UNIVERSITY OF CALIFORNIA,
IRVINE

Applications of Hierarchical Bayesian Cognitive Modeling

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Cognitive Sciences

by

Colin Nicholas Kupitz

Dissertation Committee:
Associate Professor Joachim Vandekerckhove, Chair
Professor Michael D. Lee
Professor Jeffrey N. Rouder

2020

DEDICATION

For Alex – I could not have done this without you; and for Thomas, the light of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the Department of Cognitive Sciences, the School of Social Sciences, the Cognition and Individual Differences lab, and grant #1230118 from NSF's Methods, Measurements, and Statistics for financial support.

I also thank my adviser, Joachim Vandekerckhove, for his invaluable guidance and unwavering support throughout my time at UC Irvine. I would also like to thank Michael Lee for acting as my unofficial adviser at times, and my dissertation committee members Michael Lee and Jeff Rouder, for their advice and support.

I thank Julie Golomb for her guidance prior to my time at UCI, and for helping me to continue the research I started with her at OSU during my time at UCI. I would also like to thank Carina Thiemann, Avni Bapat, and Anna Shafer-Skelton, for their contributions to that research.

I would like to thank Susanne Jaeggi and Martin Buschkuehl for their contributions and input regarding the working memory training study that serves a major foundation for the research in my disseration.

I would finally like to thank my parents, Steven and Ann Kupitz, and my wife's parents, Lynn and Shelly Pearce, for helping to support my family throughout this process.

Chapter 1 of this dissertation was coauthored with Martin Buschkuehl, Susanne Jaeggi, John Jonides, Pritih Shah, and Joachim Vandekerckhove, and is published in the *Proceedings of the 37th Annual meeting of the Cognitive Science Society.* Chapters 2 and 3 were coauthored with Joachim Vandekerckhove.

CURRICULUM VITAE

Colin Nicholas Kupitz

EDUCATION

Doctor of Philosophy in Cognitive Sciences	**est. 2020**
University of California, Irvine	*Irvine, CA*
Master of Science in Cognitive Sciences	**est. 2020**
University of California, Irvine	*Irvine, CA*
Bachelor of Science in Neuroscience and Chemistry	**2012**
University of Nevada, Reno	*Reno, Nevada*

RESEARCH EXPERIENCE

Graduate Research Assistant	**2014–2019**
University of California, Irvine	*Irvine, California*
Lab Manager: OSU Vision & Cognitive Neuroscience Lab	**2012–2014**
Ohio State University	*Columbus, Ohio*

TEACHING EXPERIENCE

Teaching Assistant	**2014–2019**
University of California, Irvine	*Irvine, California*

REFEREED JOURNAL PUBLICATIONS

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P. P. P., Hawkins, G. E., Heathcote, A., Holmes, W. R., Krypotos, A.-M., Kupitz, C. N., Leite, F. P., Lerche, V., Lin, Y.-S., Logan, G. D., Palmeri, T. J., Starns, J. J., Trueblood, J. S., van Maanen, L., van Ravenzwaaij, D., Vandekerckhove, J., Visser, I., Voss, A., White, C. N., Wiecki, T. V., Rieskamp, J., & Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review, 26(4),* 1051–1069.

Bapat, A. N., Shafer-Skelton, A., Kupitz, C. N., & Golomb, J. D. (2017). Binding object features to locations: Does the “spatial congruency bias” update with object movement? *Attention, Perception, & Psychophysics, 79(6),* 1682–1694.

Shafer-Skelton, A., Kupitz, C. N., & Golomb, J. D. (2017). Object-location binding across a saccade: A retinotopic spatial congruency bias. *Attention, Perception, & Psychophysics, 79(3),* 765–781.

Shafer-Skelton, A., Kupitz, C., Tausif, A., & Golomb, J. (2015). Feature binding and eye movements: object identity is bound to retinotopic location regardless of stimulus complexity. *Journal of Vision, 15(12),* 1062.

Golomb, J. D., Kupitz, C. N., & Thiemann, C. T. (2014). The influence of object location on identity: A “spatial congruency bias.” *Journal of Experimental Psychology: General, 143(6),* 2262.

McCarthy, J. D., Kupitz, C., & Caplovitz, G. P. (2013). The binding ring illusion: Assimilation affects the perceived size of a circular array. *F1000Research,* 2.

REFEREED CONFERENCE PUBLICATIONS

Kupitz, C. N., Buschkuehl, M., Jaeggi, S. M., Jonides, J., Shah, P., & Vandekerckhove, J. (2015). A diffusion model account of the transfer-of-training effect. *Proceedings of the 37th Annual meeting of the Cognitive Science Society.*

ABSTRACT OF THE DISSERTATION

Applications of Hierarchical Bayesian Cognitive Modeling

By

Colin Nicholas Kupitz

Doctor of Philosophy in Cognitive Sciences

University of California, Irvine, 2020

Associate Professor Joachim Vandekerckhove, Chair

The hierarchical Bayesian approach to cognitive modeling often provides a quality of inference that cannot be matched with other analytical methods. In addition, the general approach is quite flexible, and can be utilized to great effect in many analytical settings. I illustrate these qualities in two applications of hierarchical Bayesian cognitive models. In the first application, I revisit a transfer-of-training study. First, I discuss a hierarchical cognitive model that describes the transfer-of-training data. I then illustrate how that hierarchical cognitive model can be further extended in order to create a cognitive latent variable model. Critically, this cognitive latent variable model directly models the latent effects of training and transfer on the cognitive parameters that drive participant behavior. I then provide an in depth analysis to illustrate how this cognitive latent variable model provides a quality of inference that far surpasses more standard analytical approaches. In the second application, I perform a cognitive meta-analysis on the spatial congruency bias literature. To do this, I extend the hierarchical Bayesian cognitive model into an integrative data analysis, creating a Model-based Integrative Data AnalysiS (MIDAS). Using this model, I create a model that is capable of simultaneously estimating cognitive effects at the individual, within-experiment, and between-experiment levels, which is what allows us to estimate cognitive effects in a meta-analytical setting.

Chapter 1

Introduction

The standard analytical approach in the Psychological sciences is statistical significance tests based on simple summary statistics. However, these traditional analyses often cannot distinguish qualitatively different sources of variability in the data. For example, suppose we were testing some generic treatment effect, and observe that participants in the treatment group respond more quickly that those in the control condition. Were they faster at processing task relevant information? Or, perhaps, have they cognitively adapted to show less caution on the task (either by shifting criterion or a change in speed-accuracy tradeoff)? Could it be the result of multiple changes in cognition?

These questions serve to illustrate the lack of interpretability that occasionally plagues traditional analyses – the results derived from summary statistics often lack psychological meaning. In a sense, traditional analysis can only answer *what* behavioral changes occurred – never *why* those changes happened. Making matters worse, this problematic interpretability is often exacerbated by averaging artefacts, which can lead to biased estimates and errors in inference (Heathcote, Brown, & Mewhort, 2000; see also Clark, 1973).

In recent years, some researchers have begun to address these issues through the application of cognitive process models. (e.g., Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Vandekerckhove, Tuerlinckx, & Lee, 2011; van Vugt & Jha, 2011).

Critically, these cognitive models are able to distinguish qualitatively different sources of variability in a way traditional analyses cannot. Cognitive models provide analyses with interpretable results, providing a mechanism of inference capable of distinguishing *why* any observed behavioral changes occurred.

In this dissertation, we primarily use cognitive process models as a form of measurement model – in particular, we focus on the Wiener Diffusion model (WDM), first described by Stone (1960). In essence, we utilize the WDM to decompose participant behavioral data into psychologically meaningful parameters. We then construct more complex models using those WDM parameters as a foundation, by introducing hierarchical and/or latent structural effects. Of particular note, all of our models are constructed in a fully Bayesian framework, which affords a number of statistical and inferential benefits, some of which we describe in Chapters 2 and 3. For a more comprehensive review of the advantages of a fully Bayesian framework, see (Lee & Wagenmakers, 2014; Vandekerckhove, Rouder, & Kruschke, 2018; Etz & Vandekerckhove, 2018).

In Chapter 2, we introduce a simple hierarchical extension of the WDM, creating a Hierarchical Cognitive Model (HCM), in order to capture the cognitive processes and their alterations in a Transfer-of-Training study. We also demonstrate how the HCM approach offers conclusions with an increased quality of psychological interpretability.

In Chapter 3, we first improve upon this hierarchical cognitive model, creating a more complex HCM to describe the transfer-of-training effects. Next, we extend this new HCM further, and create a Cognitive Latent Variable Model (CLVM) to describe these transfer-of-training effects. We then provide an in-depth analysis of how the CLVM improves upon the HCM, and further, how the CLVM affords conclusions regarding the transfer-of-training effects at the individual level. Finally, we demonstrate how the quality of these inferences cannot be obtained with other analytical approaches.

In Chapter 4, we demonstrate the inherent flexibility of the CLVM approach, extending it into the realm of Integrative Data Analysis, meaning that we include all the trial-level data from a number of experiments within the same encompassing model. With this extension, we create a Model-Based Integrative Data AnalysiS (MIDAS), and demonstrate how the MIDAS affords the ability to perform meta analyses of the cognitive effects driving behavior across multiple experiments. To provide a proof-of-concept for the MIDAS, we use this approach in order to meta analyze a number of experiments from the Spatial Congruency Bias literature, which we describe in more detail within Chapter 4.

Chapter 2

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

Abstract

We revisit a transfer-of-training study to analyze the data with a cognitive model. 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.

2.1 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, 2017). 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 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 et al., 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.

2.2 Data

We will revisit data by Buschkuehl et al. (2017). Here we describe only the subset of data that we will use. Other measures are described in Buschkuehl et al. (2017).

2.2.1 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.

2.2.2 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 independently generated 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.

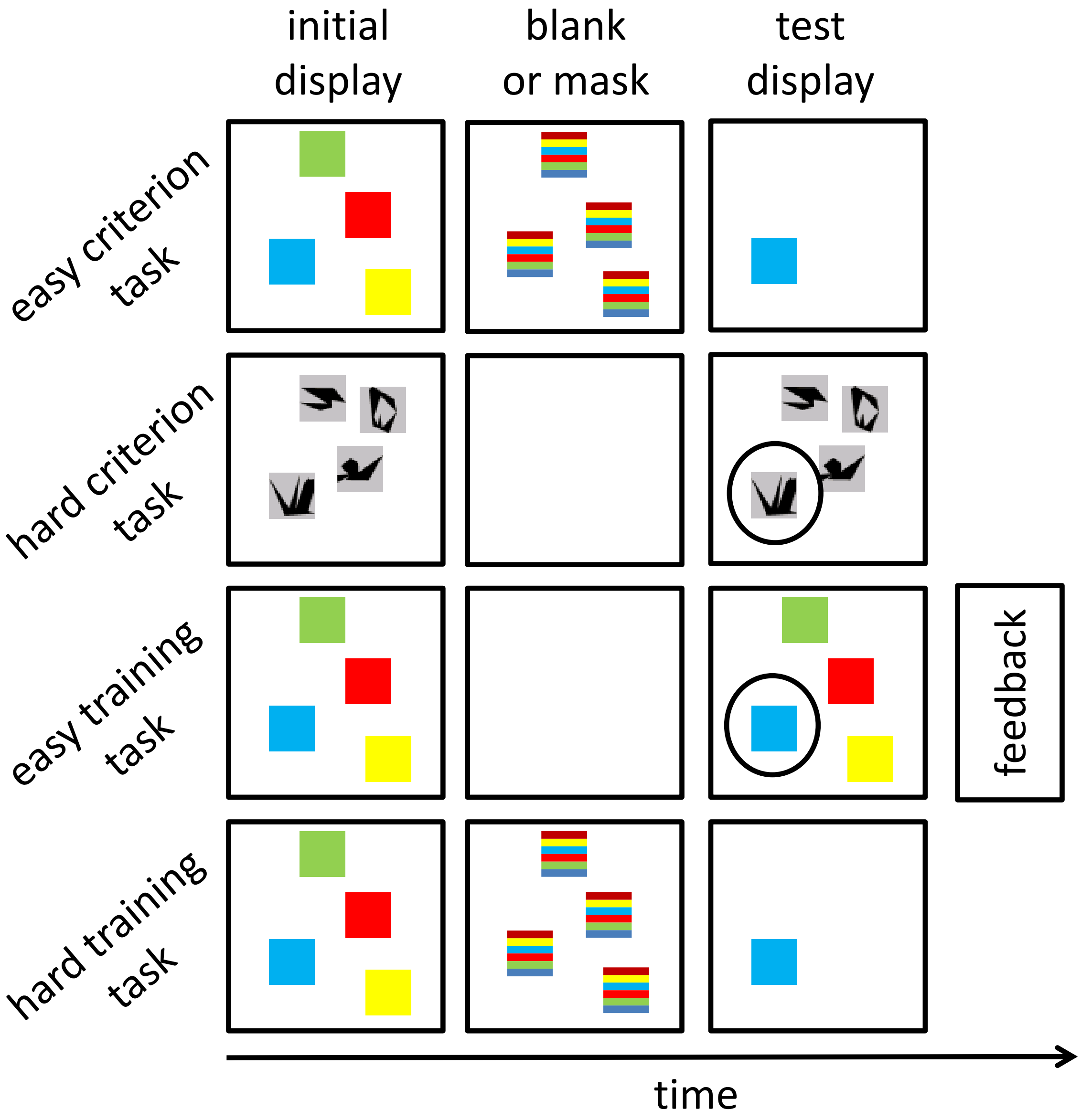

Figure 2.1: Example trials for each of the four tasks. The easy and the hard criterion tasks differ in the type of stimulus (color squares vs. shapes), the presence of masks, and the number of items remaining in the test display. The easy and hard training tasks differ only in the the presence of masks, the number of items remaining, and the presence of feedback. Note that the hard training task and easy criterion task are the same.

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 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).

2.3 Diffusion model

Our modeling analysis is based on an hierarchical diffusion model for two-choice reaction times (Vandekerckhove et al., 2011), extending a model first described by Stone (1960) and popularized in a slightly extended form by Ratcliff (1978).

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 2.2 illustrates the process. The parameters of interest are α, the amount of information required before a decision is made (which captures the speed-accuracy trade-off); β, the a-priori bias that a participant might have towards one or the other response; τ, the non-decision time including time for encoding the stimulus and executing the motor response; and δ, 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, δ), while simultaneously accounting for non-skill based changes in task performance and speed (using α, β, and τ as nuisance variables).

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., μ_4^τ 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., γ_5^β is the average offset in a-priori bias β 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 θ, and the likelihood function $\mathbf{y}_{kipt} \sim W(\alpha, \beta, \tau, \delta)$. The

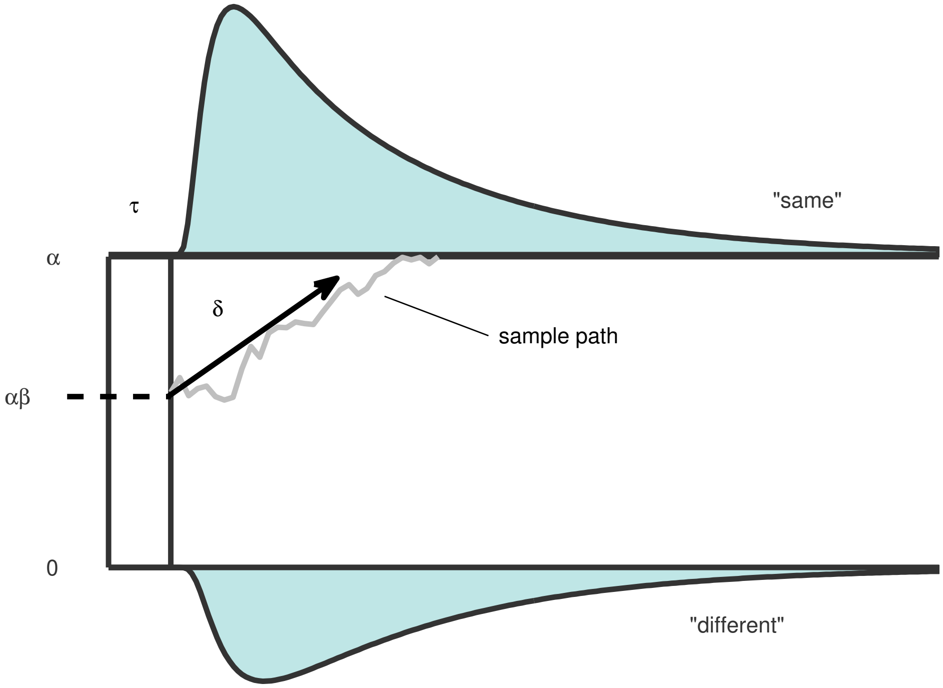

Figure 2.2: A graphical representation of the Wiener diffusion model. The accumulation process begins at evidence value $\alpha\beta$ and unfolds with an average increase of δ per second until a boundary at α or 0 is reached. τ is an additive time constant for nondecisional processes. The shaded area is the model-predicted probability density function over response and response time, $W(\alpha, \beta, \tau, \delta)$.

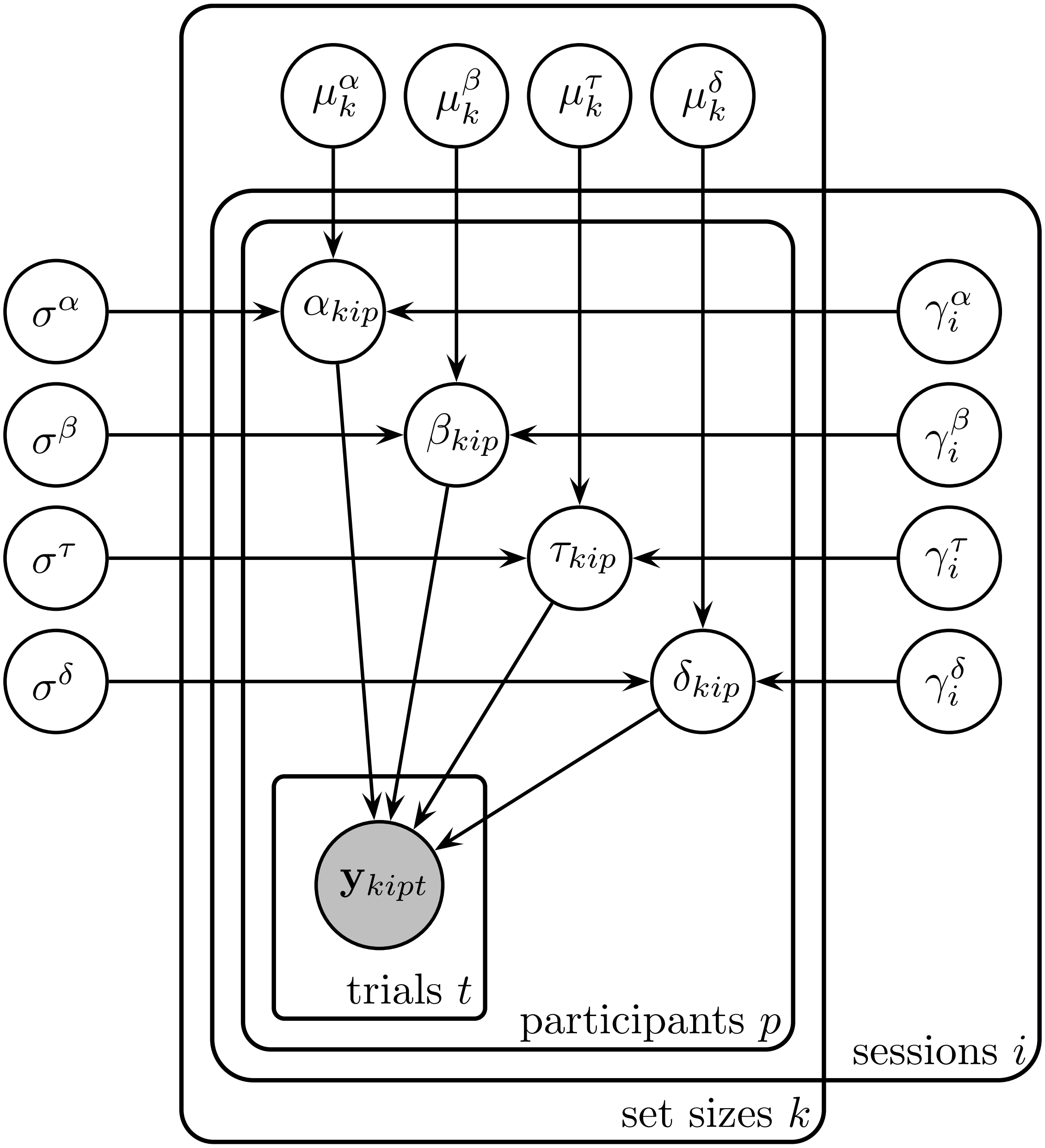

Figure 2.3: A graphical representation of our exploratory hierarchical diffusion model. Parameters μ indicate the set-size-specific population mean of each parameter; parameters γ indicate the effect of session on each parameter; and parameters σ indicate the between-person variability in each parameter. Node $\mathbf{y}_{kipt}$ is the t^{th} choice response time data point for participant p in session i with set size k. For example, the supposed distribution of δ_{kip} is normal with mean $\mu_k^\delta + \gamma_i^\delta$ and standard deviation σ^δ, 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.

likelihood function is defined as the first passage time distribution of a Wiener process with constant boundaries (e.g., Tuerlinckx, 2004).

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 2.3 gives a graphical model representation of the model we used. In this graph, variables are represented by nodes, downstrean (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.

2.4 Modeling results

2.4.1 Training

Posterior distributions of the parameters are displayed in Figure 2.4. 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 α decreases as participants begin to trade accuracy for speed. The a-priori bias level β stays constant and around 0.5, as induced by the experimental paradigm. Nondecision time τ steadily decreases over sessions. Drift rate δ 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.[1]

The right panels in Figure 2.4 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 from Hick's law), and underscoring that set size was taken into account in our analyses.

2.4.2 Transfer

Figure 2.5 shows 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 α and non-decision time τ. Additionally, we also see a stronger increase in drift rate δ. This is particularly interesting given that δ 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.

2.5 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 prac-

[1] We omit from here the details of a second model in which 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 μ_k^δ, person-specific error term ε_p^δ, regression weight ζ, and i the session number. In this model, $p(\zeta < 0|\mathbf{y}) \approx 0.007$, indicating a positive linear 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 slightly less pronounced 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 $\hat{\zeta}_{\text{HARD}} \approx .015$).

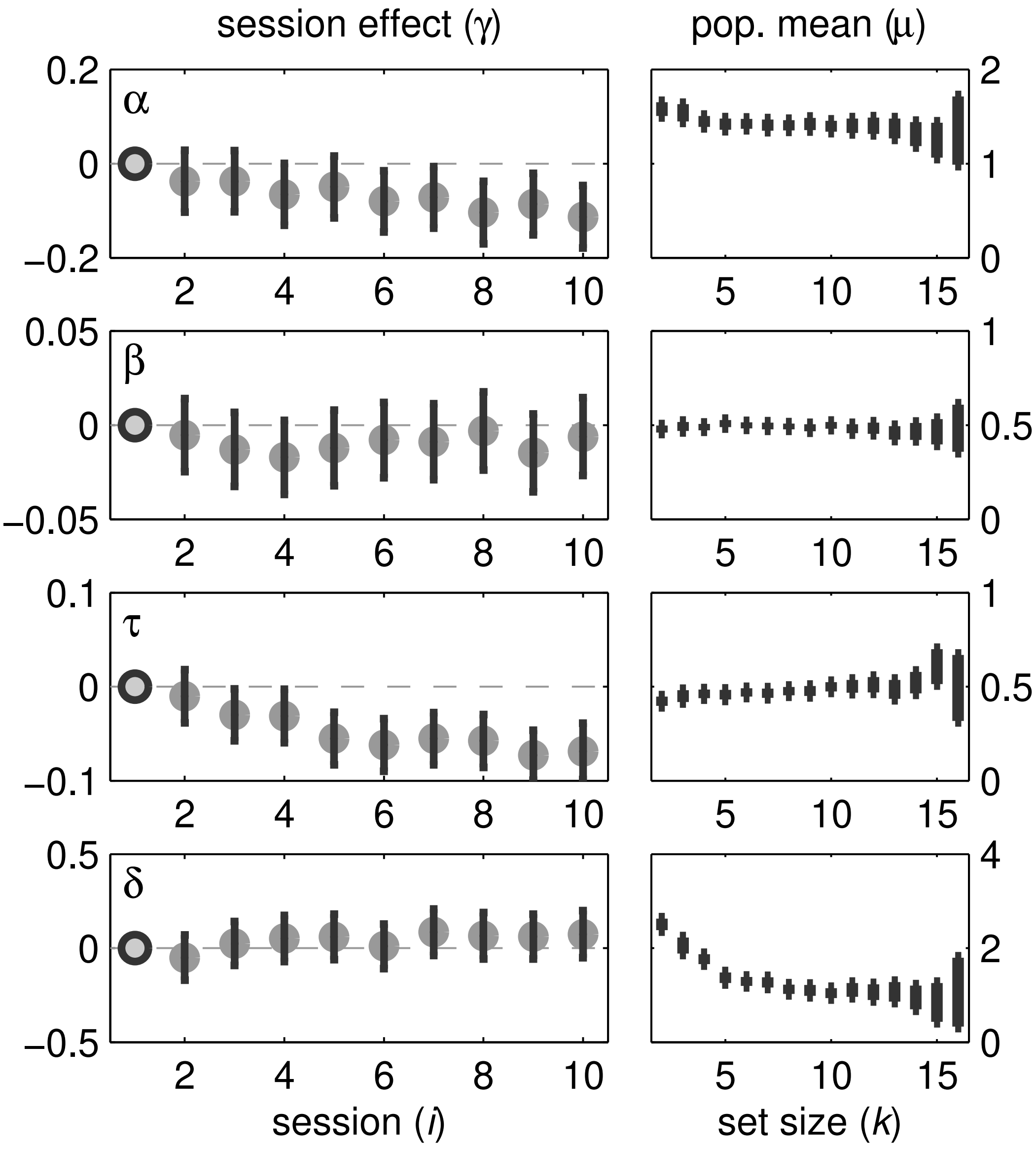

Figure 2.4: Right panels: Posterior distributions of the population means μ_k of the four diffusion model parameters as a function of set size k. Posterior uncertainty, indicated by the 95% credibility interval, is larger for the highest set sizes because few participants reached that level of difficulty. 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. Left panels: Posterior distributions of the session-specific offset terms γ_i as a function of sessions i. The leftmost marker is the first session, which is marked differently because it was the only training session held in the lab. Ignoring the first session, we observe a decrease in boundary separation α and in nondecision time τ, and a slight increase in drift rate δ.

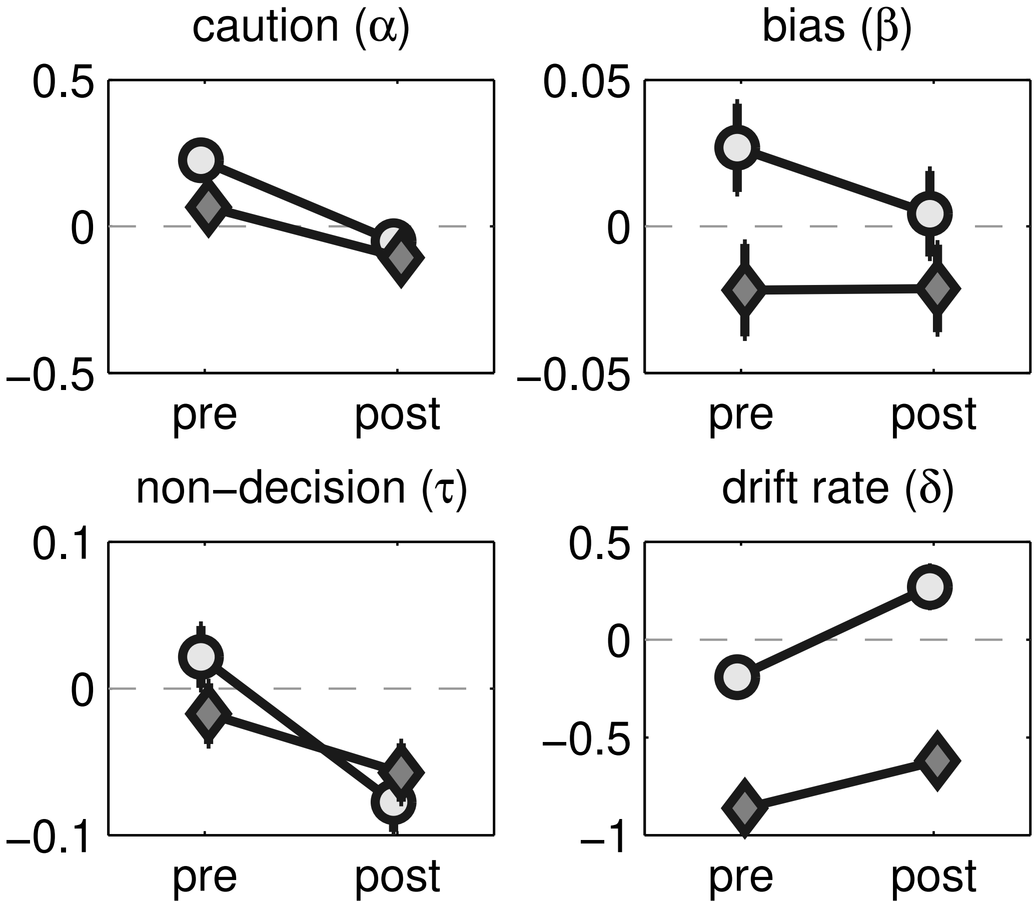

Figure 2.5: Diffusion model parameters estimates from the transfer tasks. Session is given on the horizontal axes. Circles represent the easy criterion task, diamonds the hard criterion task. Top left: α is seen to start above the reference level in the pre-training test and to end below it in the post-training test. Top right: β 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 stab1le. Bottom left: τ start level with the reference point but decreases markedly after training. Bottom right: δ 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—δ increases after training as well, indicating transfer of training.

tice effect consists of simultanous 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 et al., 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.

2.6 Authors' Note

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. This project was partly supported by grants #1230118 from NSF's Methods, Measurements, and Statistics panel and #48192 from the John Templeton Foundation to Joachim Vandekerckhove.

Chapter 3

A cognitive latent variable approach to the transfer-of-training effect

3.1 Abstract

We revisit a transfer-of-training study using a cognitive latent variable 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. In line with previous findings regarding the effects of practice, our diffusion model analysis indicates that participants simultaneously adapted speed-accuracy trade-off, increased their non-decisional response speed, and increased their speed of information processing. We observed that all three of these adaptations transferred to a similar but untrained outcome task. Our findings were further supported by an extended model: With an hierarchical regression model, the magnitude of individual-level transfer to the untrained task was related to individual task improvement across the diffusion parameters.

3.2 Introduction

Cognitive training is an increasingly popular research subject, and yet it has become mired in controversy (e.g., Dougherty, Hamovitz, & Tidwell, 2015; Jaeggi et al., 2014; Morrison & Chein, 2011; Oberauer et al., 2003; Rode et al., 2014). The ideal goal of cognitive training is to improve a portion of the cognitive processes underlying performance on a trained task (*training*), in hopes that improvements will also be observed, to some degree, in non-trained tasks (*transfer-of-training*, hereafter: transfer).

Many of these studies have focused on training working memory (WM), the cognitive system that transiently stores and manipulates information relevant to complex cognitive tasks (Baddeley, 1992). These studies are likely motivated by the pervasive nature of WM in psychological literature, as WM has been reported to influence a large number of cognitive processes, such as reasoning ability (Kyllonen & Christal, 1990), reading comprehension (Daneman & Carpenter, 1980), and fluid intelligence (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Engle, Tuholski, Laughlin, & Conway, 1999).

Unfortunately however, the field of WM training is currently quite divided in its opinion on the efficacy of WM training as a device for transfer. A number of studies have shown successful transfer of training to unrelated tasks (Klingberg et al., 2005; Chein & Morrison, 2010; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008), while a similar number have shown successful training that failed to transfer to other tasks (Thompson et al., 2013; Redick et al., 2013; Chooi & Thompson, 2012).

Customarily, WM literature has focused on measures of accuracy, speed, or capacity. The standard approach to analyzing these measures is statistical significance tests based on simple summary statistics. However, these traditional analyses often cannot distinguish qualitatively different sources of variability in the data. This analytical shortcoming is particularly obvious when studying transfer data. For example, we might observe that participants who are exposed to a training intervention on average respond more quickly in the last session compared to the first—Were they faster at processing task relevant information? Or, perhaps, have they cognitively adapted to show less caution on the task (either by shifting criterion or a change in speed-accuracy tradeoff)? Could it be the result of multiple changes in cognition?

These questions serve to illustrate the lack of interpretability that occasionally plagues traditional analyses – the results derived from summary statistics often lack psychological meaning. In a sense, traditional analysis can only answer *what* behavioral changes occurred – never *why* those changes happened. Making matters worse, this problematic interpretability is often exacerbated by averaging artefacts, which can lead to biased estimates and errors in inference (Heathcote et al., 2000; see also Clark, 1973).

In recent years, some researchers have begun to address these issues through the application of cognitive process models. (e.g., Dutilh et al., 2009; Rouder et al., 2005; Vandekerckhove et al., 2011; van Vugt & Jha, 2011). Critically, these cognitive models are able to distinguish qualitatively different sources of variability in a way traditional analyses cannot. Cognitive models provide analyses with interpretable results, providing a mechanism of inference capable of distinguishing *why* any observed behavioral changes occurred.

In the present paper, we will focus on the use of a *hierarchical cognitive model* (HCM) and a *cognitive latent variable model* (CLVM), which combine the interpretability of cognitive models with the statistical power afforded by hierarchical and latent variable models. Cognitive models are attempts to explain a complex cognitive process by decomposing it into psychologically interpretable parameters. Hierarchical models are statistical constructs that allow for individual differences in observations, pooling information across groups of participants. Latent variable models (LVM) take this one step further, and use the correlational pattern between multiple basic tasks to uncover *structured individual differences* among participants—they infer unobserved factors that support stable differences across those individuals. On their own, each approach has limitations: hierarchical models and LVMs do not allow for process-based conclusions, and cognitive models cannot make inferences about the higher-order latent properties of a system. However, these approaches can be combined, by using cognitive model to decompose a complex data set into interpretable variables that are fed into a hierarchical model or LVM, which provides a process-driven and psychologically interpretable HCM (e.g. Rouder et al., 2005) or CLVM (Vandekerckhove, 2014).

These hybrid models are particularly well suited to analyzing training and transfer effects. The HCM affords conclusions about the effects of WM training *on the underlying cognitive processes* in both training and transfer measures, at the population level. Furthering this, restructuring the HCM into a CLVM provides an even more appealing approach, as the CLVM provides estimates of individual training effects, and further, estimates of *direct individual transfer*.

In the current study, we revisit a visual WM (VWM) training and transfer dataset. The VWM task was the *change-detection task* (CDT), popularized by e.g., Luck and Vogel (1997). In this paradigm, an array is presented, participants hold it in memory for a short period of time, a second array is presented, and participants make a 'same' or 'different' judgment. We reanalyze the data with both an HCM, and a CLVM, and find that WM training does indeed improve some of the underlying cognitive processes, and notably, these improvements are not wholly constrained to the trained task.

3.3 Data and Tasks

We revisit data by Buschkuehl et al. (2017), describing only the subset of data that we will use. We have $N = 34$ participants, divided randomly and evenly into two training groups ($N = 17$ each).

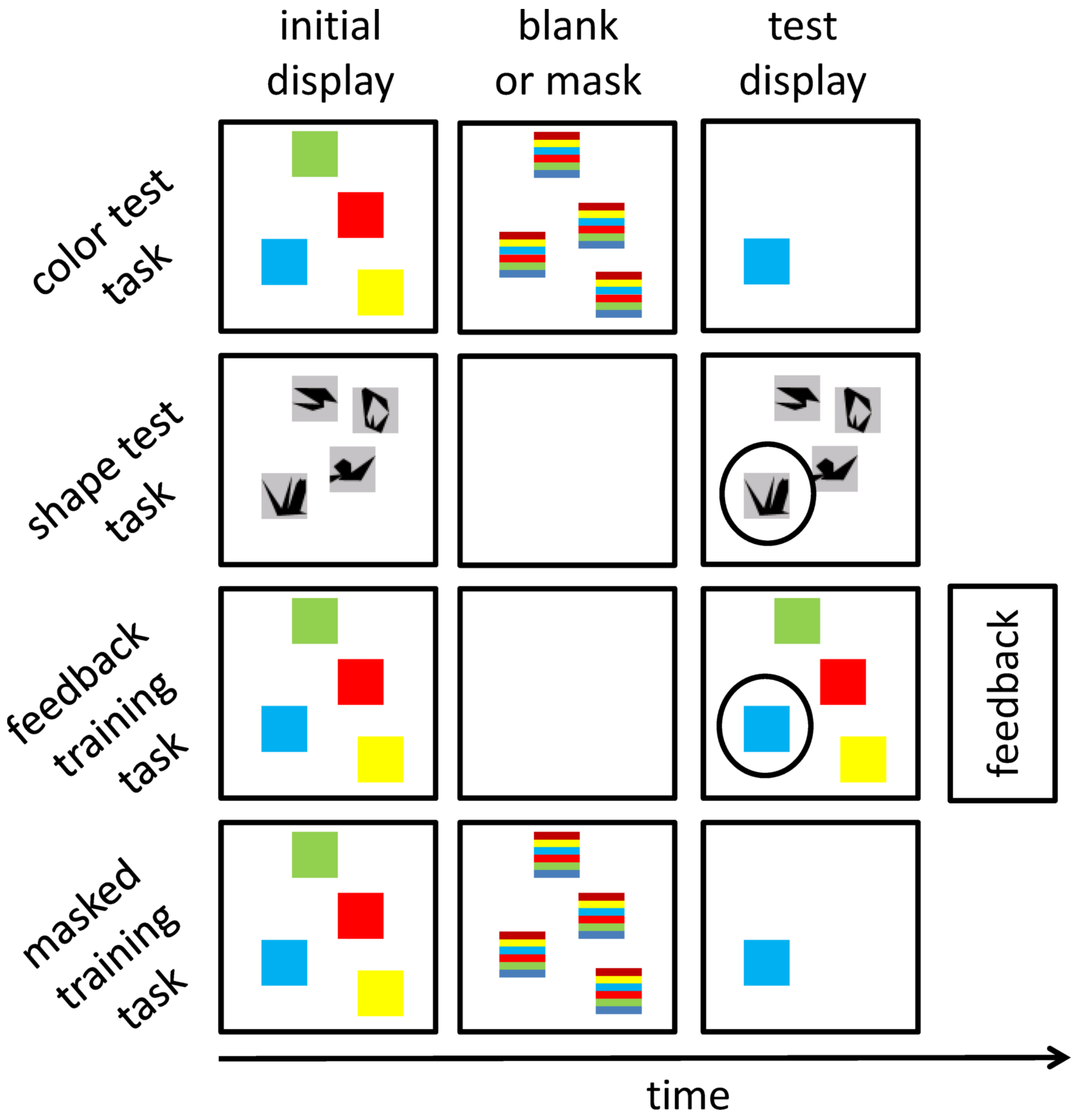

Figure 3.1: Example trials for each of the four tasks. The first two rows are the test tasks, which differ in the type of stimulus (color squares vs. shapes), masking, and test array configuration. The bottom two rows are the training tasks, which differ in masking, test array configuration, and the presence of feedback. Both groups of participants did both tests tasks in a pre- and post-test session. The Feedback training group trained for ten sessions on the feedback training task; the Masked training group trained for ten sessions on the masked training task. Note that the masked training task and color test task are the same. Figure taken from Kupitz et al. (2015), used with permission.

3.3.1 Participants

Forty-five participants were recruited for the study from two university communities, and were randomly assigned to one of two interventions. Four participants withdrew from the study following the pre-test session; five participants were excluded from the analyses due to irregularities in their training schedules; two participants were excluded for failing to complete all of the pre- and post-test tasks. This left a total of 17 participants in each of the two training groups.

3.3.2 Procedure and Tasks

The discussion of the experimental procedure is based on Kupitz et al. (2015).[1] Participants performed two 'test' tasks both before and after training. The test tasks were a color CDT (hereafter color test) and a shape CDT (hereafter shape test), both described later. The difference from pre- to post-test in the color test and shape test is used to obtain measures of transfer.

Participants were randomly assigned to one of two training groups. One training group was assigned a CDT with a feedback mechanism (hereafter feedback task); the other training group was assigned a CDT with a visual mask included (hereafter masked training). The tasks are described below, and an illustration is provided in figure Figure 3.1.

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. Participants were asked to complete ten training sessions within 14 days, and were only allowed to complete two training

[1] We change the name of the tasks for ease of discussion, but closely follow the structure of Kupitz et al. (2015)'s description here otherwise.

sessions in one day if they had not trained on the previous day. Participants were required to send the training data generated after each session via email to the laboratory. Following the training period, participants were tested again in the laboratory on the criterion tasks in order to evaluate the impact of the intervention.

Color Test

Each trial of the color test task began with a fixation cross presented in the center of the screen for 1,000ms. Then, an array of colored squares was presented on a screen with a dark gray background for 250ms, followed by a 200ms blank screen, and then a set of masks for 700ms, directly covering the colored square display locations. Each mask was a colored striped square newly generated at each trial from the colors used in the array 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 judgment was made by the participant. There was a 1,000ms break between trials.

The computer program provided task instructions, and then the participants performed 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, feedback was no longer provided. The task consisted of 150 randomly ordered test trials: 15 change trials and 15 no-change trials for each of the possible set sizes, 2, 4, 6, 8, and 10.

Shape Test

The shape test task was an altered version of color test described above. The stimulus array was presented for 500ms and followed by a 1,000ms blank screen. During the judgment phase the entire array was shown again; the shape to be judged was indicated by a black circle.

Instead of colored squares, random black shapes were used (as in Jaeggi et al., 2003, but black in color and smaller in size). Participants were asked to indicate change or no-change for the encircled shape. After a judgment was made, the next trial began immediately.

Feedback Training

The feedback training task was similar to the color test task described above with some notable 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. Finally, 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. The initial training session began with a set size of 2. Performance was evaluated after each block, and if accuracy was higher than 85%, the set size was increased by one; similarly, if the accuracy dropped below 70%, set size was reduced by one. Set size was reduced by 2 at the beginning of every new training session (a 'warm-up' period). The program had a maximum set size of 20, but no participants reached a set size higher than 16.

Masked Training

The masked training task was identical to the shape test described above. Thus, it differed from the feedback task in a few key ways: no feedback provided was provided, a mask was utilized, and only one of the squares was shown in the test display (to preclude any context or configuration effects).

One training group was assigned a Change Detection Task (CDT) with a feedback mechanism (hereafter feedback training); the other training group was assigned a CDT with a visual mask included (hereafter masked training). All participants performed two ‘test’ tasks both before and after training. The test tasks were a color CDT (hereafter color test) and a shape CDT (hereafter shape test). The difference from pre- to post-test in each test task is used to obtain measures of transfer. An illustration of each task is provided in figure Figure 3.1. The post-test was administered within 14 days of the pre-test, and participants completed 10 training sessions within that timeframe.

Data preprocessing

As this is a retrospective analysis, we must account for the fact that the experiment was designed without a model-based analysis in mind. Accordingly, we went through extensive data cleaning steps, cutting a non-trivial portion of the data. Broadly, our data cleaning strategy served to assure that all remaining data are germane to the model we will fit. The data points that we remove are those about which the model makes no predictions (because other processes are at work). We believe such data-curation steps are not only reasonable, but necessary, for proper model based inference.

First, we removed warm-up trials (the first 40 trials) in each training session. Second, we removed data from the first of the training sessions. This session which was the only one to take place in a laboratory setting and displayed qualitatively different behavior from the other sessions (see also Kupitz et al., 2015). Finally, to guard our estimates against “fast outliers” (Ratcliff, 1993), particularly in the conditions where only few data points were collected, we removed the fastest 5% of RTs within any given condition. We do not report here the results of models that do not receive this exact outlier treatment, other than to note that leaving these trials in the dataset does not produce any meaningful changes in the *population-level* inferences obtained – these cuts serve entirely to improve *individual-level* inferences.

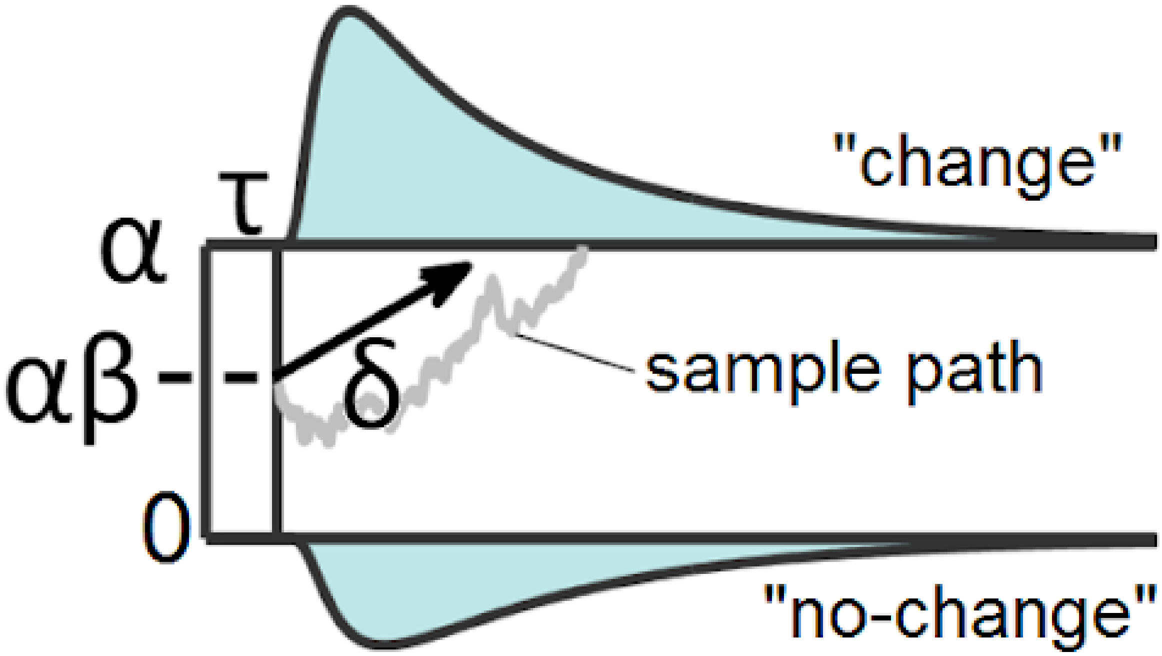

Figure 3.2: A graphical representation of the Wiener diffusion model. Evidence is accumulated at an average rate of δ units per second, and begins at the value $\alpha\beta$. The process stops when evidence reaches the boundary at α or 0, and then a response is made. The additive time constant τ accounts for nondecisional processes. The shaded area is the model-predicted probability density function over response and response time, $W(\alpha, \beta, \tau, \delta)$. Figure taken from Vandekerckhove (2009), used with permission.

3.4 The Diffusion Model as a Measurement Model

All of our analyses are based on an hierarchical extension of the diffusion model for two-choice reaction times, first introduced by Vandekerckhove et al. (2011; see also Vandekerckhove, Tuerlinckx, & Lee, 2008) as an extension of the model described by Stone (1960). The later analyses introduce latent parameters into the hierarchy to create a *cognitive latent variable model* (CLVM; Vandekerckhove, 2014).

The diffusion model assumes that participants accumulate information sequentially, and only execute a response once enough information has been gathered; the process is illustrated in Figure 3.2. The parameters of interest are α, the total amount of information required to make a decision (which captures cognitive strategies that lead to speed-accuracy trade-off); β, the a-priori bias favoring one of the choices over the other; τ, the non-decision time, including both stimulus encoding and motor response execution time (Nunez, Srinivasan, & Vandekerckhove, 2015; Nunez, Vandekerckhove, & Srinivasan, 2017); and δ, the *drift rate* or rate of information accumulation within a trial. Thus, we can account for changes in task performance and speed that arise due to improved generalizable skill (δ), improved task-specific skill (τ), and non-skill based changes (α, β) through this parameterization.

Notational Conventions

We will use θ to represent any of the four cognitive parameters, α, β, τ, and δ. Subscripts will index the level of any of the independent variables in which the parameter may be nested: We will use k for set size, s for session, b for block (a grouping of sessions), p for participant, g for training group ($g = 1$ feedback training group, $g = 2$ masked training group), and t for task ($t = 1$ for training, $t = 2$ for color test, $t = 3$ for shape test).

Finally, superscripts will index latent variable relationships. For example, if μ is the population mean of some distribution, and δ is a cognitive parameter, then μ^{δ} is the mean of the distribution from which parameter δ_k is drawn. The index k refers to a given set size, and the lack of other subscripts indicates that this parameter does not depend on block, session, participant, or trial.

3.5 Hierarchical Diffusion Model

In our first model, we will decompose the observed parameters into constituent additive components. Separately for each task t, we will assume a fixed effect μ^{θ}_{tk} of set size for each parameter θ, meaning that each set size k has a unique mean value for each parameter (e.g., $\mu^{\tau}_{t=1,k=3}$ is the average nondecision time for trials with set size 3 in the training condition).

We are predominantly interested in the overall effect of training on each task. As such, we now segment the training sessions into blocks – this reduces the effect of session-to-session variability within participants, thus providing well informed estimates of the training effect at the population level. We create an early block, $b = 1$ (sessions 2–4), a middle block (sessions 5–7), $b = 2$ and a late block, $b = 3$ (sessions 8–10). This smoothing strategy is somewhat informed by our earlier observation that session-to-session changes tend to be noisy but small on average (Kupitz et al., 2015). Conveniently, this organization allows us to use the middle block to ensure the model remains identifiable, while still obtaining well informed estimates of the setsize effects, as the overall training effects for the training tasks are best represented as the change from ‘early’ (block 1) to ‘late’ (block 3) training.

We assume a fixed offset γ^{θ}_{bgt} for each parameter at each combination of block b, group g, and task t (e.g., $\gamma^{\beta}_{b=1,g=1,t=1}$ is the average offset from the mean a-priori bias β in block 1 of the training task for the feedback training group). For reasons of model identification, the training offsets are defined relative to the middle training block for the appropriate training group (so offset $\gamma_{b=2,t=1} = 0$ for all parameters). Similarly, for both groups, the post-test offsets are defined relative to the pre-test (so offset $\gamma_{b=1,t=2} = 0$ and $\gamma_{b=1,t=3} = 0$).

We also allow for a fixed participant effect ρ^{θ}_{bpt}, which is drawn from a normal distribution with mean γ^{θ}_{bgt} for each participant p, block b, and task t combination. Thus, this additive decomposition, applied for each diffusion model parameter θ, is fully described by the set of equations

$$\begin{aligned}
\theta_{kspt} &= \mu^{\theta}_{kt} + \rho^{\theta}_{pbgt} + \varepsilon^{\theta}_{spt} \\
\rho^{\theta}_{pbt} &\sim N\big(\gamma^{\theta}_{bgt}, \sigma^{\theta}_{\rho}\big) \\
\varepsilon^{\theta}_{spt} &\sim N\big(0, \sigma^{\theta}\big)
\end{aligned}$$

and the likelihood function $\mathbf{y}_{kspt} \sim W(\alpha_{pskt}, \beta_{pskt}, \tau_{pskt}, \delta_{pskt})$. The likelihood function is defined as the first passage time distribution of a Wiener process with constant boundaries (Feller, 1970; Navarro & Fuss, 2009).

We fit this model simultaneously to the training data and the test tasks, allowing for different session offsets for each parameter in each of the test sessions. Importantly, we also define a new parameter, ϕ^{θ}_{gt}, to represent the population training effect on parameter θ for group g in task t. For the training tasks, this is equal to the change in the offset parameter, γ, from the early to the late block. For the test tasks, this is equal to the change in the offset parameter from pre- to post-test. Together, these training effects are described by the equation

$$\psi^{\theta}_{gt} = \begin{cases} \gamma^{\theta}_{b=3,gt} - \gamma^{\theta}_{b=1,gt} & \text{if training } (t = 1) \\ \gamma^{\theta}_{b=2,gt} - \gamma^{\theta}_{b=1,gt} & \text{if test } (t = 2 \text{ or } t = 3) \end{cases} \tag{3.1}$$

That is, the training effect ψ is the difference in magnitude of the offset γ between the final and initial block, for a given parameter θ, group g, and task t. For example, given this parameterization, a negative ψ^{α} implies a decrease in caution, α, at the population level due to training.

Model Implementation Details

Using an hierarchical Bayesian framework, we implemented the model as in Vandekerckhove et al. (2011). We drew fifteen chains of 500 samples from the joint posterior distribution of all model parameters 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. Every model we discuss in this paper is implemented in this manner.

3.6 Hierarchical Diffusion Model – Results

In general, throughout our analyses we treat Set Size as a nuisance variable: we always account for it, but are uninterested in the parameter estimates. Accordingly, the modeling constraints due to Set Size, and the Set Size results have been moved to the appendix.

3.6.1 Training Effects

Figure 3.3 displays the posterior distributions of the training effect, ψ^{θ}, for both groups across all three tasks.

Feedback training group

Training effects for the feedback training group are shown in blue in Figure 3.3. Boundary separation, α, decreases over the course of training. This effect is even more pronounced in transfer to both test tasks. A-priori bias, β, shows virtually no change in the training. There may be a small shift in bias towards the "no-change" decision in the color test. Nondecision time, τ, may decrease over the training. A definite decrease is observed with a much larger magnitude in both the color and shape tests, which suggests that there is a training effect for most of the population, and that the effect transfers to both test tasks. Drift rate, δ, does not change over the course of training. There may be a small increase in the drift rate from pre- to post-test in both test tasks, but it if does exist it is not strong.

Masked training group

Training effects for the masked training group are shown in red in Figure 3.3. Boundary separation, α, may decrease slightly as a result of training at the population level. However, there is a notable decrease in caution for the color test. A-priori bias, β, shows virtually no change in the training. There may be a small shift in bias towards the "no-change" decision in the color test. Nondecision time, τ, may decrease over the training. A definite decrease is observed with a much larger magnitude in both the color and shape tests, which suggests that there is a training effect for most of the population, and that the effect transfers to both test tasks. Drift rate, δ, appears to show a slight decrease over training. This decrease may be driven by the small decrease in α. However, at the population level, it appears that the drift rate shows a small increase in the color task, and a definite increase in the shape task.[2]

[2] The decrease observed in δ here is in direct contrast to the earlier findings for this project, in Kupitz et al., 2015. We believe this difference is likely driven (at least partially) by the data restrictions used herein, namely the removal of the training session in the laboratory, as well as the 'warm-up' blocks from each at-home training session. This suggest that the majority of changes in δ at the population level occur early in training.

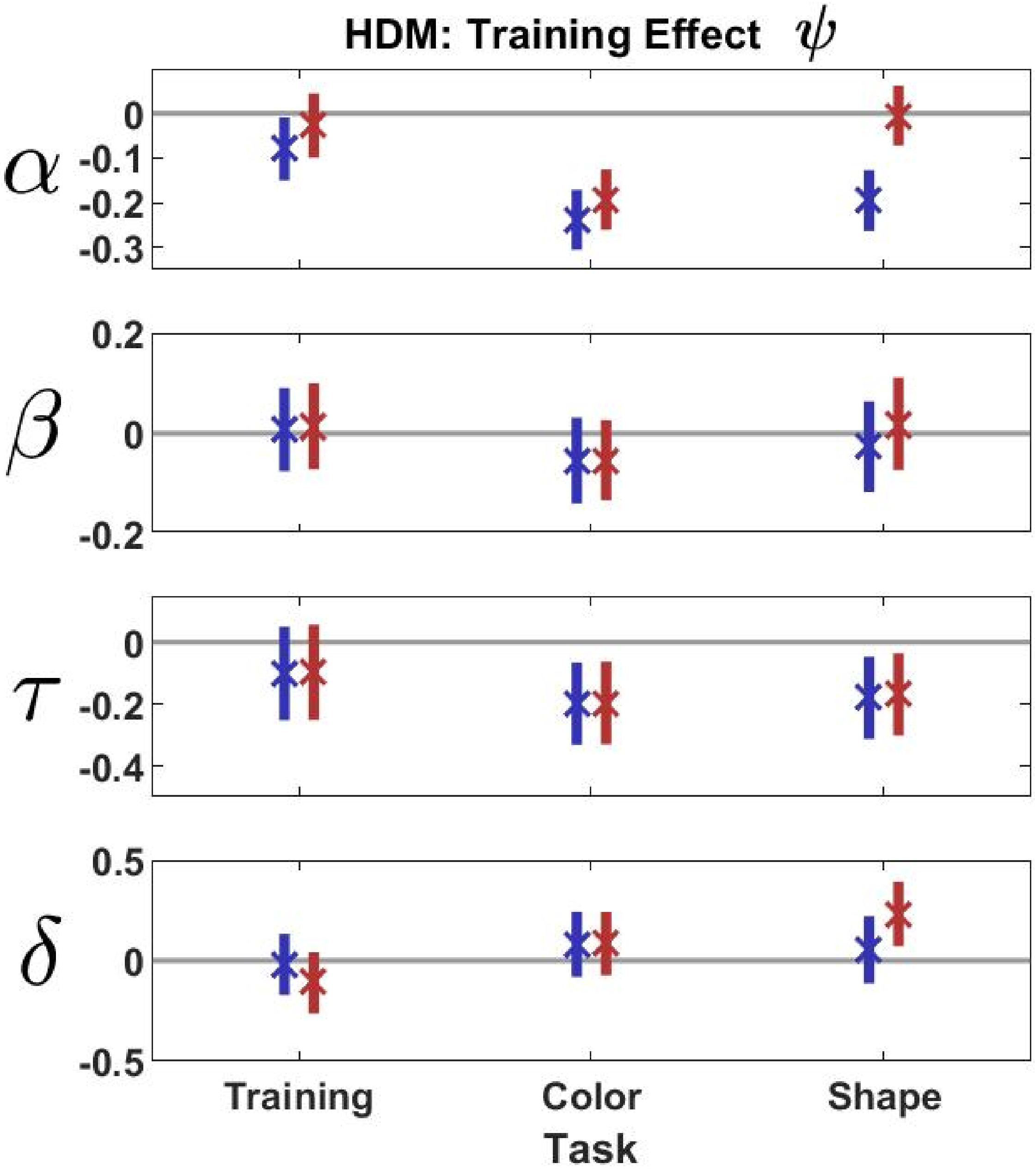

Figure 3.3: Results of the Hierarchical Diffusion Model (HDM). Posterior distributions of the population mean training effect of the four diffusion model parameters, for each group g (Feedback in Blue, Masked in Red) across all tasks t. Posterior uncertainty is indicated by the 99% credibility interval, with the mean of the posterior distributions shown as X's.

3.7 Cognitive Latent Variable Model

The results of the Hierarchical Diffusion Model (HDM) analysis at the population level are clear. However, with the structure of the previous model we cannot make any claims as to what degree, if any, the training effect *directly transfers* to the test tasks. Rather, we can only observe the effect of training on each task, and assume that similar effects imply transfer-of-training. Further, these results have only been concerned with population level effects, but there are large individual differences in training and transfer effects.

To address these issues, we make the following model modifications to the Hierarchical Diffusion Model (HDM). We estimate individual level training effects, ι_p^θ. We then introduce a linear regression structure, with ψ_{gt}^θ as the intercept, and κ_{gt}^θ as the slope. We use ι_p^θ as the regressors, producing estimates of individual level transfer effects to each separate test task, λ_{pt}^θ, as the regressands. The addition of these new latent structures creates a Cognitive Latent Variable Model (CLVM). Critically, it is the CLVM structure that allows us to make inferences about direct transfer-of-training at the individual level.

The equation updates to make these model changes are as follows:

$$
\begin{aligned}
\psi_{gt}^\theta &\sim N\big(mean = 0, SD = 2^{-1/2}\big) \\
\kappa_{gt}^\theta &\sim N\big(mean = 0, SD = 2^{-1/2}\big) \\
\iota_p^\theta &= \rho_{b=3,pt}^\theta - \rho_{b=1,pt}^\theta && (3.2) \\
\lambda_{pt}^\theta &\sim N\big(\psi_{gt}^\theta + \kappa_{gt}^\theta \times \iota_{pg}^\theta, \sigma_\rho^\theta\big) && (3.3)
\end{aligned}
$$

where $\iota_{\cdot g}^\theta$ is the population average training effect on θ for group g. We chose to estimate

λ_{pt}^{θ} as a draw from a Normal distribution centered on the regression line, with standard deviation $\sigma_{\lambda}^{\theta}$, to allow for individual differences in degree of transfer-of-training.

With this design, the intercept ψ_{gt}^{θ} represents the population average transfer effect for θ, the change from pre- to post-test for group g in task t. The slope κ_{gt}^{θ} then represents the degree to which direct transfer-of-training is observed. For example, a positive value of κ_{gt}^{θ} implies that participants who show a larger than average training effect ι_{p}^{θ} would also show a larger than average transfer effect λ_{pt}^{θ}.

3.8 Cognitive Latent Variable Model – Results

3.8.1 Training Effects – Population Level

First, we compare our average population level training effects to those obtained in the basic HDM, to determine if the model updates had any effect on the basic population results. For the training task, ψ should be virtually identical between the HDM and CLVM. For the test tasks, ψ now represents the intercept in the transfer-of-training linear regression design; accordingly, we should only observe major changes in ψ, from the HDM to the CLVM, when the CLVM results support direct transfer-of-training (captured in the new slope parameter, κ).

Feedback training group

Training effects for the feedback training group are shown in blue in the left side of Figure 3.4. These effects are quite consistent with those from the HDM. In the training task, as predicted we observe that there is essentially no difference in the ψ parameter between the HDM and CLVM results. In the Color test, there are no noteworthy changes in the population level training effects from the HDM to the CLVM. In the Shape test, we observe that at the population level, the negative training effect on α observed in the HDM is still present in the CLVM, but with a decrease in magnitude. As expected, this difference is driven by a direct transfer-of-training effect (we discussed all direction transfer-of-training effects in depth in the following section). Similarly, we find that the decrease in τ observed in the HDM is no longer present in the CLVM, and again, this is driven by a direct transfer-of-training effect.

Masked training group

Training effects for the masked training group are shown in red in the left side of Figure 3.4. These effects are quite consistent with those from the HDM. In the training task, as predicted we observe that there is essentially no difference in the ψ parameter between the HDM and CLVM results. In the Color test, we observe the following at the population level. The effect of training on *a-priori* bias, β, is definitively negative in the CLVM, meaning that participants were predisposed to responding "no-change" before the trial occurred over the course of training. As in the Feedback training group, this change in effect is driven by a direct transfer-of-training effect.

3.8.2 Direct Transfer-of-Training Results

We now discuss the posterior distributions for the regression slope parameters κ^{θ}_{gt}, provided on the right side of Figure 3.4. As a reminder, the regression design implies that these slope values represent the population average rate at which a change in an individual's training effect (ι^{θ}_{p}) directly influences their transfer effect (λ^{θ}_{pt}). We discuss these results separately for each training group and test task combination.

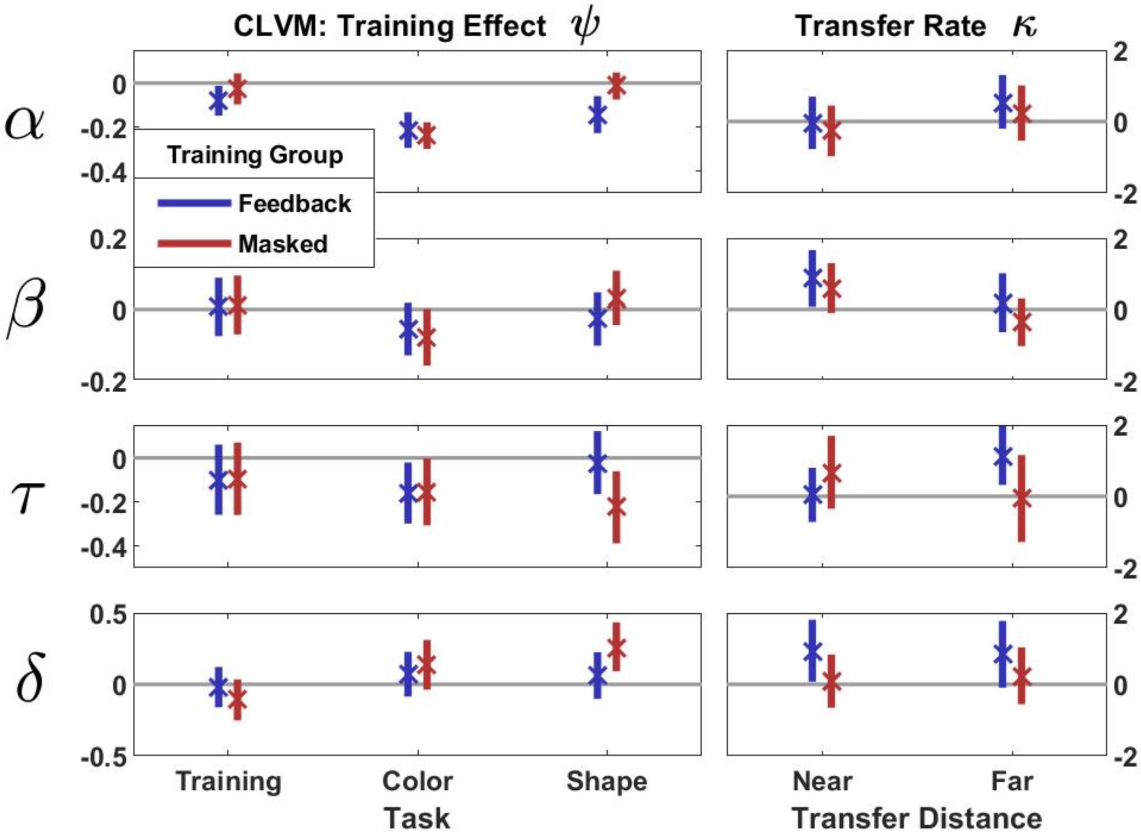

Figure 3.4: Results of the Cognitive Latent Variable Model (CLVM). Posterior distributions of the population mean training effect of the four diffusion model parameters, for each group g across all tasks t. The left portion of this figure is a direct analogue to figure 3.3, and intended to illustrate there is little difference between the two. Posterior uncertainty is indicated by the 95% credibility interval, with the mean of the posterior distributions shown as X's.

Results for Direct Transfer for the Feedback Group in the Color Test The feedback group regression slope parameters κ^{θ}_{gt} are shown in blue in the right side of Figure 3.4. Boundary separation, κ^{α}, has a large degree of uncertainty, and its 95% CI is centered close to 0. There is no direct transfer here. A-priori bias, κ^{β}, is entirely positive with a mean of approximately 1. There is direct transfer-of-training at the individual level here. Nondecision time, κ^{τ}, has a large degree of uncertainty, and its 95% CI is centered close to 0. There is no direct transfer here. Drift rate, κ^{δ}, is entirely positive with a mean of approximately 1. There is direct transfer-of-training at the individual level here. Critically, this means that the subjects who improved at the training task also demonstrated an proportional improvement in the color test task.

Results for Direct Transfer for the Masked Group in the Color Test The masked group regression slope parameters κ^{θ}_{gt} are shown in red in Figure 3.4. Boundary separation, κ^{α}, has a large degree of uncertainty, and its 95% CI is centered close to 0. There is no direct transfer here. A-priori bias, κ^{β}, is entirely positive. There is direct transfer here. Nondecision time, κ^{τ}, has a large degree of uncertainty but suggests a small, positive effect. This is weak evidence for direct transfer. Drift rate, κ^{δ}, has a large degree of uncertainty, and its 95% CI is centered close to 0. There is no direct transfer here.

Results for Direct Transfer for the Feedback Training Group in the Shape Test Boundary separation, κ^{α}, is predominantly positive, but does include 0. This is weak evidence of direct transfer-of-training at the individual level. A-priori bias, κ^{β}, has a large degree of uncertainty, and its 95% CI is centered close to 0. There is no direct transfer here. Nondecision time, κ^{τ}, is entirely positive with a mean of approximately 1. There is direct transfer-of-training at the individual level here. Drift rate, κ^{δ}, is almost entirely positive with a mean of approximately 1. There is direct transfer-of-training at the individual level here.

Results for Direct Transfer for the Masked Training Group in the Shape Test In all four WDM parameters, the direct transfer-of-training estimates for the Masked training group are centered close to 0. There is no evidence of direct transfer-of-training at the individual level for the Masked Training group.

3.8.3 Model Diagnostics: Posterior Predictive Distributions

We evaluate model fit through posterior predictive assessment (Gelman, Meng, & Stern, 1996). The ultimate goal of the CLVM is to accurately capture transfer *at the individual level*. Accordingly, our diagnostic visualizations focus on individual level posterior predictive distributions. We provide the posterior predictive distribution for the median RT (figure 3.5) and accuracy (figure 3.6) for each participant in each task.

These figures show that the model is acceptably capturing the observed behavior, at the individual level. Only a handful of the observed participant median RTs or accuracies fall outside the posterior predictive distribution interval.

3.9 The Unique Benefits of a CLVM

In the previous section, we presented a suitable model to describe participant behavior in this transfer-of-training study. Now, we illustrate how our unique approach – the CLVM – allows for scientific inferences that are not possible in the more traditional approaches. Indeed, we will show that when compared to other analysis paradigms, the inferences afforded by a CLVM are *stronger statistically and more interesting theoretically.*

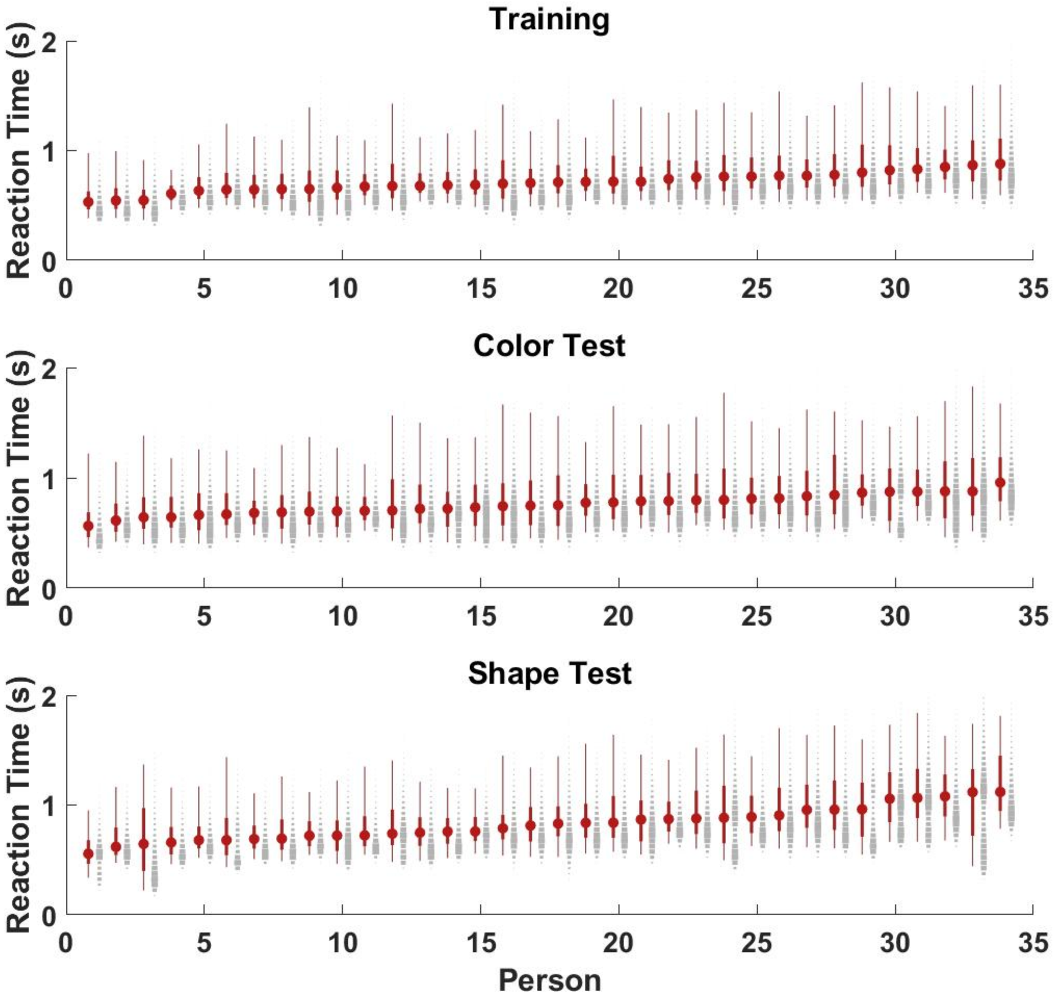

Figure 3.5: PPD of the median RT for each participant, in each task. Red X is true median RT. In black, the PPD mean, 95% and 99% credibility intervals, respectively the X, thick, and thin bars.

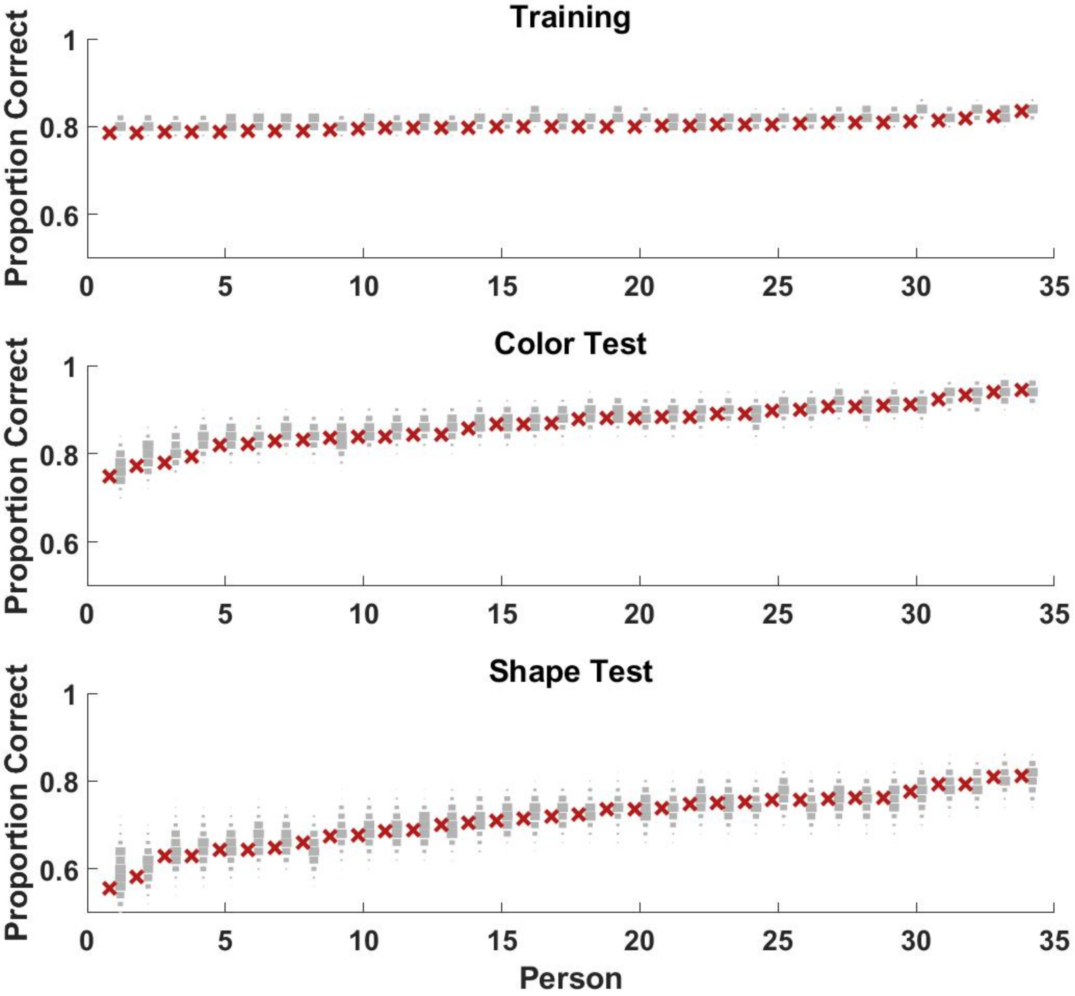

Figure 3.6: PPD of the accuracy for each participant, in each task. Red X is true accuracy. In black, the PPD mean, 95% and 99% credibility intervals, respectively the X, thick, and thin bars.

3.9.1 The Role of Hierarchy

The first critical advantage conferred by our analysis is a direct result of the well designed hierarchical structure. In the CLVM, we hierarchically combine effects for set size, task, and training group. This structure is paramount to our approach, as a well-designed hierarchy provides more accurate effect estimates at *both* the population and individual levels. To illustrate the benefits of a hierarchical design, we now compare the HDM and CLVM to two new models.

The first comparison model has no hierarchy: every parameter is estimated independently. We call this the Flat model, and it is *entirely* described by the following equation:

$$\theta_{kbpt} \sim N\left(\mu^{\theta}, \sigma^{\theta}\right),$$

where both μ^{θ} and σ^{θ} are the static prior values (described in the Appendix) appropriate for each WDM parameter θ.

The second comparison model uses a very basic form of hierarchy, and thus we refer to it as Basic. In this model, we estimate a hierarchical population mean, μ^{θ}_{kbgt}, for each unique combination of task t, group g, block b, set size k. Note how this differs from the full models (the HDM and CLVM): each $tgbk$ combination has a unique, and critically, *independently estimated* mean. Each unique θ_{kbpt} is then estimated as a draw from a Normal distribution centered on μ^{θ}_{kbgt} with task dependent variance σ^{θ}_{t}. Thus, Basic is described by the following equations:

$$
\begin{aligned}
\mu^{\theta}_{bgt} &\sim N\big(\mu^{\theta}, \sigma^{\theta}\big) & \theta_{kbpt} &\sim N\big(\mu^{\theta}_{bgt}, \sigma^{\theta}_{t}\big) \\
\sigma^{\alpha}_{t} &\sim Gamma(1, 0.5) & \sigma^{\beta}_{t} &\sim Gamma(1, 0.25) \\
\sigma^{\tau}_{t} &\sim Gamma(1, 0.25) & \sigma^{\delta}_{t} &\sim Gamma(1, 1)
\end{aligned}
$$

where both μ^{θ} and σ^{θ} are static prior values identical to those used in Flat, provided in the appendix.

To illustrate the advantages of an hierarchical design, we now compare the fit of each model. To do this, once again we use posterior predictive assessment (Gelman et al., 1996). Independently for each model, we generate the PPD for the RT median for each unique combination of setsize k, block b, person p, and task t. We show scatterplots of the (mean) predicted median RT against the observed median RT, as well as the linear regression of said data, for all four models in figure 3.7.[3]

Benefits of Hierarchy Comparing Flat and Basic illustrates the power of having *any* form of hierarchy, even a basic version. By simply introducing a hierarchical structure, the model must consider all participants in two contexts simultaneously: as an individual *and* as a member of the population. By introducing a hierarchical structure, the model has new parameters in which to distribute the same amount of fundamental uncertainty that exists within the data. The result is a model that obtains better effect estimates at *both* the population and individual levels.

[3] While we use posterior means here, the posterior medians show the same patterns.

Benefits of Constraints The differences between Basic and the HDM are more subtle, but still illustrative. The HDM constrains the hierarchy of Basic, obtaining separate effect estimates for session, task, and setsize, and not allowing for the any interaction effects. This increases the overall variance of the observed PPDs, slightly decreasing the model fit. However, this minor loss in model fit is worthwhile. The overall model accuracy remains high, and introducing said hierarchical constraints of the HDM provides direct estimates of the training effects we are interested – without compromising model accuracy. This comparison illustrates that the hierarchical constraints of the HDM are reasonable, because there is no evidence of systematic misfit. A properly instantiated hierarchical structure can provide interesting effect estimates *without compromising model fit*.

Benefits of Structure Finally, the differences between the HDM and CLVM are minimal, as the CLVM is a reorganization of the model parameters with only minor alterations to the structure. This illustrates an important nuance in the CLVM approach: once the proper hierarchical structure is determined, the effect estimates (e.g. session effect γ in HDM) can often be reparameterized to introduce additional latent structures. Critically, these latent structures can be used to introduce more theoretically interesting effects – at both the individual and population level. The CLVM allows us to answer psychologically meaningful questions while maintaining model fit.[4]

[4] During our analyses, we also inspected the same PP graph comparisons between the models, when considering only one task at a time. There were no noteworthy differences between the overall model comparisons described above, and the individual task comparisons, so we omit those figures. Similarly, we inspected the same PP graph comparisons between the models, but with accuracy as the measure of interest. There were no noteworthy differences between the accuracy and RT median comparisons, so we omit those figures.

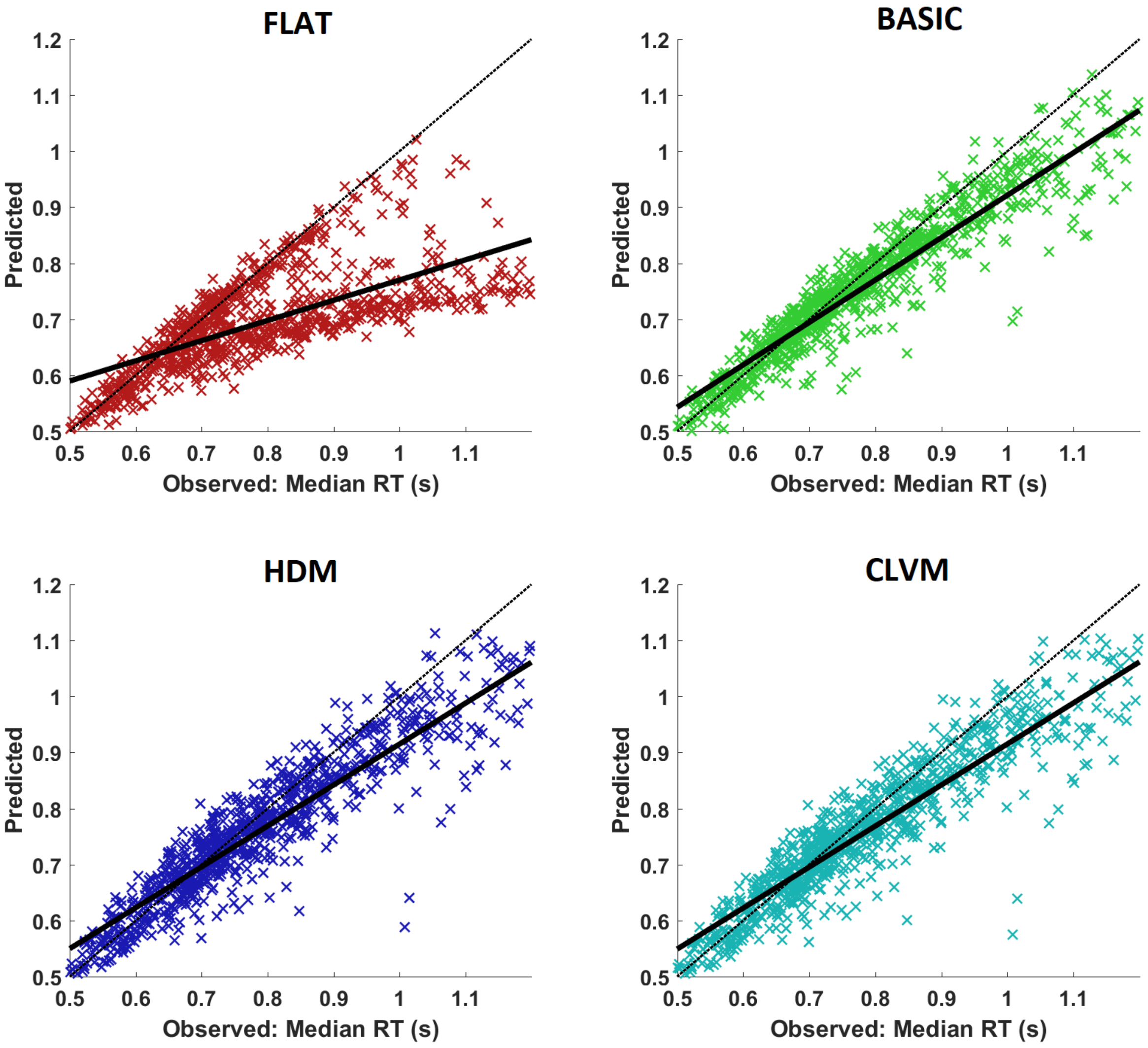

Figure 3.7: Comparison of model fits. Scatterplot showing the true behavioral RT median on the X axis, and the PP RT median on the Y axis. Each individual X represents one unique combination of setsize k, block b, person p, and task t. The thick line represents the linear regression of all shown data. The dashed line shows the 'perfect' PP line ($Y = X$).

3.9.2 The Role of Generative Cognitive Models

At their core, all cognitive models share the same goal: capturing the underlying cognitive processes driving human behavior. The major strength of cognitive models is therefore that they decompose observed behavior into parameters that have psychological interpretations – *cognitive parameters have meaning in the context of human behavior.* Removing the cognitive model foundation from our model removes the approximation of human behavior, thereby removing any psychological meaning in the parameters.

The previous section illustrated that even a well founded cognitive model, such as the diffusion model, does not prove sufficient to model the observed behavior – the model must also take into account the structured individual differences. However, the inverse of this is usually true as well: in most cases an LVM alone is insufficient to model the observed behavior in a way that is psychologically interpretable. Exceptions occur when the concept of interest is reasonably assumed to be normally distributed, and thus the mean and variance estimate can have meaningful interpretations (e.g., IQ). However, outside of such exceptions, LVM analyses often suffer from the same issue as do traditional analyses: the parameters have poor interpretability. Suppose we used an LVM, analyzing this dataset with a purely descriptive model. We could certainly draw conclusions about the structure of changes in RT and accuracy due to training, and determine if those effects show transfer. But regardless of the results, we would not be able to conclude which cognitive processes were driving the observed changes in behavior.

While this point is reasonably well demonstrated by theory alone, we also provide a more concrete demonstration here. Thus, we now compare the CLVM model to an identical model sans the assumed diffusion process. We simply refer to this new model as the LVM, and it takes a purely descriptive approach, assuming that the log of each RT measurement is an independent draw from a normal distribution about a unique condition mean, and participant specific variance. We chose to use this log-normal RT distribution as it is a relatively common approach to analyzing RT data (Ratcliff, 1993). Similarly, we treat the accuracy of each trial as a draw from a Bernoulli distribution with a unique condition mean. For both RT and accuracy, the unique condition means share the exact same hierarchical latent structure used in the CLVM.

Once again, we will compare the ability of the models to capture participant behavior by inspecting their PPDs. The RT median PPDs of the LVM shown in figure 3.8. Especially in comparison to the CLVM model fit (previously provided in Figure 3.7), it is quite clear that the LVM falls short. In virtually all participants, the LVM is overestimates the median RT. This illustrates an additional potential benefit provided by the use of cognitive models: in many cases, they are more practical. Provided the chosen cognitive model is appropriate for the data – meaning it reasonably approximates the cognitive processes driving behavior – it will often outperform simple descriptive processes.

A similar analysis modeling accuracy with a Bernoulli distribution is not shown, as it resulted in a good fit of the accuracy-only model.

3.10 Discussion

The current paper investigates the efficacy of cognitive training and transfer-of-training in a visual WM task with a novel approach, utilizing a Cognitive Latent Variable Model (CLVM). The CLVM addresses many of the shortcomings of more traditional analyses by allowing for joint conclusions about both the task related cognitive processes and their structured individual differences. The CLVM provides conclusions regarding the underlying cognitive processes involved in the task – *and critically, how those cognitive processes change with training.*

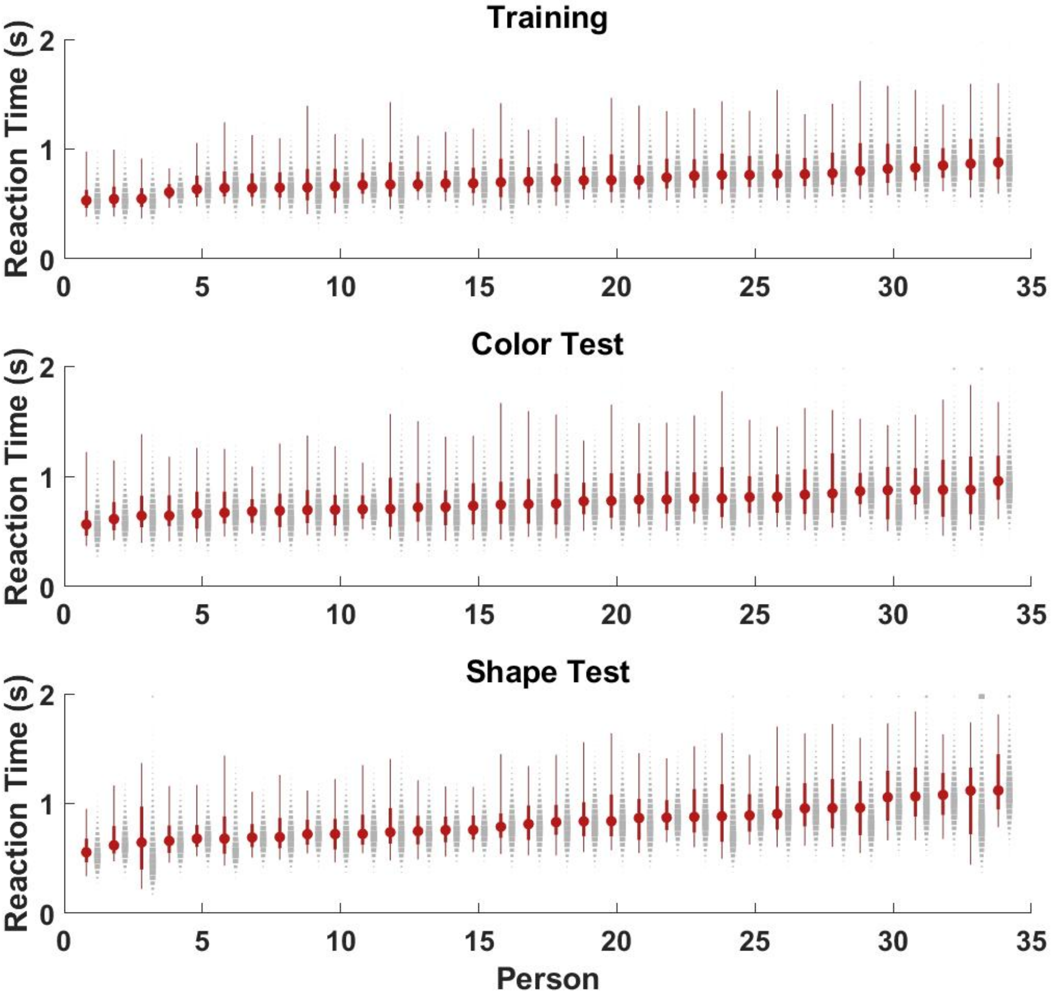

Figure 3.8: For the LVM, posterior predictive distributions of the median RT for each participant, across each task. The true median RT is shown as a red X. The mean of the PP distribution is a black X, and the uncertainty of the PP is illustrated using the 95% and 99% credibility intervals, respectively the thick and thin black bars.

The study had two training groups, the Feedback and Masked groups, both of which performed similar variants of a basic color based Change Detection Task (CDT). Both training groups also performed a pre- and post-test session for two test CDT tasks, one to measure near transfer (Color Test), and one to measure far transfer (Shape Test).

The diffusion model serves as the cognitive model basis for the analyses we performed. Importantly, the parameters of the diffusion model have semantically meaningful interpretations. Our first analysis uses a Hierarchical Diffusion Model (HDM) to investigate training effects at the population level.

In the Feedback Training Group, we observe the following. We find a negative training effect for boundary separation, α, meaning participants traded in accuracy for speed. This effect transfers to both test tasks. We find no effects regarding β, the A-priori bias. The Nondecision time, τ, has a negative training effect, meaning participants were faster at visually encoding the stimuli and/or executing their response. It is unclear how much of this effect transfers to the test tasks. Finally, we find a small but positive training effect for the drift rate, δ. This effect appears to transfer to both test tasks.

In the Masked Training Group, we observe the following. We find no notable effects for α during training, but do observe a decrease in α for the shape test. We find no effects regarding β during training, however, there is a large shift in bias towards the "change" decision in the shape test. As in the Feedback Group, we find a negative training effect for τ, meaning participants were faster at visually encoding the stimuli and/or executing their response. Similar effects were observed in both test tasks, but with varying magnitudes. Finally, we find a small but positive training effect for the drift rate, δ. This effect appears to transfer to the color test more than the shape test.

Our second analysis uses a Cognitive Latent Variable Model (CLVM), extending the HDM to investigate training and transfer-of-training effects at the individual level. Our extension involved introducing a linear regression into the model itself, using individual training effects as the regressors, and individual transfer-of-training effects as the regressands. By using this structure the estimated intercept, Ψ, captures the population training effects, while the estimated slope , ω, provides an estimate of the direct transfer-of-training rate at the individual level. At the population level, we find that the results of the CLVM agree with the population effects found in the HDM analysis.

Of note, the CLVM results also find direct transfer-of-training at the individual level, for a few select diffusion parameters. We observed that individual changes in β show direct transfer-of-training to both tests tasks. This implies that when an individual performed a strategy shift during training, they would likely perform the same shift in the test tasks. We also observed that in the masked group, individual changes in δ show direct transfer-of-training to the color test. This is unsurprising, given that the color test task is the same task that the masked group trained on. However, this finding serves to illustrate that the model design is capable of capturing direct transfer-of-training effects, should they exist.

We verified that the CLVM reasonably captures participant behavior at the individual level (and thus also the population level), by comparing Posterior Predictive Distributions of reaction times to the actual behavioral data. Overall, said PPD comparisons clearly illustrate that the CLVM is accurately capturing participant behavior.

3.11 Conclusions

We have demonstrated the unique advantages the CLVM approach has over other analyses. We have shown that a properly instantiated hierarchical design provides effect estimates that are more accurate at both the population and individual levels. Furthermore, we illustrated that a well-designed hierarchical model can be *restructured* into a latent variable model, thereby obtaining parameters that have psychologically meaningful interpretations without compromising model fit. We have also discussed why a well designed latent variable model is typically not enough on its own – in most research the parameters will have little psychological meaning, and thus should be combined with a generative cognitive model. It is for these reasons that we believe the CLVM approach is particularly appropriate in training and transfer-of-training research.

It is the unique combination of an hierarchical structure built upon a strong cognitive model foundation that makes the CLVM such a powerful analysis technique. The cognitive model basis provides to decompose the dataset into psychologically interpretable parameters, and the latent variable model design captures structured individual differences. By combining these two traditionally disparate approaches, we overcome the individual weaknesses of each. The CLVM allowed us to make conclusions about both the effects of WM training on the underlying cognitive processes, and further, the structured individual differences in those training effects – conclusions which no other analysis can provide.

3.12 Appendix – Static Model Priors

In both the Flat and Basic models, we use the following static prior values:

$$\begin{aligned}
\mu^{\alpha} &= 1.25 & \sigma^{\alpha} &= 10^{-1/2} \\
\mu^{\beta} &= 0.5 & \sigma^{\beta} &= 20^{-1/2} \\
\mu^{\tau} &= 0.1 & \sigma^{\tau} &= 20^{-1/2} \\
\mu^{\delta} &= 0 & \sigma^{\delta} &= 1
\end{aligned}$$

3.13 Appendix – Modeling Constraints due to Set Size

Another major modeling issue we faced was the treatment of setsize effects. The staircasing procedure used on setsize in addition to large individual differences precludes estimating the effect of setsize at the individual level: we must treat it as a population level effect. This is exacerbated by differences in relative frequencies of each setsize between training groups, and further, by the relatively sparse data of the pre- and post-test sessions (as compared to the training data). Thus, we decide to treat setsize as a nuisance variable: we account for it within the model, assuming that the effect of setsize is constant across all tasks.

We do this by essentially "giving up" the middle block of training sessions, not allowing for any effects at the individual level. This provides the model with a large amount of data used solely to determine the setsize effects of each parameter, thereby producing posterior effect estimates with a low degree of uncertainty.

However, by taking this approach, we reintroduced the issue of the Tau anchoring effect – this time, in the population setsize effect estimates. Removing any individual level effect from the middle sessions means that the estimate of Tau for each setsize will be anchored to the fastest RT found in that setsize, across all participants. This created the unfortunate side effect that the setsize effect estimates for Tau were not strictly increasing as setsize increased – which does not hold, theoretically (an increase in the number of items to process should strictly increase the amount of required processing time). Additionally, we only have even setsizes (2, 4, 6, 8, and 10) in the test tasks.

Given these modeling constraints, we converged on the following solution. We estimate the even setsize effects for all parameters freely, with the restriction that Tau must increase as setsize increases. We then treat each odd setsize effect as being equal to the average of the setsize below, and the setsize above. Our reasoning for this selection is two-fold. First, we only have the even setsizes for all four CDT tasks – and thus, we should obtain the most informed setsize effect estimates for these setsizes. Second, this design forces the model to account for adjacent setsizes when estimating Tau – once again removing the Tau anchoring issue.

3.14 Appendix – Hierarchical Diffusion Model Results: Set Size

Figure 3.9 displays the posterior distributions of the setsize means, μ_k^θ. Notably, both α and β are relatively stable across setsizes. As expected, τ shows a small positive slope, as the addition of stimuli should inherently increase the amount of nondecision time. Similarly, as expected δ shows a small negative slope, as the CDT becomes more difficult as the number of stimuli increases.

3.15 Authors' Note

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Greater detail can be found in Buschkuehl et al. (2017). This project was partly supported by grants #1230118 from NSF's Methods, Measurements, and Statistics panel and #48192 from the John Templeton Foundation to Joachim Vandekerckhove.

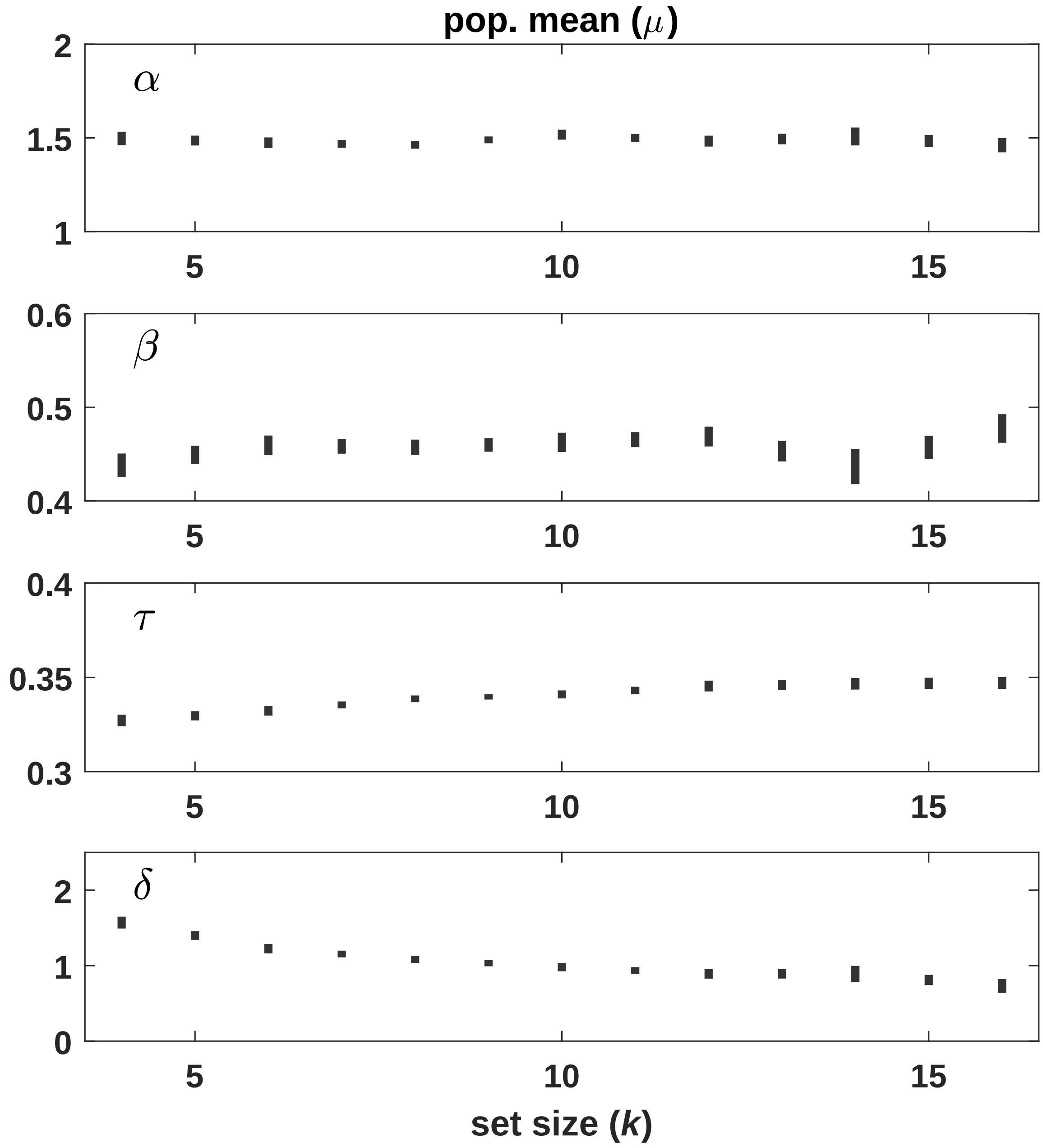

Figure 3.9: Posterior distributions of the population means μ_k of the four diffusion model parameters as a function of set size k. Posterior uncertainty, indicated by the 99% credibility interval, is larger for the highest set sizes because few participants reached that level of difficulty. There is a marked decrease in drift rate δ as set size increases, and an increase in non-decision time τ as set size increases. Both caution α and a-priori bias β appear to be relatively stable across set sizes.

Chapter 4

A model-based integrative data analysis of spatial congruency bias

4.1 Abstract

Integrative data analysis and model-based data analysis share a main goal: to draw stronger inferences from empirical data. Through Bayesian hierarchical modeling, these two methods can be combined, leading to an even more powerful method. We describe the model-based integrative data analysis strategy and apply it to a set of 15 experiments testing the spatial congruency bias effect.

4.2 Introduction

The use of meta-analysis has become increasingly prevalent in psychological research. Researchers in psychology can mean very different things by meta-analysis, including what is more generally known as a *systematic review* (Jones, 2007), an integrative data analysis

(Curran & Hussong, 2009), and a number of retrospective analyses that rely primarily on published summary statistics (Hedges, 1992; Sutton & Abrams, 2001). Regardless of the exact method used, the ultimate goal of these analyses is always the same: meta-analyses aggregate the results of multiple related studies in an effort to obtain an improved estimate of the effect of some manipulation, intervention, or some other comparison. Combining the results of multiple studies improves the precision of such estimate and consequently enhances the researcher's ability to accurately discriminate between theoretical accounts and make confident claims about the efficacy of treatments and manipulations. This improvement over standard hypothesis testing procedures makes the meta-analysis a powerful analytical tool with wide appeal. Unfortunately, however, meta-analysis also enhances some of the issues that plague much of the psychological literature.

Publication bias

One such issue is that of *publication bias*. Because manuscripts that report so-called 'positive' results (i.e., that meet the somewhat arbitrary standard of $p < .05$) are, all other things being equal, more likely to be judged suitable for publication. As a result, the scientific record is not a faithful or representative record of the many studies researchers conduct (Ioannidis, 2005; Young, Ioannidis, & Al-Ubaydli, 2008). The detrimental effects of publication bias on the literature are not only visible in present-day literature (Etz & Vandekerckhove, 2016; Rosenthal, 1979), but they were predicted 60 years ago when null hypothesis testing had become commonplace (Sterling, 1959). Despite recent attempts to mitigate publication bias, such corrective methods tend to rely on strongly assumptive statistical models (e.g., Guan & Vandekerckhove, 2016) or fail under a host of common conditions (Stanley, 2017). Unsurprisingly, these methods are applied infrequently in practice.

Interpretability

In addition, many meta-analyses suffer from the same issue seen in many statistical tests, namely a lack of interpretable parameters. In most meta-analyses, the obtained effect estimates are constructs of a purely statistical nature. This limits the potential theoretical impact of any results obtained. We also note that this particular limitation is not limited to frequentist meta-analyses, but also occurs in Bayesian meta-analyses. In most cases, statistical effects can only describe *what* changes occur in behavior, never *why* they are occurring. This particular distinction is crucially important in cognitive science.

4.2.1 Model-based Integrative Data Analysis

Although it is clearly a useful analytical tool, the previous section illustrates that meta-analysis, as commonly conducted, can suffer from both practical and theoretical issues. In the present paper, we attempt to address some of these issues by demonstrating a novel approach to performing a meta-analysis, which we refer to as a Model-based Integrative Data AnalysiS, or MIDAS.

MIDAS is based on the hierarchical Bayesian extension of cognitive models. While hierarchical Bayesian models date back to at least the 1950s (Good, 1952), their application to cognitive models is relatively recent (Lee, 2011; Rouder & Lu, 2005; Vandekerckhove et al., 2011). The model-based approach is specifically designed to allow for complex analyses on interpretable parameters, in order to provide psychologically meaningful joint conclusions regarding both task-related cognitive processes and their structured individual differences. Accordingly, MIDAS does not suffer from the lack of interpretable parameters that plague most off-the-shelf analyses, but instead provides conclusions that are psychologically meaningful. A more detailed account of the analytical benefits of such an approach can be found in Kupitz and Vandekerckhove (2019) and in Vandekerckhove et al. (2010).

MIDAS extends the hierarchical Bayesian model to also be an Integrative Data Analysis (IDA). This is done by pooling the trial level data of many studies, and analyzing all the data in a single encompassing model. In general, IDA approaches allow stronger statistical inference compared to non-IDA based analyses (Curran & Hussong, 2009), and the MIDAS is no exception. With this approach, MIDAS can be used to simultaneously capture the structured differences in the cognitive processes underlying behavior at the *individual, within experiment, and between experiment levels.* Critically, this structure allows MIDAS to perform a meta-analysis of the cognitive effects driving behavior.

4.2.2 Spatial congruency bias

In the present study, we revisit two papers that relate to the Spatial Congruency Bias (SCB) effect. The motivation of the first paper (Golomb, Kupitz, & Thiemann, 2014) was to thoroughly investigate the "binding problem." The binding problem refers to the fact that different object features (especially shape, color, and location) are coded separately in the brain, and must be integrated together in order to identify objects. In particular, this paper focused on how object-location binding occurred, over a total of six experiments.

In the first experiment, Golomb et al. (2014) found that object location influences the perception of object identity: when two objects were presented in the same location, there was a bias towards judging them as having the same identity. We will refer to this effect as the Spatial Congruency Bias or SCB. In their second experiment, Golomb et al. (2014) found that the SCB effect is somewhat gradient – the more similar the location of the two objects, the stronger the SCB effect. The third experiment found that the SCB effect also influences graded judgments of object similarity, not only same-or-different identity judgments. Together, Experiments 4 through 6 found that location uniquely influences perceptions of object identity, because the SCB effect held for color judgments, but neither color nor identity influenced judgments of location.

The second paper (Shafer-Skelton et al., 2017) focuses on how object-location binding occurs across a saccade (an eye-movement). In this study, the authors were primarily interested in determining whether the SCB effect would occur if objects were presented in the same retinotopic location (i.e., the same location on the retina), or the same spatiotopic location (i.e., the same location in the world). Over the experiments conducted within, it was concluded that the SCB effect is more prominent when the retinotopic locations are congruent. In one case, there was also a small SCB effect when the spatiotopic locations were congruent. Additionally, this study also demonstrated that the SCB effect is also found when judging face stimuli, and when judging the orientation of Gabor patches.

In the majority of these experiments, a significant reaction time (RT) priming effect was found: when objects were presented in the same location, overall RTs were reduced, a finding which agrees with many previous investigations.

The SCB effect is reasonably robust, having been observed in a number of experimental settings. Prominent examples include the addition of a saccade between stimuli, alteration of stimulus type, and altering the stimulus presentation timing. The majority of studies on the SCB use a 2-Alternative Forced Choice task, which has established cognitive models that capture participant behavior well. Altogether, we believe this makes the SCB an excellent candidate effect to illustrate the MIDAS approach to meta-analysis.

4.3 Data set

4.3.1 Task Description

All of the studies included in our analysis use a two-alternative forced choice task. Each trial followed the same overall structure, as illustrated in Figure 4.1. First, the participant focuses on a fixation point. Second, a stimulus is presented peripherally, and must be remembered. Third, the participant focuses on a second fixation point (in some experiments, there is a saccade between fixation points; in others, the fixation points are identical). Fourth, a second stimulus is presented peripherally. Finally, the participant makes a judgment: were the stimuli the same or different?

Here we revisit data from a total of 15 experiments investigating the SCB effect, with a total of $N = 256$ subjects remaining after data preparation and sanitization.[1] The experiments included herein have a number of variations in trial timing and stimuli, but the general structure of the trial is preserved in all included experiments. In particular, many of these manipulations are shared across multiple experiments. Table 4.1 contains the *integrative design matrix*, which shows the manipulations that are common between experiments and that we will be investigating.

We now describe what each row indicates in the integrative design matrix. *No-Saccade* indicates the inclusion of trials where no saccade was performed between the two stimulus presentations. *Saccade* indicates the inclusion of trials where a saccade was performed in between the two stimulus presentations. *Spatiotopic* indicates the inclusion of trials where the second stimulus shared the same location in space as the first stimulus, which we refer

[1] Our inclusion criterion were chosen *a priori*, and are as follows. First, we censored any trials wherein the reaction time was clearly not from a one-shot decision making process, in this case by requiring reaction times to be between 200ms and 2000ms. Following the removal of any trials with unacceptable reaction times, we then removed any participants who dropped below 120 total trials completed (as this is 75% of the total number of trials in the smallest experiment, which had 160 trials). This restriction resulted in the removal of six subjects from the data pool, leaving us with a total of $N = 256$.

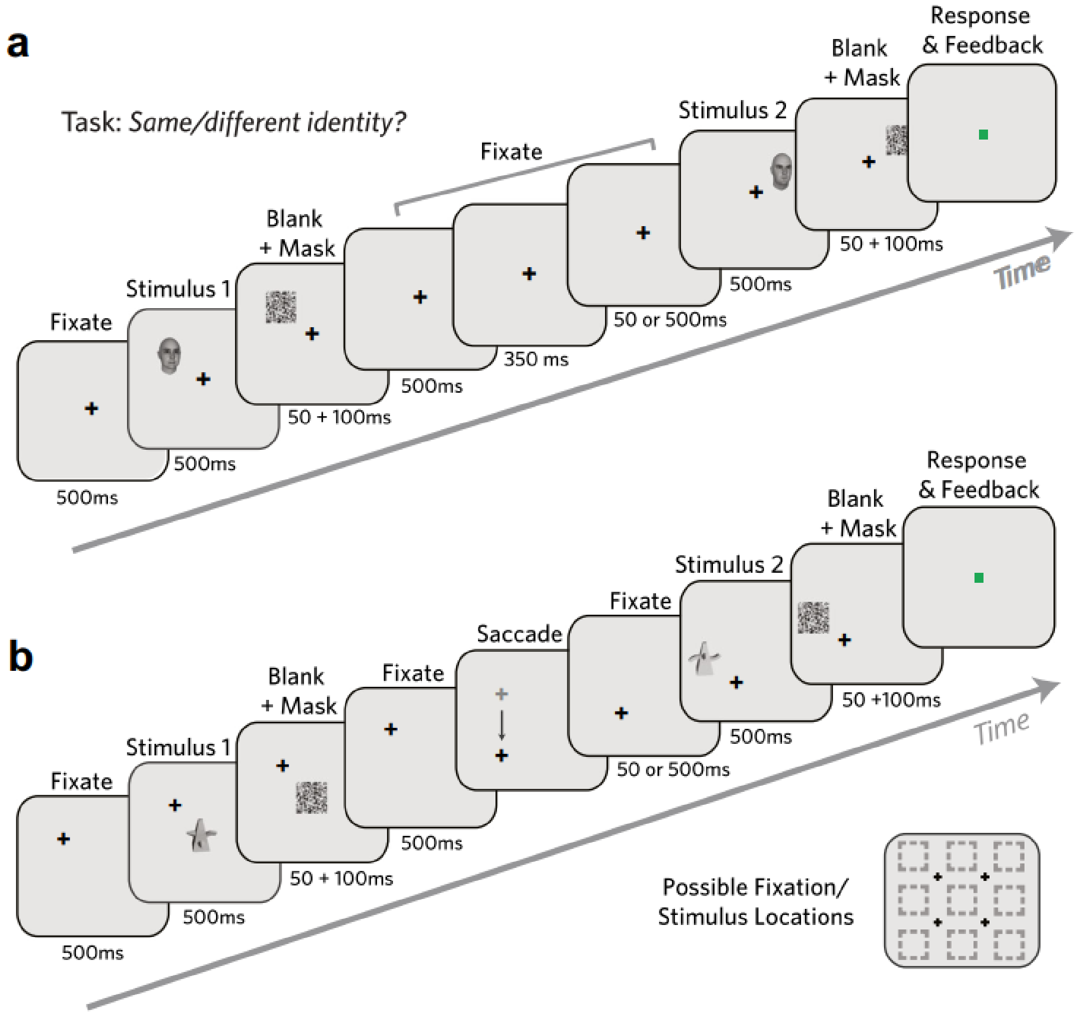

Figure 4.1: (a) An example trial in an SCB task using Faces as stimuli, without a saccade included between stimuli. Note that both the identity and location of stimulus 2 could be different from that of stimulus 1. (b) A second SCB example trial, using Morph stimuli, and including a saccade between stimulus presentations. Figure taken from Shafer-Skelton et al. (2017), used with permission.

to as spatiotopic congruency. *Retinotopic* indicates the inclusion of trials where the second stimulus shared the same retinal location as the first stimulus, which we refer to as retinotopic congruency. Note that in the no-saccade task, stimuli presented in the same location would have both retinotopic and spatiotopic congruency. *Probe 50ms* and *Probe 500ms* indicate the inclusion of trials where the second fixation time was 50ms or 500ms, respectively. *Morph, Color, Face,* and *Gabor Stimuli* indicate the type of stimuli compared in the "same" or "different" identity judgments; examples of each class of stimulus are provided in figure 4.2.

In saccade and no-saccade tasks, there was always a 500ms delay following the conclusion of the first masking stimulus. Following this delay, the no-saccade task immediately enters the second fixation stage (for the length of time indicated by Probe). In the saccade task, the participant was queued to perform a saccade following the 500ms delay, and the the second fixation stage began when the participant had successfully completed the saccade to the new point of fixation.

4.4 Goals of the MIDAS

4.4.1 Improved Psychological Interpretability

The first goal of the MIDAS is to offer results with improved psychological interpretability. In the original studies, bias measures were obtained using signal detection theory (SDT). While some may argue that SDT is a form of cognitive theory, it is not a cognitive *process* model. These SDT measures do not attempt to distinguish the cognitive processes driving each individual decision. Instead, the measures provided by SDT are summary statistics: they describe the relative proportions of hit rates and false alarm rates over the entirety of a participants data.

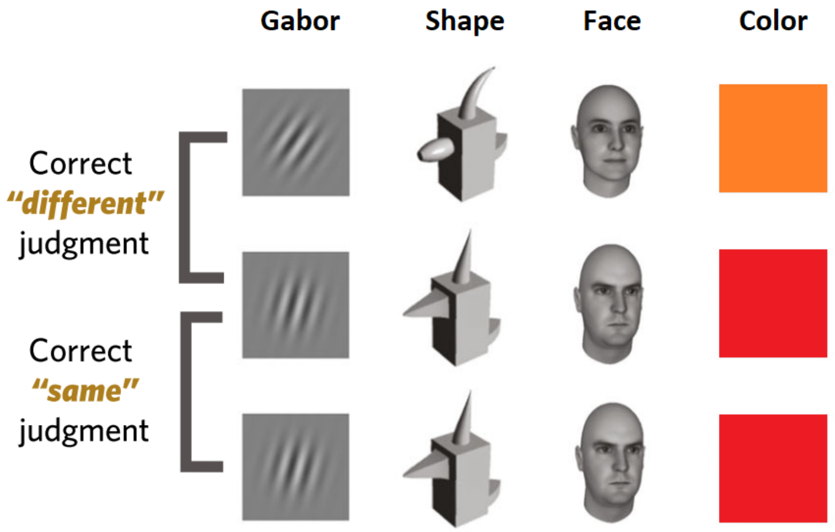

Figure 4.2: Examples of the Gabor, Shape, Face, and Color stimuli used in the SCB experiments being analyzed here. Figure taken from Shafer-Skelton et al. (2017), used with permission, and modified to include Color stimuli.

Table 4.1: Design matrix of the experimental manipulations included in each of the 15 experiments we analyze, where ✓ indicates inclusion.

Experiment ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No-Saccade	✓	✓	·	·	·	✓	·	✓	·	·	✓	✓	✓	✓	·
Saccade	✓	✓	✓	✓	✓	·	✓	·	✓	✓	·	·	·	·	✓
Spatiotopic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Retinotopic	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Probe 50ms	✓	·	✓	✓	✓	·	✓	✓	✓	✓	·	·	·	·	·
Probe 500ms	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Shape Stimuli	✓	✓	✓	·	·	·	·	·	·	·	·	·	✓	✓	✓
Color Stimuli	·	·	·	✓	✓	·	·	·	·	·	✓	✓	·	·	·
Face Stimuli	·	·	·	·	·	✓	✓	·	·	·	·	·	·	·	·
Gabor Stimuli	·	·	·	·	·	·	·	✓	✓	✓	·	·	·	·	·

Thus, we believe the interpretability of the results can be improved with use of a cognitive process model. By using a cognitive process model – in this case, the diffusion model – we will obtain estimates of bias *within the cognitive processes driving decisions.*

Further, our estimates of bias will be driven by both response and reaction time, simultaneously.The original studies analyzed the reaction time data separately, and again, using simple summary statistics. The simultaneous inclusion of reaction time data above and beyond the accuracy data provides for a much richer data set from which we can draw stronger conclusion merely by virtue of the increased size of the data set.

4.4.2 Analytical Goals

Our primary analytical goal is to investigate the underlying cognitive effects that drive the spatial congruency bias. To that end, we will construct a model that simultaneously considers effects at the individual level, within-experiment level, and between-experiment level.

Our secondary analytical goal is to improve the statistical validity of the results. The results of the original experiments were obtained with the analysis of simple summary statistics. Here, our analysis is based on a Bayesian hierarchical model. With this approach, we avoid summarizing any of the data. Instead, the trial level data directly informs our ultimate effect estimates through a hierarchy. This approach ensures that we do not lose any of the variability inherent in the data, as it propagates through the model in the form of uncertainty.

4.5 Constructing the MIDAS

Cognitive Model Foundation

The experimental data we analyze herein comes in the form of two-choice reaction times. Accordingly, we use an hierarchical extension of the diffusion model first introduced by Vandekerckhove et al. (2011; see also Vandekerckhove et al., 2008) to serve as the cognitive model foundation upon which we construct the full MIDAS.

The key assumption of the diffusion model is that participants gather information sequentially, executing a response only when enough information has been gathered; the process is illustrated in Figure 4.3. There are four parameters of interest, each of which captures a unique portion of the cognitive processes underlying behavior. The first is α, the total amount of information required to make a decision, which typically captures cognitive strategy shifts, such as a speed-accuracy tradeoff. Next is β, the a-priori bias favoring one of the choices over the other. Then τ represents the non-decision time, including both stimulus encoding and motor response execution time (Nunez et al., 2015, 2017). Finally, δ is the *drift rate* or rate of information accumulation within a trial. In essence, we use this parameterization in order to transform participant behavioral data into psychologically meaningful constructs, upon which we can perform more advanced analyses on these psychologically meaningful parameters.

Choice of Bias Measure

In the diffusion model, there are two potential sources of response bias. The first is simply β, the *a priori* bias towards one response over another. The second is often referred to as drift criterion, and is the difference in mean drift rates between some experiment conditions (Leite & Ratcliff, 2010; Ratcliff, 1985). In essence, drift criterion is the bias in the evidence accumulation process itself. Here, we are interested in determining how the second stimulus location influences judgments of object identity. Accordingly, the drift criterion measure is a more appropriate choice of bias measure. We also note that drift criterion is analogous to the bias measure obtained using SDT and is thus especially apropos here.

Parameter Restrictions

From a theoretical standpoint, we would not expect the alteration of stimulus location within a particular trial to change β, as it is meant to capture the response bias *that exists before the decision begins.* Similarly, we believe that α, the amount of evidence required for a response, should not be influenced by stimulus location. Accordingly, within our models we restrict both α and β to only vary by experiment, participant, and task (saccade or no-saccade).

Notational Conventions

We will use θ to represent any of the four cognitive parameters, α, β, τ, and δ. When appropriate, subscripts will index the level of any of the independent variables in which the parameter may be nested.

4.6 Model-Based Integrative Data Analysis

In our first model, we focus on analyzing how object location influences judgments of identity. For all four of the WDM parameters, α, β, τ, and δ, we estimate the following effects. For readability, we do not index by these parameters unless there are differences. Separately for the saccade $s = 1$ and no-saccade $s = 0$ tasks, we assume a fixed effect μ_s to serve as the task baseline. This baseline parameter is the only structural parameter in which we treat the four WDM parameters differently, and use prior values chosen by researcher experience. We also assume a fixed effect ζ_c for each stimulus class, c, (Shape, Color, Face, Gabor, 1 through 4 respectively). For model identifiability, these ζ_c are defined relative to the Shape ($c = 1$) stimulus class (so $\zeta_{c=1} = 0$). We also assume a fixed offset effect for each experiment e and saccade s combination, η_{es}, which is drawn from a normal distribution centered on 0. We also assume a fixed effect of each participant p, ρ_p, in order to account for individual differences, drawn from a normal distribution centered on 0.

$$\begin{aligned} \mu_s^{\alpha} &\sim N(1.5, 1) & \mu_s^{\beta} &\sim N(0.5, 1) \\ \mu_s^{\tau} &\sim N(0.2, 1) & \mu_s^{\delta} &\sim N(0, 1) \\ \zeta_c &\sim N(0, 0.2) & \zeta_{c=1} &= 0 \\ \eta_{es} &\sim N(0, \sigma_\eta) & \sigma_\eta &\sim Gamma\,(0.5, 0.5) \\ \rho_p &\sim N(0, \sigma_\rho) & \sigma_\rho &\sim Gamma\,(1, 1) \end{aligned}$$

For the following parameters, we only estimate effects for the non-decision time, τ, and the drift rate, δ. Our rationales for this choice are described in the Parameter Restrictions section above. In each experiment e we assume a fixed effect ϕ_e, which captures the effect of reducing the second fixation time f from 500ms to 50ms, and is drawn from a normal distribution centered on the latent effect ϕ.

In the no-saccade task, the second stimulus can either be presented in a different location, or the same location. Accordingly, we treat the different location condition as the baseline, and allow for a fixed effect of spatial congruency $\lambda_{1,e}$ in each experiment e, which are drawn from a latent mean Λ_1.

In the saccade task, the second stimulus can presented in one of three locations of interest: different location, same spatiotopic location, or same retinotopic location. As in the no-saccade task, we treat the different location condition as the baseline, and allow for a fixed effect of spatiotopic congruency, $\lambda_{2,e}$, and a fixed effect of retinotopic congruency, $\lambda_{3,e}$, in each experiment e. Again, these are drawn from latent means Λ_2 and Λ_3 respectively.

$$\begin{array}{lll} \phi_e \sim N(\phi, \sigma_\phi) & \phi \sim N(0, 0.2) & \sigma_\phi \sim \mathit{Gamma}\,(0.5, 0.5) \\ \lambda_{el} \sim N(\Lambda_l, \sigma_{\lambda_l}) & \Lambda_l \sim N(0, 0.2) & \sigma_{\lambda,l} \sim \mathit{Gamma}\,(1, 1) \end{array}$$

For τ only, we assume a random effect ε_{seclfp} for each unique combination of effects, drawn from a normal distribution centered on 0. We include this as a way to capture measurement error in the non-decision time. The estimates for each τ parameter are strongly anchored to the fastest response present in a given condition, and here we do not have the level of data per condition requisite to ensure that the real non-decision time has been captured. For δ, the non-location effects are positive for “same” identity trials, and negative for “different” identity trials, so that these effects represent the rate of evidence accumulation towards the appropriate boundary. This ensures that the location effects, λ^{δ}_{el}, are drift criterion effects, our estimate of bias. Altogether, these effects are combined to provide the final WDM parameter estimates

$$
\begin{aligned}
\varepsilon_{seclfp} &\sim N(0, \sigma_\varepsilon) \qquad \sigma_\varepsilon \sim Gamma\,(1, 1) \\
\alpha_{secp} &= \mu_s^\alpha + \zeta_c^\alpha + \eta_{es}^\alpha + \rho_p^\alpha \\
\beta_{secp} &= \mu_s^\beta + \zeta_c^\beta + \eta_{es}^\beta + \rho_p^\beta \\
\tau_{seclfp} &= \mu_s^\tau + \zeta_c^\tau + \eta_{es}^\tau + \rho_p^\tau + \phi_e^\tau + \lambda_{el}^\tau + \varepsilon_{seclfp} \\
\delta_{seclfp} &= (\mu_s^\delta + \zeta_c^\delta + \eta_{es}^\delta + \rho_p^\delta + \phi_e^\delta) + \lambda_{el}^\delta \quad \text{for ``same'' identity} \\
\delta_{seclfp} &= -(\mu_s^\delta + \zeta_c^\delta + \eta_{es}^\delta + \rho_p^\delta + \phi_e^\delta) + \lambda_{el}^\delta \quad \text{for ``different'' identity}
\end{aligned}
$$

which are used in the likelihood function $\mathbf{y}_{seclfp} \sim W(\alpha_{secp}, \beta_{secp}, \tau_{seclfp}, \delta_{seclfp})$. The likelihood function is defined as the first passage time distribution of a Wiener process with constant boundaries (Feller, 1970; Navarro & Fuss, 2009). Note that while we forego indicators or additional indexing for readability, ϕ is set to zero when it is not a fast probe trial, and λ is set to zero when it is a different location trial.

4.7 MIDAS Results

Primarily, we are interested in determining the presence or absence of the various location congruency effects, and if present, their direction. To that end, we will be presenting the various location effect estimates across the experiment pool, λ_{el}, but primarily focus on the estimated latent effect across experiments, Λ_l, which serves as the latent mean from which the experiment specific effects are estimated. For each of these Λ_l we intentionally use an uninformative prior centered on zero, to allow the data to maximally drive our conclusions. Given such a prior distribution, if the posterior distribution is entirely positive, we will strongly conclude that there is a true effect in the positive direction, as it implies the data is providing absolutely no evidence that the effect is zero or negative. The same holds in reverse for the negative direction. Finally, if the posterior distribution straddles zero, then we may conclude that there is no effect, that it is of negligible magnitude, or that the data do not allow strong conclusions either way.[2]

No-Saccade, Spatial Congruency Effect

First we present the estimated spatial congruency effects for the No-Saccade task in figure 4.4. For τ, the general trend of λ_{e1}^{τ} across experiments is negative, suggesting a decrease in non-decision time. For δ, the general trend of λ_{e1}^{δ} across experiments is positive, suggesting the presence of a drift criterion effect.

In both τ and δ, we observe a degree of between-experiment heterogeneity in the λ_{e1} effect estimates. In fact, a number of these λ_{e1}^{θ} estimates are not conclusive effects – meaning they include 0 as a possible effect size in their 95% credibility intervals – in essence implying we would conclude the absence of an effect, if we were to analyze that experiment independently. However, one of the driving motivations behind any meta-analyses is to account for between-experiment heterogeneity, and in doing so increase the certainty in the final effect estimate. In MIDAS this comes in the form of our hierarchical effect estimates: we estimate the latent effect, Λ_1^{θ}, and assume the experiment specific effects, λ_{e1}^{θ}, will follow a normal distribution centered on Λ_1^{θ}, with variance $\sigma_{l=1}^{\lambda,\theta}$.

These estimated latent spatial congruency effects, Λ_1^{θ}, are presented in blue on figure 4.4. We observe that Λ_3^{τ} is completely negative, indicating that spatial congruency reduces non-decision time, which in this case is likely due to a reduction in processing time. We also

[2] This approach is suboptimal. In general, we prefer the use of Bayes factor model comparison for the purpose of deciding whether an effect is positive, negative, or zero. To that end, we implemented a version of "does-everyone" testing as proposed by Haaf and Rouder (2017). However, the procedure turned out to be numerically unstable, so we omit the results here. We default to estimating the treatment effect in the style of Gelman (2004).

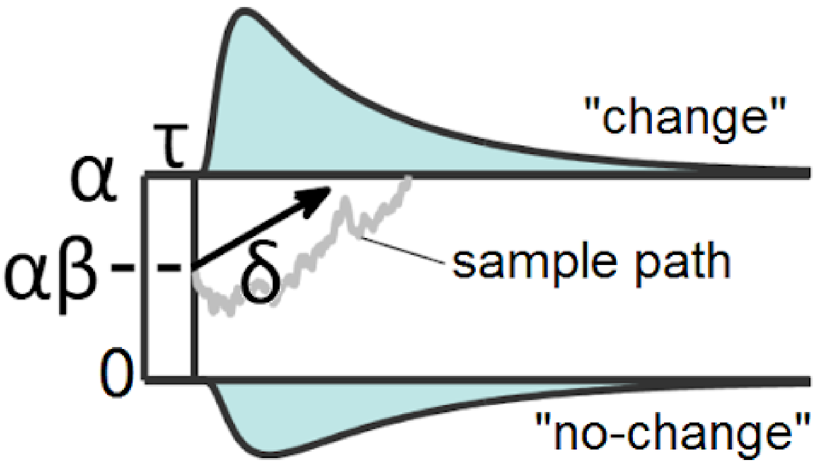

Figure 4.3: A graphical representation of the Wiener diffusion model. Evidence is accumulated at an average rate of δ units per second, and begins at the value $\alpha\beta$. The process stops when evidence reaches the boundary at α or 0, and then a response is made. The additive time constant τ accounts for nondecisional processes. The shaded area is the model-predicted probability density function over response and response time, $W(\alpha, \beta, \tau, \delta)$. Figure taken from Vandekerckhove (2009), used with permission.

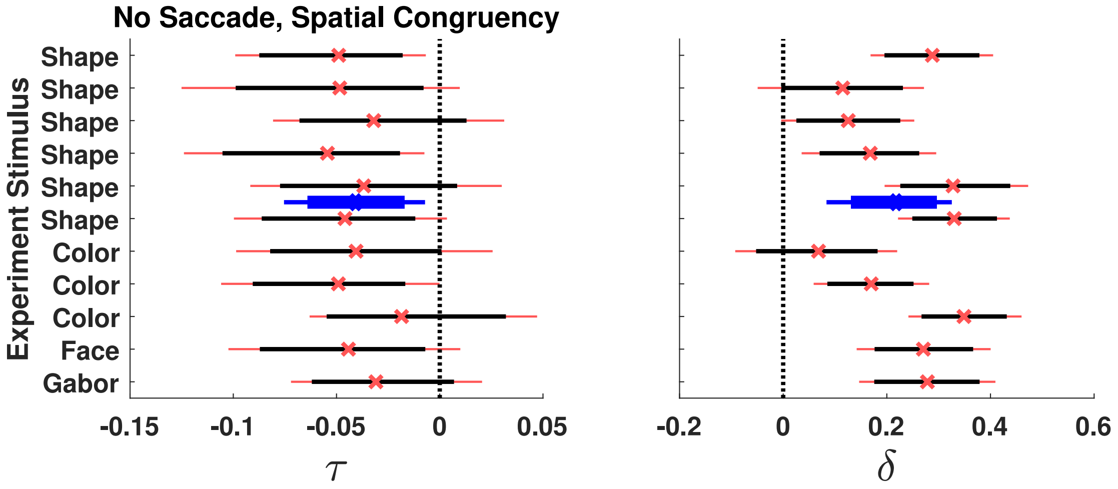

Figure 4.4: In the No Saccade task, the estimated effect λ_1 for each experiment, which represents the effect of complete spatial congruency. The estimated effects for each experiment are presented as black and red bars, which are the 95% and 99% credibility intervals, respectively. The estimated latent effect, Λ_1, is shown in blue (thick and thin bars 95% and 99% credibility intervals, respectively). Λ_1 is the latent mean from which each λ_{1e} is drawn.

observe that Λ_1^δ is completely positive, and thus find that spatial congruency has a positive drift criterion effect, and therefore induces a bias in the decision making process towards responding 'same.' Notably, both of these latent effect estimates are completely unambiguous, as they give no posterior weight to the possibility of zero as an effect size. Furthermore, they also show a reduced amount of uncertainty relative to the experiment specific effects. These observations demonstrate how the MIDAS approach can provide strong conclusions about cognitive effects.

Saccade, Spatiotopic Congruency Effect

The estimated spatiotopic congruency effects for the Saccade task are presented in figure 4.5. As before, there is a degree of between-experiment heterogeneity in the experiment specific λ_{e2}. Despite this, all of the experiment specific effects λ_{e2} and both latent effect estimates Λ_2 include 0 in their posteriors. Accordingly, we find that spatiotopic congruency has no meaningful effect on judgments of object identity.

Saccade, Retinotopic Congruency Effect

The estimated retinotopic congruency effects for the Saccade task are presented in figure 4.6. As expected, there is a degree of between-experiment heterogeneity in the experiment specific λ_{e3}. We find that there is no effect of retinotopic congruency on non-decision time, as the estimated latent effect Λ_2^τ includes and has a mean close to 0. We find that the estimated latent effect, Λ_3^δ is completely positive, and therefore conclude that retinotopic congruency induces a bias in the cognitive process towards responding 'same.'

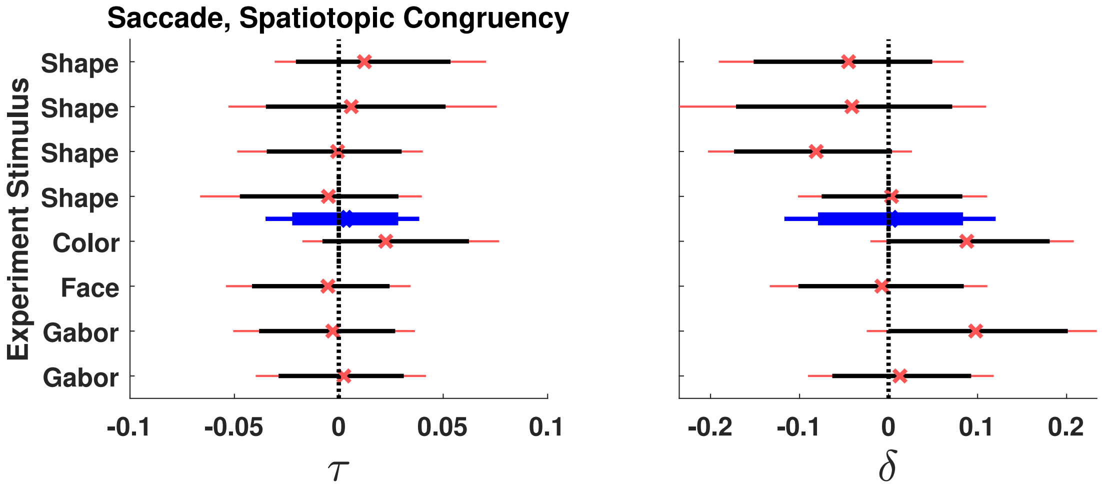

Figure 4.5: In the Saccade task, the estimated effect λ_2 for each experiment, which represents the effect of spatiotopic congruency following a saccade. The estimated effects for each experiment are presented as black and red bars, which are the 95% and 99% credibility intervals, respectively. The estimated latent effect, Λ_2, is shown in blue (thick and thin bars 95% and 99% credibility intervals, respectively). Λ_2 is the latent mean from which each λ_{e2} is drawn.

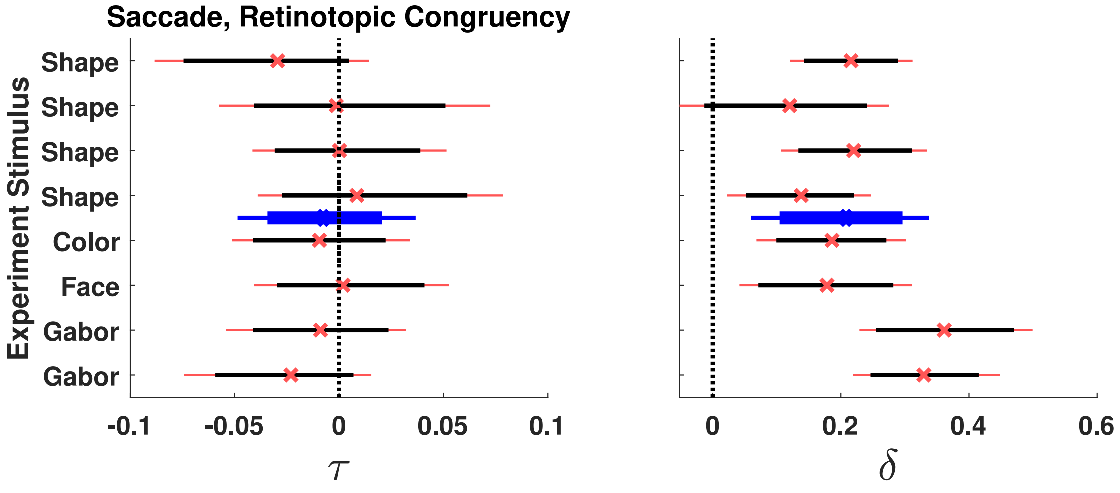

Figure 4.6: In the Saccade task, the estimated effect λ_7 for each experiment, which represents the effect of retinotopic congruency following a saccade. The estimated effects for each experiment are presented as black and red bars, which are the 95% and 99% credibility intervals, respectively. The estimated latent effect, Λ_3, is shown in blue (thick and thin bars 95% and 99% credibility intervals, respectively). Λ_3 is the latent mean from which each λ_{3e} is drawn.

4.8 Meta-analytical Conclusions of MIDAS

With the MIDAS, we have validated the original SCB effect findings in a meta-analytical setting. We estimated the effect of spatial congruency across all studies, and found evidence of a cognitive effect on judgments of object identity. We find that presenting objects in the same location creates a positive drift criterion effect, thereby inducing a bias towards judging those objects as identical. Similarly, we also observed that spatial congruency reduces the non-decision time involved in those judgments, likely due to a reduction in encoding time, which again validates the RT priming effects observed in the original study.

We have also analyzed how object location influences identity judgments over a saccade. In the original studies, it was concluded that following a saccade, or eye movement, the SCB is primarily driven by retinotopic congruency. Here, the MIDAS finds that retinotopic congruency has a cognitive effect, a positive shift in drift criterion that induces a bias towards judging objects as the same. Once again, we have validated the findings of the original studies in a meta-analytical setting.

4.9 Model Diagnostics: Posterior Predictives

We evaluate model fit through posterior predictive assessment (Gelman et al., 1996). Although we include measures of individual differences, the ultimate goal of the MIDAS is to accurately capture behavior at the experiment level. Accordingly, our diagnostic visualizations focus on experiment level posterior predictive distributions. For each experiment, we provide the posterior predictive distribution for the median RT, mean RT, and accuracy in figure 4.7. The lack of direct coverage in the median RT and mean RT measures indicate it is likely there are additional effects that the model is not accounting for at present. However, in general the MIDAS captures the pattern of differences in the measures very well, as in all three measures the predicted distributions correlate almost perfectly with observed values. As the MIDAS is an attempt to model experimental differences in a parsimious manner, we believe this indicates an acceptable degree of model fit.

4.10 Discussion

In this study, we have developed a novel approach to performing meta-analyses in the behavioral sciences, which we call the Model-based Integrative Data AnalysiS, or MIDAS. To serve as an example of the benefits afforded by MIDAS, we have reanalyzed a number of experiments that investigated the Spatial Congruency Bias effect.

In our first MIDAS model, we validated the results of Golomb et al. (2014), which found that spatial congruency results in both reduced reaction times, and a bias towards judging objects as identical (the SCB effect). In a meta-analytical setting, we found strong evidence that spatial congruency both produces a reaction time priming effect, and further, induces a true bias in the cognitive process towards judging two stimuli as being the same identity – regardless of if they actually were the same object.

Similarly, we also validated the results of Shafer-Skelton et al. (2017), which found that after a saccade, retinotopic congruency induced a bias towards judging objects as identical. In a meta-analytical setting, we found strong evidence that retinotopic congruency induces a true bias in the cognitive process towards judging two stimuli as being the same identity. However, we further found strong evidence that spatiotopic congruency induces the same bias, but to a smaller magnitude. This finding is in contrast to the original study by Shafer-Skelton et al. (2017), wherein they found that spatiotopic congruency had a significant bias effect in only case. This finding in particular illustrates one of the benefits of the MIDAS approach: it is more capable of discerning small but true effects than standard analytical

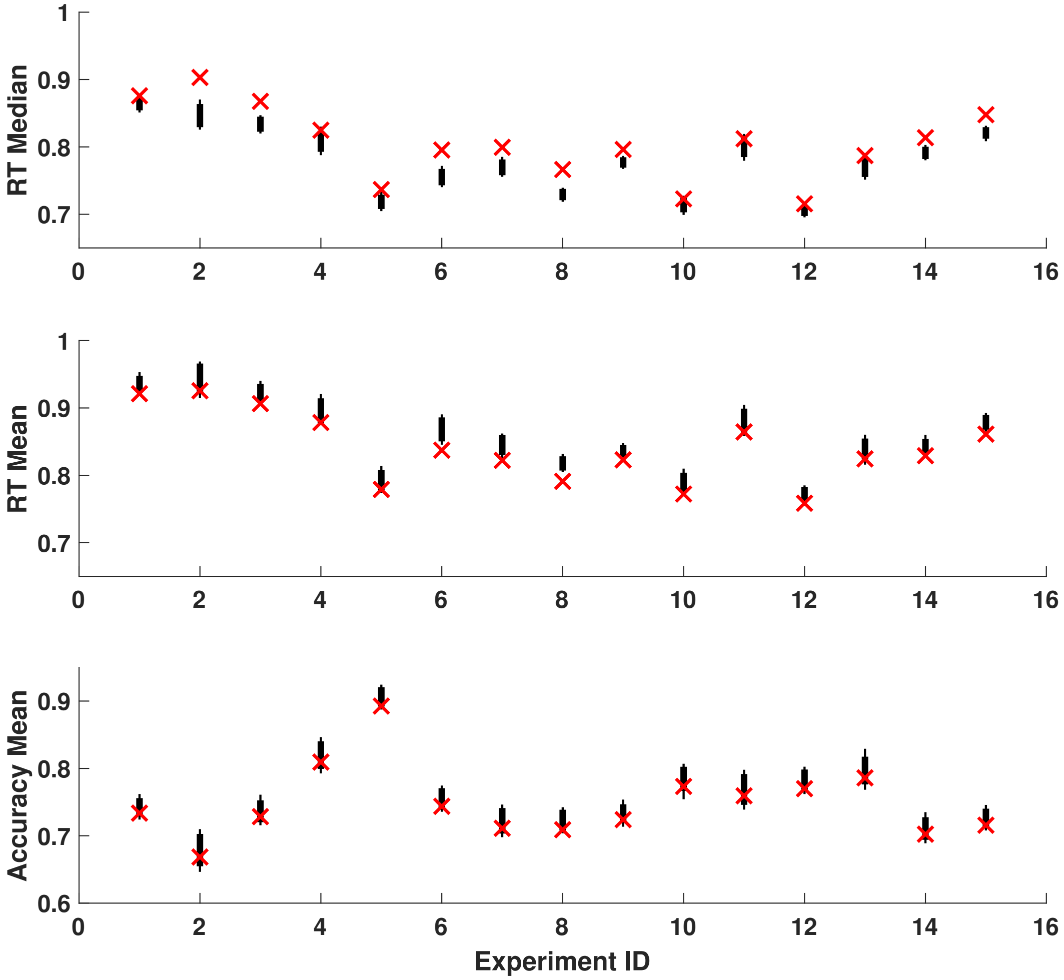

Figure 4.7: Posterior predictives for RT median, RT mean, and accuracy, across all included experiments. Thick and thin black bars indicate the 95% and 99% credibility intervals,

approaches.

Importantly, these conclusions were made possible by the unique approach of MIDAS. With MIDAS, within one hierarchical model we are able to simultaneously account for the structured differences in effects at the individual, within-experiment, and between-experiment levels. In addition, we implement this hierarchy in a Bayesian framework, and thus are able to fully account for the uncertainty in our estimates at every level of data. Together, this provides estimated distributions of the true effect with unparalleled statistical rigor. The implementation of this model was made possible by pooling all the trial level data across experiments, illustrating how the MIDAS benefits from being a form of Integrative Data Analysis.

Finally, the MIDAS is further improved by the cognitive model foundation we utilize. Primarily, this comes in the form of improved interpretability, as we are able to discuss the estimated effects of interest in terms of their psychological meaning – how they alter the cognitive processes underlying behavior. In the original studies, signal detection theory was used to obtain bias estimates for each individual. Here, the MIDAS provides estimates of the bias in the decision making process, using *both* response patterns and reaction times simultaneously. Although neither the SDT nor WDM can distinguish between perceptual and response level biases, we still believe that the bias effects provided by the WDM are more meaningful, as they are provided in the context of a generative cognitive process model, and attempt to account for the full participant behavior.

4.11 Conclusion

In this study, we have described a novel approach to performing meta-analyses in the behavioral sciences, the Model-based Integrative Data AnalysiS, or MIDAS. Critically, the MIDAS provides the ability to draw meta-analytical conclusions about cognitive effects.

4.12 Appendix

4.12.1 Additional MIDAS Parameter Estimates

The effect estimates provided here were primarily included ensure that we accurately accounted for the general design differences between the included experiments. The estimated effects of reducing Probe from 500ms to 50ms, ϕ_e, are presented in figure 4.8. Generally, this reduction in probe time produces an increase in non-decision time, although there is a large degree of between experiment heterogenity in this effect.

The estimated effect of altering stimulus type from Shape to either Color, Face, or Gabor, ζ_c, are presented in figure 4.9.

4.13 Authors' Note

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. This project was partly supported by grants #1230118 from NSF's Methods, Measurements, and Statistics panel and #48192 from the John Templeton Foundation to Joachim Vandekerckhove.

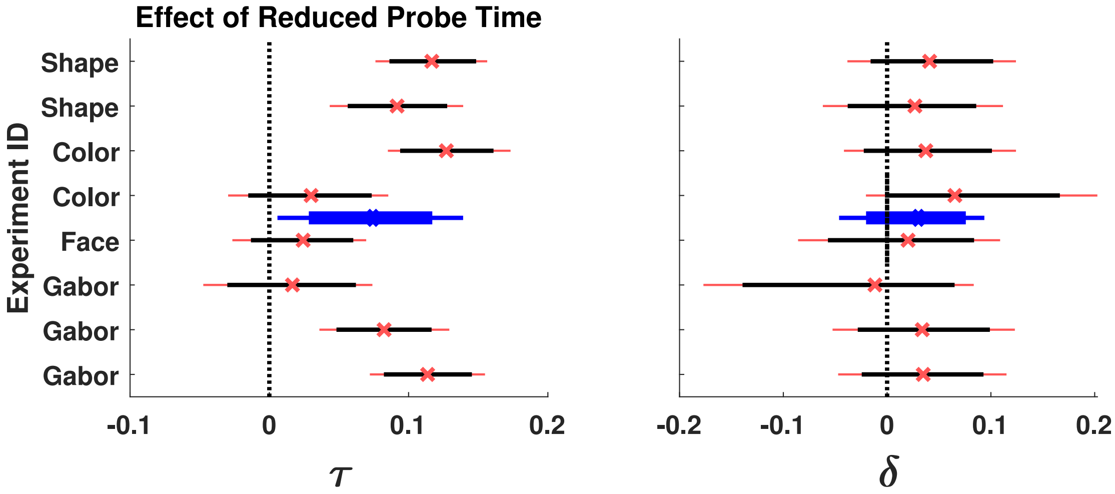

Figure 4.8: Estimated effect ϕ_e for each experiment, which represents the effect of altering the probe duration from 500ms to 50ms. The estimated effects for each experiment are presented as black and red bars, which are the 95% and 99% credibility intervals, respectively. The estimated true effect, Φ, is shown in blue (thick and thin bars 95% and 99% credibility intervals, respectively).

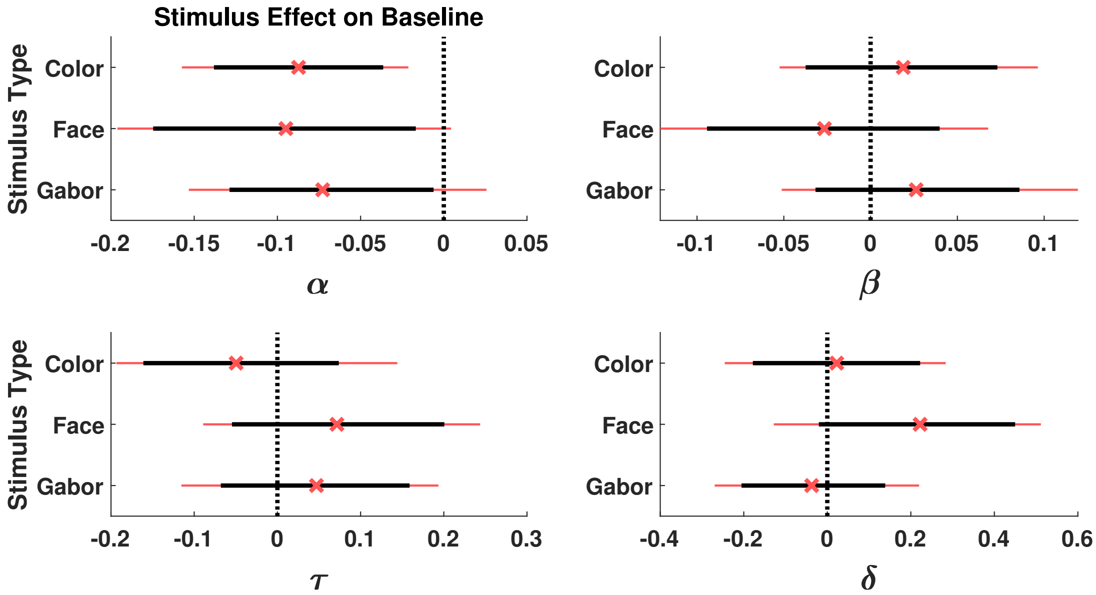

Figure 4.9: Estimated effect ζ_c for each experiment, which represents the effect of altering the stimulus from Shape to the indicated type. The estimated effects for each experiment are presented as black and red bars, which are the 95% and 99% credibility intervals, respectively.

Chapter 5

Conclusion

In this dissertation, we have demonstrated that hierarchical Bayesian cognitive modeling is an extremely flexible analytical framework, and further, that it commonly affords a depth of inference that cannot be matched by more standard analytical methods.

In Chapter 2, we offer a demonstration of a basic hierarchical cognitive model (HCM), by revisiting a previous study on working memory training and transfer. Using the Wiener Diffusion Model (WDM) as a foundation, we extend it hierarchically in a Bayesian framework, and in doing so capture *cognitive* training and transfer effects at the population level.

In Chapter 3, we first improved upon the aforementioned HCM, by creating a more complex HCM to capture the training and transfer effects. We then further extended this analysis to create a Cognitive Latent Variable Model (CLVM), which provided us with the capability to drawn inferences regarding the cognitive training and transfer effects of both the population, and individual participants, simultaneously. We then compare these CLVM inferences to those obtained using less comprehensive analytical approaches, in essence stripping out the various major pieces of the CLVM approach. By doing this, we provided a concrete illustration of how each piece of the CLVM analytical approach contributes to the overall strength of analysis, and critically, we demonstrate why the CLVM approach provides inferences that cannot be obtained with more traditional analyses.

In Chapter 4, we demonstrated the very flexible nature of the CLVM analytical approach. In particular, we brought the CLVM into the realm of Integrative Data Analysis by pooling all of the trial-level data from a number of separate experiments investigating the same psychological effect. In doing so, we created a novel approach to meta analyses in cognitive science: the Model-based Integrative Data AnalysiS, or MIDAS. We then provided a proof-of-concept for the MIDAS using a number of experiments from the spatial congruency bias literature. The MIDAS approach afforded us the capability to draw meta analytical conclusions regarding the cognitive effects driving behavior across multiple experiments.

References

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Buschkuehl, M., Jaeggi, S. M., Mueller, S. T., Shah, P., & Jonides, J. (2017). Training change detection leads to substantial task-specific improvement. *Journal of Cognitive Enhancement*, *1*(4), 419–433.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review*, *17*(2), 193–199.

Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*(6), 531–542.

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.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, *14*(2), 81–100. Retrieved from `https://doi.org/10.1037/a0015914` doi: 10.1037/a0015914

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450–466.

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2015). Reevaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin & Review*, 1–11.

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

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of experimental psychology: General*, *128*(3), 309.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*, e0149794. doi: 10.1371/journal.pone.0149794

Etz, A., & Vandekerckhove, J. (2018). Introduction to bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*(1), 5–34.

Feller, W. (1970). *An introduction to probability theory and its applications: Vol. I.* New York: John Wiley & Sons.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, *99*, 537–545.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

Golomb, J. D., Kupitz, C. N., & Thiemann, C. T. (2014). The influence of object location on identity: A "spatial congruency bias.". *Journal of Experimental Psychology: General*, *143*(6), 2262.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*(1), 107–114. Retrieved from http://www.jstor.org/stable/2984087

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic bulletin & review*, *23*(1), 74–86.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological methods*, *22*(4), 779.

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.

Hedges, L. V. (1992, December). Meta-analysis. *Journal of Educational Statistics*, *17*(4), 279–296. Retrieved from https://doi.org/10.3102/10769986017004279 doi: 10.3102/10769986017004279

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

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., & Gutbrod, K. (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.

Jones, R. (2007, April). Strength of evidence in qualitative research. *Journal of Clinical Epidemiology*, *60*(4), 321–323. Retrieved from https://doi.org/10.1016/j.jclinepi.2006.06.001 doi: 10.1016/j.jclinepi.2006.06.001

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., … Westerberg, H. (2005). Computerized training of working memory in children with adhd-a randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*(2), 177–186.

Kupitz, C., Buschkuehl, M., Jaeggi, S., Jonides, J., Shah, P., & Vandekerckhove, J. (2015). A diffusion model account of the transfer-of-training effect. In R. Dale et al. (Eds.), *Proceedings of the 37th annual conference of the cognitive science society* (pp. 1219–1224). Austin, TX: Cognitive Science Society.

Kupitz, C., & Vandekerckhove, J. (2019). *A cognitive latent variable approach to the transfer-of-training effect.* (Manuscript in preparation)

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge university press.

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, *72*, 246–273.

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.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*(4), 222–230.

Nunez, M. D., Srinivasan, R., & Vandekerckhove, J. (2015). Individual differences in attention influence perceptual decision making. *Frontiers in psychology*, *8*. Retrieved from `http://www.frontiersin.org/quantitative_psychology_and_measurement/10.3389/fpsyg.2015.00018/abstract`

Nunez, M. D., Vandekerckhove, J., & Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. *Journal of Mathematical Psychology*, *76*(Part B), 117–130.

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. doi: http://dx.doi.org/10.1016/S0160--2896(02)00115--0

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. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, *92*, 212–225.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532.

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.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., ... Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359.

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.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223.

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.

Shafer-Skelton, A., Kupitz, C. N., & Golomb, J. D. (2017). Object-location binding across a saccade: A retinotopic spatial congruency bias. *Attention, Perception, & Psychophysics*, *79*(3), 765–781.

Stanley, T. D. (2017, February). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science*, *8*(5), 581–591. Retrieved from `https://doi.org/10.1177/1948550617693062` doi: 10.1177/1948550617693062

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

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

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*(4), 277-303.

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., ... others (2013). Failure of working memory training to enhance cognition or intelligence. *PloS one*, *8*(5), e63614.

Tuerlinckx, F. (2004). The efficient computation of the cumulative distribution and probability density functions in the diffusion model. *Behavior Research Methods, Instruments, & Computers*, *36*, 702–716.

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. (2009). *Extensions and applications of the diffusion model for two-choice response times.* (Unpublished doctoral dissertation)

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). *Bayesian methods for advancing psychological science.* Springer.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision–making. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 1429–1434). Austin, TX: Cognitive Science Society.

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.

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS medicine*, *5*(10), e201.

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