed-accuracy trade-off); $\beta$, the a-priori bias that a participant might have towards one or the other response; $\tau$, the non-decision time including time for encoding the stimulus and executing the motor response; and $\delta$, the "drift rate" or rate of information accumulation within a trial. Importantly, this parameterization gives us a representation of skill at the task (in the form of the drift rate variable, $\delta$), while simultaneously accounting for non-skill based changes in task performance and speed.

In our model, we will decompose the observed parameters into constituent components. For all parameters, we will assume a fixed effect of set size, so that each set size has its own mean value for each parameter (e.g., $\mu_4^\tau$ is the average nondecision time for trials with set size 4). We additionally assume an average fixed offset for each parameter in each session (e.g., $\gamma_5^\beta$ is the average offset in a-priori bias $\beta$ in session 5), relative to the first

training session (so $\gamma_1 = 0$ for all parameters). Finally, we assume a random participant effect, so that each participant gets an additional term to indicate their unique level of each parameter relative to the group mean. This term will be a draw from a normal distribution with mean 0. Taken together, the model is fully described by the set of structural equations

$$\begin{aligned} \theta_{kip} &= \mu_k^\theta + \gamma_i^\theta + \varepsilon_p^\theta \\ \varepsilon_p^\theta &\sim N\left(0, \sigma^\theta\right), \end{aligned}$$

for each diffusion model parameter $\theta$, and the likelihood function $\mathbf{y}_{kipt} \sim W(\alpha, \beta, \tau, \delta)$. The likelihood function is defined as the first passage time distribution of a Wiener process with constant boundaries.

We fit this model simultaneously to the training data and the criterion tasks, allowing for different session offsets for each parameter in each of the criterion sessions.

We implemented the model in an hierarchical Bayesian framework, as in Vandekerckhove et al. (2011). Figure 3 gives a graphical model representation of the model we used. In this graph, variables are represented by nodes. Downstream (i.e., "receiving") nodes are probabilistically dependent on upstream nodes, shaded nodes are observed variables, and unshaded nodes unobserved variables. Plates indicate 'loops' over sets of similar nodes.

We drew eight chains of 1000 samples from the joint posterior distribution of all parameters of the hierarchical diffusion model using a freely available extension of the Bayesian computation program JAGS (Wabersich & Vandekerckhove, 2014). Convergence of the Monte Carlo chains was confirmed with the typical $\hat{R} < 1.1$ criterion.

## Modeling results

### Training

Posterior distributions of the parameters are displayed in Figure 4. The left panels in the figure show the progression of the parameter over sessions. The first session is used as a reference point. The pattern of behavior is clear for each parameter. Over sessions, boundary separation $\alpha$ decreases as participants begin to trade accuracy for speed. The a-priori bias level $\beta$ stays constant and around 0.5, as induced by the experimental paradigm. Nondecision time $\tau$ steadily decreases over sessions. Drift rate $\delta$ shows a slight decrease going from the first to the second session (presumably due to the change in context from the laboratory to the participant's home) but rapidly stabilizes. A slight upward trend is visible.

In a second analysis, the increase of drift rate over sessions two through ten was modeled as a linear function: $\delta_{pik} = \mu_k^\delta + \zeta(i-6) + \varepsilon_p^\delta$, with set-size mean $\mu_k^\delta$, person-specific error term $\varepsilon_p^\delta$, regression weight $\zeta$, and $i$ the session number. In this model, the posterior of regression weight $p(\zeta < 0|\mathbf{y}) \approx 0.007$, indicating a positive trend with mean a posteriori estimate (MAPE) $\hat{\zeta} \approx .011$.

We conducted a third analysis in which we took into account the difference between the "hard training" and "easy training" participant groups. The results were qualitatively similar between the two groups, with the exception that the learning effect on drift rate was smaller in the "easy training" group ($p(\zeta_{\text{EASY}} < 0|\mathbf{y}) \approx 0.071$, MAPE $\hat{\zeta}_{\text{EASY}} \approx .010$) than in the "hard training" group ($p(\zeta_{\text{HARD}} < 0|\mathbf{y}) \approx 0.008$, MAPE $\hat{\zeta}_{\text{HARD}} \approx .015$).

The right panels in Figure 4 show the mean of each parameter per set size. These results are not important to our discussion, save for knowing that the parameters behave in expected ways (most stay relatively constant, except for drift rate, which decreases as expected with increasing task difficulty), and underscoring that set size was taken into account in our analyses.
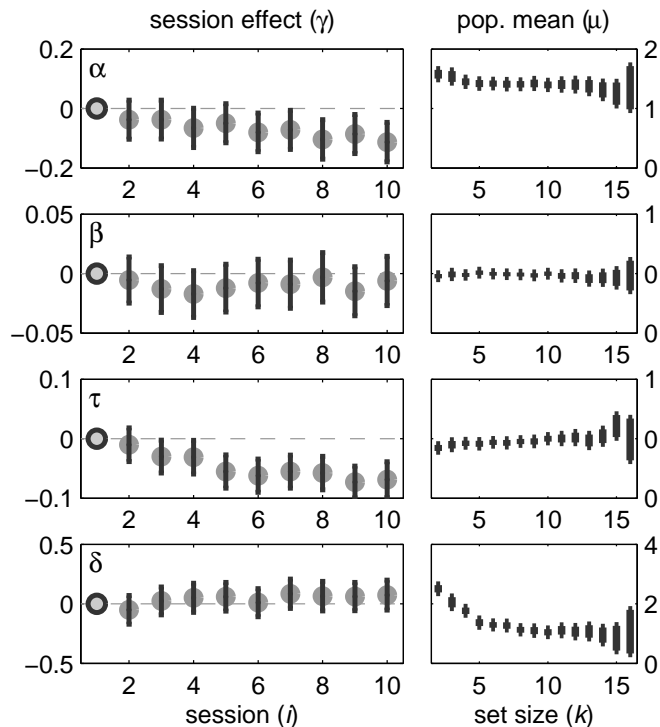


Figure 4: <u>Right panels</u>: Posterior distributions of the population means $\mu_k$ of the four diffusion model parameters as a function of set size $k$. Posterior uncertainty, indicated by the 99% credibility interval, is larger for the highest set sizes because few participants reached that level of difficulty. The panels show little systematic effects, except for a marked decrease in drift rate from set size 2 to 5. This shows that task difficulty increases with set size, but levels off around 5. <u>Left panels</u>: Posterior distributions of the session-specific offset terms $\gamma_i$ as a function of sessions $i$. The leftmost marker is the first session, which is singled out because it was the only training session held in the lab. Ignoring the first session, we observe a decrease in boundary separation $\alpha$ and in nondecision time $\tau$, and a slight increase in drift rate $\delta$.
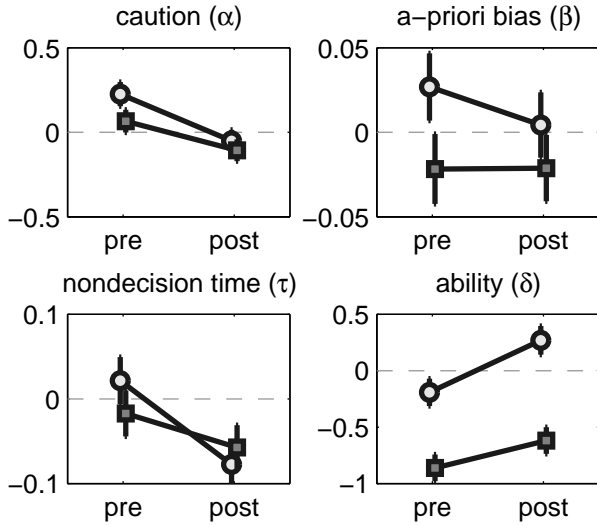
Figure 5: Diffusion model parameter estimates, with 99% credibility intervals, from the transfer tasks. Session is given on the horizontal axes. **Circles** represent the easy criterion task; **Squares** represent the hard criterion task. Top left: $\alpha$s are seen to start above the reference level in the pre-training test and to end below it in the post-training test. Top right: $\beta$s start slightly above the reference level in the pre-training for the easy criterion task and below it for the hard criterion task, with the former decreasing and the latter stable. Bottom left: $\tau$s start level with the reference point but decreases markedly after training. Bottom right: $\delta$ for the easy criterion task starts below the reference level in the pre-training test and ends above it in the post-training test. This is expected because this task is very similar to the training task. Interestingly, for the hard criterion task—which is less similar—$\delta$ increases after training as well, indicating transfer of training.

## Transfer

Figures 5 and 6 show similar results for the criterion tasks. When we compare the pre- and post-test data for the easy (circles) and hard (diamonds) criterion tasks, we find the same changes in boundary separation $\alpha$ and non-decision time $\tau$. Additionally, we also see a stronger increase in drift rate $\delta$. This is particularly interesting given that $\delta$ is most readily interpreted as a higher-level "ability" (e.g., Vandekerckhove, Verheyen, & Tuerlinckx, 2010; Pe, Vandekerckhove, & Kuppens, 2013) which should be less sensitive to specific properties of the task.

## Discussion

Two findings are of note. First, the diffusion model analysis indicates that the improvement seen during the training phase of the experiment is a multicomponential effect: The practice effect consists of simultanous
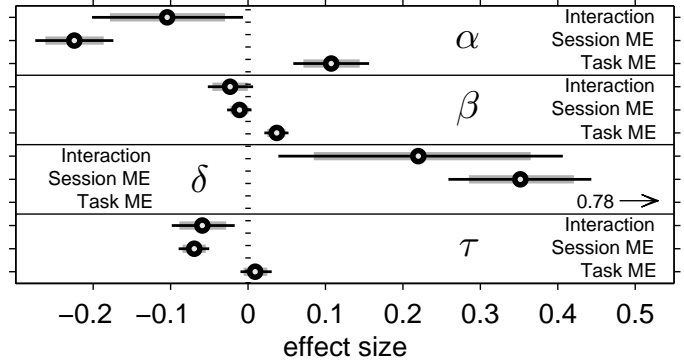


Figure 6: Posterior distributions of effect size estimates corresponding to the training effects in Fig. 5. The Session main effect (ME) is the parameter in the second session minus that in the first. The Task ME is the performance in the easy task minus the hard. All effect sizes are expressed in the parameter's original units. **Circles** represent mean effect size and thick and thin bars the 95% and 99% credibility intervals, respectively. Consistent learning effects are seen in caution $\alpha$, nondecision time $\tau$ and drift rate $\delta$, while bias $\beta$ is the most stable parameter. An interaction effect indicates that the ability parameter $\delta$ increases more for the easy transfer task than for the hard transfer task (which is less similar to the training tasks).

changes in cognitive strategy (the amount of information required to make a decision), motor and encoding time (nondecision time), and—to a lesser degree—task ability (drift rate). Given that drift rate has been associated with fluid intelligence (Ratcliff, Schmiedek, & McKoon, 2008; van Ravenzwaaij, Brown, & Wagenmakers, 2011), this strikes us as the most practically significant finding. This finding is also in line with previous results from cognitive models of practice and learning (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009).

More importantly, the transfer of training effect is seen in the parameters of the diffusion model. On the one hand, we see changes in the boundary separation parameter and the non-decision time. These two parameters are typically interpreted as cognitive strategy (speed/accuracy tradeoff), and speed of stimulus preprocessing and motor response, respectively. In the latter parameter, we expect to see transfer of training to closely related tasks (i.e., tasks that rely on similar stimulus configurations that require similar perceptual encoding), with diminishing effect the more unrelated the tasks become. On the other hand, we also observe an increase in drift rate parameter from the first testing occasion to the last. This parameter is commonly interpreted as a higher level cognitive ability, more distant from superficial task properties. Hence, training in this parameter is expected to transfer more easily to "distant" tasks (i.e.,

tasks that rely on different stimulus configurations), relative to the other parameters of the diffusion model. In future studies, we will explicitly manipulate the distance between tasks to test this hypothesis.

Finally, we should point out that this type of conclusion was made possible through the use of a cognitive psychometric model. Future work will include the application of a more sophisticated cognitive-psychometric model in which individual differences in training effect size will be used to forecast transfer effect size.

# References

Buschkuehl, M., Jaeggi, S., Mueller, S., Shah, P., & Jonides, J. (under review). Training change detection leads to substantial task-specific improvement.

Buschkuehl, M., Jaeggi, S. M., Mueller, S. T., Shah, P., & Jonides, J. (2014). Training on change detection leads to substantial task-specific improvements. In *Poster presented at the 26th annual convention of the association for psychological science, san francisco, ca.*

Clark, H. H. (1973). The language–as–fixed–effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*, 1026–1036.

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

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

Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., et al. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: functional magnetic resonance imaging study. *NeuroImage*, *19*(2), 210–225.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, *18*(1), 46–60.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167 – 193.

Pe, M., Vandekerckhove, J., & Kuppens, P. (2013). A diffusion model account of the relationship between the emotional flanker task and depression and rumination. *Emotion*, *13*, 739–747.

Ratcliff, R., Schmiedek, F., & McKoon, G. (2008). A diffusion model explanation of the worst performance rule for reaction time and IQ. *Intelligence*, *36*, 10–17.

Rode, C., Robson, R., Purviance, A., Geary, D. C., & Mayr, U. (2014). Is working memory training effective? A study in a school setting. *PloS one*, *9*(8), e104796.

Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975–5979.

Stone, M. (1960). Models for choice–reaction time. *Psychometrika*, *25*, 251–260.

van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*(3), 381–393.

van Vugt, M. K., & Jha, A. P. (2011). Investigating the impact of mindfulness meditation training on working memory: A mathematical modeling approach. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(3), 344–353.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two–choice response times. *Psychological Methods*, *16*, 44–62.

Vandekerckhove, J., Verheyen, S., & Tuerlinckx, F. (2010). A crossed random effects diffusion model for speeded semantic categorization data. *Acta Psychologica*, *133*, 269–282.

Wabersich, D., & Vandekerckhove, J. (2014). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavior Research Methods*, *46*, 15–28.

Zhang, W., & Luck, S. J. (2011). The number and quality of representations in working memory. *Psychological Science*, *22*(11), 1434–1441.

# Authors' Note