Bayesian Explorations in Mathematical Psychology

Dóra Matzke

ISBN .....
This publication is typeset in LaTeX using the Memoir class.
Printed by Ipskamp Drukkers B.V., Enschede
Cover by Harm Erbé

Bayesian Explorations in Mathematical Psychology

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 5 december 2014, te 12.00 uur

door

Dóra Matzke
geboren te Boedapest, Hongarije

Promotiecommissie

Promotor:	Prof. dr. E.-J. Wagenmakers
Co-promotor:	Prof. dr. C. V. Dolan
Overige Leden:	Dr. D. Kellen
	Prof. dr. M. D. Lee
	Prof. dr. G. D. Logan
	Prof. dr. K. R. Ridderinkhof
	Prof. dr. H. L. J. van der Maas
	Dr. I. Visser

Faculteit der Maatschappij—en Gedragswetenschappen

This dissertation is dedicated to my father (1944-2004).
Látod, papa, megy ez a matek.

Contents

Chapter 1

Introduction

How can we best understand data obtained from psychological experiments? Throughout this dissertation, I will argue that this is best done by means of formal mathematical models. The goal of mathematical modeling is to capture regularities in the data using parameters that represent separate statistical of psychological processes. Mathematical models come in many flavors. Consider, for instance, the following two-choice task: You have to indicate with a button press whether an arrow presented on the computer screen points to the left or to the right. You perform this simple task, say, five times, and produce the following response times (RTs): 440, 409, 326, 482, and 511 ms. We may, for instance, model these five RTs with a normal distribution, $\text{RT}_i \sim \text{Normal}(\mu, \sigma)$, for $i = 1, ..., 5$, and use the sample mean (i.e., 433.6 ms) and the sample standard deviation (71.39 ms) as estimators for μ and σ; it is not a very sophisticated model, but it is model nevertheless.

If additional to the mean and the standard deviation, we also want to describe the —most likely right-skewed— shape of your RT distribution, we have to dig a bit deeper and rely on descriptive RT models such as the ex-Gaussian distribution. RT distributions, however, only provide a descriptive summary of the data and do not account for the psychological processes that underlie performance in a given experimental paradigm. If we want to relate the observed responses to psychological processes, we need to rely on more sophisticated —cognitive— models that go beyond statistical description. Performance in our hypothetical RT experiment, for instance, may be accounted for by the diffusion model (Ratcliff, 1978), one of the most prominent RT models with parameters that correspond to well-defined cognitive processes, such as the rate of information accumulation and response caution.

Let us now make the hypothetical RT experiment slightly more difficult: On some of the trials, the arrow is followed by a tone that tells you to withhold —inhibit– your response on that trial. We are no longer interested in the left-right responses, rather we would like to measure the time required to stop the primary response to the arrow. In this situation, we necessarily have to go beyond observable measurements; after all, stopping latencies cannot be observed directly. We may, for instance, formalize our assumptions about the cognitive processes that underlie successful response inhibition using the independent horse race model (Logan & Cowan, 1984). If we conceptualize response inhibition as a horse race between a go process and a stop process, we can derive and estimate the unobservable latency of stopping. Even better, if we are willing to make parametric assumptions about the distribution of the finishing times of the go and the stop process, we can estimate the entire distribution of stopping times.

Mathematical models can take a variety of forms and this dissertation mirrors this diversity. In addition to descriptive and process models of performance in two-choice RT tasks, I will focus on multinomial processing tree models for the analysis of categorical data as well as well-known statistical models, such as the t test, analysis of variance, (partial) correlations, structural equation models, and mediation analysis.

Once we chose a mathematical model for our data, we need to estimate the model parameters and assess the degree to which the chosen model provides an adequate description of the data. How can we best analyze psychological data sets that are described with mathematical models? Throughout this dissertation, I will argue that this is best done by means of Bayesian inference. In Bayesian inference, uncertainty is expressed in terms of probability. We start with a prior probability distribution that quantifies our existing beliefs about the state of the world. The prior is then updated by the incoming data using Bayes' rule to yield the posterior probability distribution. The posterior quantifies our updated beliefs about the state of the world and takes account of the prior as well as the observed data. In fact, the "...Bayesian approach is a common sense approach..." (Edwards, Lindman, & Savage, 1963, p. 195) that describes how rational decision makers should revise their opinion after considering incoming data. Although contemporary psychology heavily relies on frequentist inference, the application of Bayesian methods in psychological research has increased greatly over the past decade. This increase is partly fueled by the development of Markov chain Monte Carlo sampling (MCMC; Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996), the introduction of Bayesian statistical software, such as WinBUGS, and the availability of the Bayesian equivalent of popular hypothesis tests (e.g., Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Grasman, & Wagenmakers, 2012).

In this dissertation, I will focus on two applications of Bayesian inference: parameter estimation and model selection. With respect to parameter estimation, I will demonstrate that the Bayesian approach can be extremely valuable for hierarchical models where maximum likelihood estimation becomes practically difficult. With respect to model selection, I will rely on Bayes factors to measure statistical evidence in the context of competing cognitive models as well as standard statistical tests, such as the t test. I will make a case for the use of Bayesian model selection for the following reasons. First, Bayesian inference —as opposed to frequentist inference— does not depend on the intention with which the data were collected and hence does not require adjustments for sequential testing. Second, Bayesian hypothesis testing —as opposed to its frequentist counterpart— enables researchers to measure evidence in favor of the null hypothesis. Finally, Bayes factors —as opposed to p values— quantify the probability of the data under one hypothesis relative to the other; that is, Bayes factors tell researchers what they arguably would like to know in the first place when they engage in hypothesis testing.

The focus on Bayesian inference does not imply that frequentist solutions are always inappropriate nor that we are unable to reach valid conclusions relying on classical statistics. In fact, throughout the dissertation, the reader will stumble upon a number of p values and various other concepts related to classical null hypothesis testing. I will nevertheless argue that the Bayesian approach provides important theoretical and practical advantages that make it particularly suited for tackling problems that research psychologists face on a daily basis.

In the remainder of the introduction, I will give an overview and a brief description of the problems to be addressed in the coming chapters. The dissertation encompasses variety of topics; the glue that holds the diversity of topics together is the commitment to mathematical modeling and principled statistical inference.

1.1 Chapter Outline

Part I. The Analysis of Response Time Distributions

The first part of the dissertation focuses on modeling RTs —observed and unobserved— in two-choice tasks with descriptive RT models, such as the ex-Gaussian and the shifted Wald distributions.

In Chapter 2, I investigate the validity of the cognitive interpretation of the parameters of the ex-Gaussian and shifted Wald distributions. The ex-Gauss and the shifted Wald are popular RT distributions to summarize RT data for speeded two-choice tasks. The parameters of these distributions are often interpreted in terms of specific cognitive processes. We study the validity of this interpretation by relating the parameters of the ex-Gaussian and shifted Wald distributions to those of the Ratcliff diffusion model (Ratcliff, 1978), a successful model whose parameters have a well-established cognitive interpretation (e.g.,Voss, Rothermund, & Voss, 2004). The results clearly demonstrate that the ex-Gaussian and shifted Wald parameters do not correspond uniquely to parameters of the diffusion model.

In Chapter 3, I introduce a Bayesian parametric approach for the estimation of stopping latencies (SSRTs) in the stop-signal paradigm, a popular experimental paradigm to study response inhibition. Based on the horse race model (Logan & Cowan, 1984), several methods have been developed to estimate SSRTs. However, none of these approaches allow for the accurate estimation of the entire distribution of SSRTs. Here we introduce a Bayesian parametric approach that addresses this limitation. The new method assumes that SSRTs are ex-Gaussian distributed and uses MCMC sampling to obtain posterior distributions for the model parameters. We present the results of a number of parameter recovery and robustness studies and apply the approach to published data from a stop-signal experiment.

In Chapter 4, I present BEESTS, an efficient and user-friendly software implementation of the Bayesian parametric approach introduced in Chapter 3. BEESTS comes with an easy-to-use graphical user interface and provides users with summary statistics of the posterior distribution of the parameters as well various diagnostic tools to assess the quality of the parameter estimates. We illustrate the use of BEESTS with published stop-signal data.

Part II. Multinomial Processing Tree Models

The second part of the dissertation focuses on parameter estimation and model selection for multinomial processing tree (MPT) models. MPT models are theoretically motivated stochastic models for categorical data. Due to their simplicity, MPT models have been applied to a variety of areas in cognitive psychology (e.g., Batchelder & Riefer, 1999).

In Chapter 5, I introduce a Bayesian approach that accounts for parameter heterogeneity in MPT models as a result of differences between participants as well as items. Traditionally, statistical analysis for MPT models is carried out on aggregated data, assuming homogeneity in participants and items (Hu & Batchelder, 1994). In many applications, however, it is reasonable to treat both participant and items effects as random and base statistical inference on unaggregated data. Here we focus on a crossed-random effects extension of the pair-clustering model (Batchelder & Riefer, 1980), one of the most extensively studied MPT models for the analysis of free recall data. We apply the crossed-random effects pair-clustering model to novel experimental data featuring the manipulation of word frequency.

In Chapter 6, I present various procedures for model comparison in the context of MPT models. The topic of quantitative model comparison has received —and continues to receive— considerable attention (Pitt & Myung, 2002). Here we focus on two popular information criteria, the AIC ("an information criterion", Akaike, 1973) and the BIC ("Bayesian information criterion", G. Schwarz, 1978), on the Fisher information approximation of the minimum description length principle (MDL; Grünwald, 2007), and on Bayes factors as obtained from importance sampling (Hammersley & Handscomb, 1964).

Part III. Correlations, Partial Correlations, and Mediation

The third part of the dissertation focuses on estimating and testing observed and unobserved (partial) correlations.

In Chapter 7, I examine the power to reject the hypothesis of perfect correlation in the context of higher-order structural equation models. In higher-order factor models, general intelligence (g) is often found to correlate perfectly with lower-order common factors, suggesting that g and some well-defined cognitive ability, such as working memory, may be identical. Here we investigate the power to reject the equivalence of g and lower-order factors using artificial data sets, based on realistic parameter values and on the results of selected publications. The results of the power analyses indicate that most case studies that reported a perfect correlation between g and a lower-order factor were severely underpowered to detect the distinctiveness of the two factors.

In Chapter 8, I discuss a Bayesian method for correcting the correlation coefficient for the uncertainty of the observations. The Pearson product-moment correlation coefficient can be severely underestimated when the observations are subject to measurement noise. Various approaches exist to correct the estimation of the correlation in the presence of measurement error, but none are routinely applied in psychological research. Here we outline a Bayesian correction method for the attenuation of correlations proposed by Behseta, Berdyyeva, Olson, and Kass (2009). We illustrate the Bayesian correction with two empirical data sets and demonstrate that its application can substantially increase the correlation between noisy observations.

In Chapter 9, I discuss a default Bayesian hypothesis test for mediation. In order to quantify the relationship between multiple variables, researchers often carry out a mediation analysis. In such an analysis, a mediator (e.g., knowledge of healthy diet) transmits the effect from an independent variable (e.g., classroom instruction on healthy diet) to a dependent variable (e.g., consumption of fruits and vegetables). Almost all mediation analyses in psychology use frequentist parameter estimation and hypothesis testing techniques. Here we describe a default Bayesian hypothesis test based on the Jeffreys-Zellner-Siow approach (Rouder et al., 2009).

Part IV. Improving Research Practice

The fourth and final part of the dissertation focuses on suboptimal research practices in psychology.

In Chapter 10, I introduce a novel variant of proponent-skeptic collaboration that focuses on the association between horizontal eye movements and episodic memory. A growing body of research suggests that horizontal saccadic eye movements facilitate the retrieval of episodic memories in free recall and recognition memory tasks. Nevertheless, a minority of studies have failed to replicate this effect. Here we attempt to resolve the inconsistent results by introducing a novel variant of proponent-skeptic joint research. The proposed approach combines the features of adversarial collaboration (Kahneman, 2003) and purely confirmatory preregistered research (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). As anticipated by the skeptics, the results of a series of Bayesian hypothesis tests indicate that horizontal eye movements did not improve free recall performance in the joint experiment.

In Chapter 11, I describe a comparison of the statistical evidence provided by p values, effect sizes, and default Bayes factors. Statistical inference in psychology heavily relies on p value significance testing. This traditional approach, however, has been widely criticized. Here we present a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. The comparison yields two main results. First, although p values and default Bayes factors almost always agree about what hypothesis is better supported by the data, the measures often disagree about the strength of this support. That is, 70% of the p values that fall between .01 and .05 correspond to Bayes factors that indicate that the data are no more than three times more likely under the alternative hypothesis than under the null hypothesis. Second, effect sizes can provide additional evidence to p values and default Bayes factors.

In the twelfth and final chapter, I focus on multiway analysis of variance (ANOVA). Many empirical researchers do not realize that the common multiway ANOVA harbors a multiple comparison problem. We illustrate the use of the sequential Bonferroni (Hartley, 1955) correction for multiple comparison and show that its application often alters the conclusions drawn from ANOVA designs.

Part I

The Analysis of Response Time Distributions

Chapter 2

Psychological Interpretation of the Ex-Gaussian and Shifted Wald Parameters: A Diffusion Model Analysis

This chapter has been published as:
Dora Matzke and Eric-Jan Wagenmakers (2009).
Psychological interpretation of the ex-Gaussian and shifted Wald Parameters: A diffusion model analysis.
Psychonomic Bulletin & Review, 16, 798–817.[1]

Abstract

A growing number of researchers use descriptive distributions such as the ex–Gaussian and the shifted Wald to summarize response time data for speeded two–choice tasks. Some of these researchers also assume that the parameters of these distributions uniquely correspond to specific cognitive processes. We studied the validity of this cognitive interpretation by relating the parameters of the ex–Gaussian and shifted Wald distributions to those of the Ratcliff diffusion model, a successful model whose parameters have a well–established cognitive interpretation. In a simulation study, we fitted the ex–Gaussian and shifted Wald distributions to data generated from the diffusion model by systematically varying its parameters across a wide range of plausible values. In an empirical study, the two descriptive distributions were fitted to published data that featured manipulations of task difficulty, response caution, and a priori bias. The results clearly demonstrate that the ex–Gaussian and shifted Wald parameters do not correspond uniquely to parameters of the diffusion model. We conclude that researchers should resist temptation to interpret changes in the ex–Gaussian and shifted Wald parameters in terms of cognitive processes.

[1] The final publication is available at `http://link.springer.com/article/10.3758%2FPBR.16.5.798`.

2.1 Introduction

The analysis of response times (RT) has a long history in cognitive psychology (e.g., Hohle, 1965; Luce, 1986; Ratcliff & McKoon, 2008; Townsend & Ashby, 1983). To draw inferences about mental processes, researchers originally relied on measures of central tendency such as mean RT and median RT. As it became clear that these measures may lose important information (e.g., Heathcote, Popiel, & Mewhort, 1991), a growing number of researchers has started to use mathematical and statistical models that can accommodate not just mean RT, but also the shape of entire RT distributions.

Primary among the statistical models that facilitate the analysis of RT distributions are the ex–Gaussian and the shifted Wald. Changes in the parameters of these distributions may be used to summarize the effects of experimental manipulations. For instance, Leth-Steensen, King Elbaz, and Douglas (2000) found that children with ADHD differed from age–matched controls specifically in the ex–Gaussian parameter that captures the tail of the RT distribution.

Although the ex–Gaussian and the shifted Wald distributions are sometimes used as purely descriptive tools (see, e.g., Wagenmakers, van der Maas, Dolan, & Grasman, 2008), many researchers go one step further and assume that changes in the parameters of these distributions map on to changes in specific cognitive processes. For instance, Kieffaber et al. (2006) argued that changes in the Gaussian component of the ex–Gaussian distribution reflect changes in attentional cognitive processes, whereas changes in the exponential component reflect changes in intentional cognitive processes. The purpose of our study is to examine whether this mapping from parameters to processes is warranted. To this end, we attempt to link the parameters of the descriptive distributions to those of the Ratcliff diffusion model (Ratcliff, 1978). The diffusion model provides a theoretical account of performance in speeded two–choice tasks and it has been successfully applied across a wide range of paradigms. Most importantly, the parameters of the diffusion model correspond to well–defined psychological processes such as the rate of information accumulation (influenced by task difficulty or subject ability), response caution, a priori bias, and the time taken up by processes unrelated to decision making (e.g., encoding and motor processes). The association between the diffusion model parameters and the psychological processes that they are supposed to represent has been confirmed in numerous experiments (e.g., Voss et al., 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

The outline of this article is as follows. In the first and second section we describe the ex–Gaussian and shifted Wald distributions, respectively. In the third section we briefly explain the diffusion model and previous research that studied the association between the diffusion model parameters and the parameters of the two descriptive distributions. The fourth section describes the simulation study in which we systematically varied parameters of the diffusion model to study the corresponding changes in the parameters of the descriptive distributions. In the fifth section we apply the descriptive distributions to recently published lexical decision data that feature manipulations of task difficulty, response caution, and a priori bias (Wagenmakers, Ratcliff, et al., 2008). The sixth section concludes our investigation.

2.2 The Ex–Gaussian Distribution

The ex–Gaussian distribution results from the convolution of a Gaussian and an exponential distribution and can be described by three parameters: μ and σ, the mean and the standard deviation of the Gaussian component, and τ, the mean of the exponential component. Roughly, μ and σ reflect the leading edge and τ reflects the tail of the distribution. The ex–Gaussian distribution has a positively skewed unimodal shape and generally produces an excellent fit to empirical RT distributions. Figure 2.1 shows changes in the shape of the ex–Gaussian distribution as a result of changes in the ex–Gaussian parameters μ, σ, and τ. The probability density function of the ex–Gaussian is given by

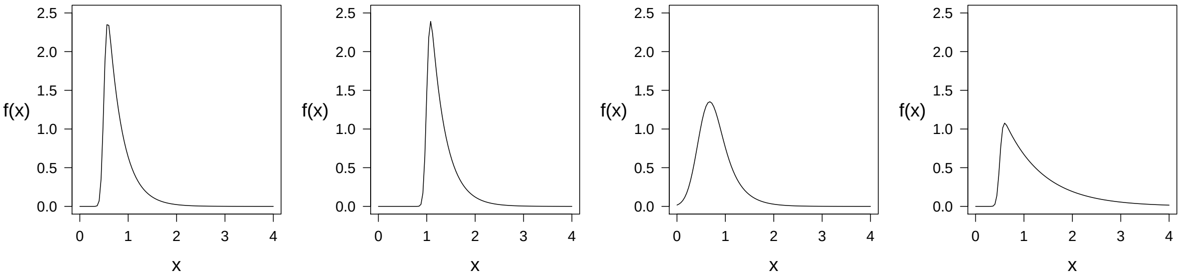
$$f(x|\mu,\sigma,\tau) = \frac{1}{\tau\sqrt{2\pi}} \exp\left(\frac{\sigma^2}{2\tau^2} - \frac{x-\mu}{\tau}\right) \int_{-\infty}^{[(x-\mu)/\sigma]-(\sigma/\tau)} \exp\left(-\frac{y^2}{2}\right) dy \tag{2.1}$$

and its mean and variance are

$$E(x) = \mu + \tau \tag{2.2}$$

and

$$Var(x) = \sigma^2 + \tau^2. \tag{2.3}$$



(a) Default parameter set (b) Increasing μ (c) Increasing σ (d) Increasing τ

Figure 2.1 *Changes in the shape of the ex–Gaussian distribution as a result of changes in the ex–Gaussian parameters μ, σ, and τ.* The parameter sets used to generate the distributions are μ=0.5, σ=0.05, τ=0.3 (panel *a*), μ=1, σ=0.05, τ=0.3 (panel *b*), μ=0.5, σ=0.2, τ=0.3 (panel *c*), and μ=0.5, σ=0.05, τ=0.8 (panel *d*).

Originally, the ex–Gaussian distribution was thought to represent the durations of two successive components of cognitive processing. In particular, Hohle (1965) suggested that the exponential component represents “the decision and perceptual portion of an RT”, whereas the Gaussian component reflects “the time required for organization and execution of the motor response” (p. 384). Although it may be tempting to associate these particular cognitive processes with the ex–Gaussian parameters, Hohle’s interpretation of the ex–Gaussian parameters has been frequently challenged.

First, there is disagreement as to which processing mechanism should be attributed to the two ex–Gaussian components. McGill (1963) and McGill and Gibbon (1965), for example, suggested that residual motor latency corresponds to the exponential component of the ex–Gaussian, not to the Gaussian component. This interpretation is diametrically opposed to that of Hohle (1965).

Second, the rationale underlying Hohle’s (1965) interpretation of the two ex–Gaussian components has been criticized. Hohle based his interpretation on the finding that the mean of the Gaussian component, μ, was somewhat more sensitive to manipulations of foreperiod duration, but the mean of the exponential component, τ, was more sensitive to manipulations of stimulus intensity. Because foreperiod duration was assumed to influence motor response time and stimulus intensity was assumed to influence decision time, Hohle concluded that the Gaussian component must reflect residual time and the exponential component the decision portion of the RT.

Luce (1986, pp. 100–102), however, has argued that Hohle's (1965) data are equally consistent with many other decompositions of RT, and that the influence of foreperiod duration can in fact be attributed to either component of RT. Also, experimental results often fail to support the differential sensitivity of the ex–Gaussian parameters to manipulations that are assumed to influence the decision component of RT. For example, the manipulation of word frequency, which is assumed to affect the decision portion of RT, commonly influences the τ as well as the μ parameter (e.g., Andrews & Heathcote, 2001; Plourde & Besner, 1997; Yap & Balota, 2007; Yap, Balota, Cortese, & Watson, 2006). Table 2.1 summarizes the effects that experimental manipulations have on the ex–Gaussian parameters, on the basis of a literature review of 54 studies.[2]

Third, and most important, it has been argued that the ex–Gaussian distribution lacks a plausible theoretical basis, and that it is consequently unable to account for the psychological mechanisms that drive performance (see, e.g., Heathcote et al., 1991; Luce, 1986). Specifically, the Gaussian component necessarily assigns positive probability to negative RTs, a conceptual inadequacy that highlights the fact that the ex–Gaussian distribution can never correspond to a plausible cognitive process model.

Table 2.1 The Effects of Experimental Manipulations on the Ex–Gaussian Parameters.

Experimental manipulation	N	μ	σ	τ
Word frequency (high or low)	18	16	7	12
Flanker task condition (no flanker, neutral, congruent, and incongruent)	14	13	7	4
Age (children, young adults, and older adults)	11	8	8	7
Length of study list	13	6	2	10
Number of stimulus presentations	9	1	0	8
Stimulus quality (clear or degraded)	9	8	3	8
Stroop task condition (neutral, congruent, and incongruent)	8	8	4	6
Study position of probe items	12	12	1	9
Local/global task^a condition (neutral, congruent, and incongruent)	6	4	2	1
Output position of recalled items	5	0	0	5
Animacy of stimulus	4	1	1	2
Length of retention interval	4	1	0	4
Nonword type (pseudohomophone, legal, and illegal)	4	4	0	4
Cue–to–target stimulus onset asynchrony	3	3	0	0
Interstimulus interval	3	3	3	3
Speed–accuracy instruction	2	2	0	2

Note. Table 2.1 summarizes the results of an extensive literature review, covering 54 applications of the ex–Gaussian distribution. The summary is created by selecting the most frequently used experimental manipulations encountered in the literature and tallying how often, out of N attempts, the manipulations influenced each ex–Gaussian parameter. The criterion that researchers used to evaluate whether a given experimental manipulation influenced the ex–Gaussian parameters varied across the experiments. The criterion either was one of < 0.1, $p < 0.05$, $p < 0.001$, or was based on visual inspection of the changes in parameter values. ^{a} An example of a congruent stimulus in the local/global task is the letter H, constructed with small Hs. An example of an incongruent stimulus is the letter H, constructed with small Zs. An example of a neutral stimulus is a circle, constructed with small Hs.

As a consequence of its problematic theoretical underpinning, some researchers have adopted a cautious attitude and warned against the cognitive interpretation of the ex–Gaussian parameters.

[2] An extensive overview of the effects of experimental manipulations on the ex–Gaussian parameters, including interaction effects, is available in the supplemental materials at `http://dora.erbe-matzke.com/publications.html`.

As Heathcote et al. (1991) stated, “Although the ex–Gaussian model describes RT data successfully, it does so without the benefit of an underlying theory.” (p. 346). Consistent with this view, the ex–Gaussian distribution has been sometimes used as an economical three–parameter summary of RT data and as a tool to evaluate the predictions of competing cognitive models beyond the level of mean RT (e.g., Heathcote et al., 1991; Hockley, 1982, 1984; Ratcliff, 1978, 1993; Ratcliff & Murdock, 1976).

Other researchers, however, have not been so cautious and persisted on the substantive interpretation of the ex–Gaussian parameters. Rohrer and Wixted (1994), for example, interpreted the Gaussian component as “a brief initiation that precedes retrieval” and the exponential component as “an ongoing search” (p. 512–513). Balota and Spieler (1999) related the Gaussian component to “more stimulus driven automatic (nonanalytic) processes”, and the exponential component to “more central attention demanding (analytic) processes” (p. 34). Kieffaber et al. (2006) interpreted μ in terms of attentional and τ in terms of intentional cognitive processes (p. 348). As a final example, Gordon and Carson (1990) argued that the “lumped sensory input/motor output component” of RT has a Gaussian distribution, and that the “decisional phase” of RT has an exponential distribution (p. 150; see also Madden et al., 1999; Possamaï, 1991; Rotello & Zeng, 2008 for similar interpretations). Table 2.2 gives an overview of the cognitive interpretations attributed to the ex–Gaussian parameters;[3] since the σ parameter is rarely given a cognitive interpretation, it is omitted from the overview. As can be seen in Table 2.2, there is some consistency in the cognitive interpretation of the ex–Gaussian parameters: Lower–order processes are generally ascribed to μ and higher–order processes are generally ascribed to τ. Note, however, that the precise interpretation of the ex–Gaussian parameters varies considerably across researchers.

To summarize, the ex–Gaussian distribution provides a description of empirical RT data that is accurate but lacks a plausible theoretical rationale. Despite this limitation, the ex–Gaussian parameters have often been interpreted in terms of underlying cognitive processes. The following section introduces the shifted Wald distribution, a descriptive distribution that has the potential to provide parameters that are theoretically more meaningful.

2.3 The Shifted Wald Distribution

The Wald distribution (Wald, 1947) represents the density of the first passage times of a Wiener diffusion process toward a single absorbing boundary (see Figure 2.2). This distribution can be characterized by two parameters: γ, reflecting the drift rate of the diffusion process, and α, reflecting the separation between the starting point of the diffusion process and an absorbing barrier. In the RT context, the Wald distribution is often supplemented with a positive parameter θ that shifts the entire RT distribution. The shifted Wald has a positively skewed unimodal shape that generally produces an excellent fit to empirical RT distributions. Figure 2.3 shows changes in the shape of the shifted Wald distribution as a result of changes in the parameters α, θ, and γ. The probability density function of the shifted Wald is given by

$$f(x|\alpha,\theta,\gamma) = \frac{\alpha}{\sqrt{2\pi(x-\theta)^3}} \exp\left\{-\frac{[\alpha-\gamma(x-\theta)]^2}{2(x-\theta)}\right\}, \tag{2.4}$$

where $x > \theta$, and its mean and variance are

$$E(x) = \theta + \alpha/\gamma \tag{2.5}$$

[3] Specific quotations are available in the supplemental materials.

Table 2.2 Cognitive Interpretations Attributed to the Ex–Gaussian Parameters

Authors	μ	τ
Balota and Spieler (1999)	stimulus driven automatic (nonanalytic) processes	central attention demanding (analytic) processes
Blough (1988, 1989)	component of RT unrelated to stimulus variables (e.g., neural transmission and motor response)	momentary probability of target detection/search component of RT
Epstein et al. (2006); Leth-Steensen et al. (2000)	–	attentional lapses
Gholson and Hohle (1968a, 1968b)	–	response choice latency/response competition
Gordon and Carson (1990); Hohle (1965); Madden et al. (1999); Possamaï (1991); Rotello and Zeng (2008)	duration of residual processes (e.g., sensory and motor processes)	durations of the decisional phase of RT
Kieffaber et al. (2006)	attentional cognitive processes	intentional cognitive processes
Penner-Wilger, Leth-Steensen, and LeFevre (2002)	retrieval processes	nonretrieval/procedure use
Rohrer (1996, 2002); Rohrer and Wixted (1994); Wixted, Ghadisha, and Vera (1997); Wixted and Rohrer (1993)	initial pause preceding the retrieval of the first response	mean recall latency/ongoing memory search
Schmiedek, Oberauer, Wilhelm, Suss, and Wittmann (2007)	–	higher cognitive functioning (e.g.,working memory and reasoning)
Spieler, Balota, and Faust (1996)	–	more central processing component

Note. A dash indicates that the parameter is not given any cognitive interpretation.

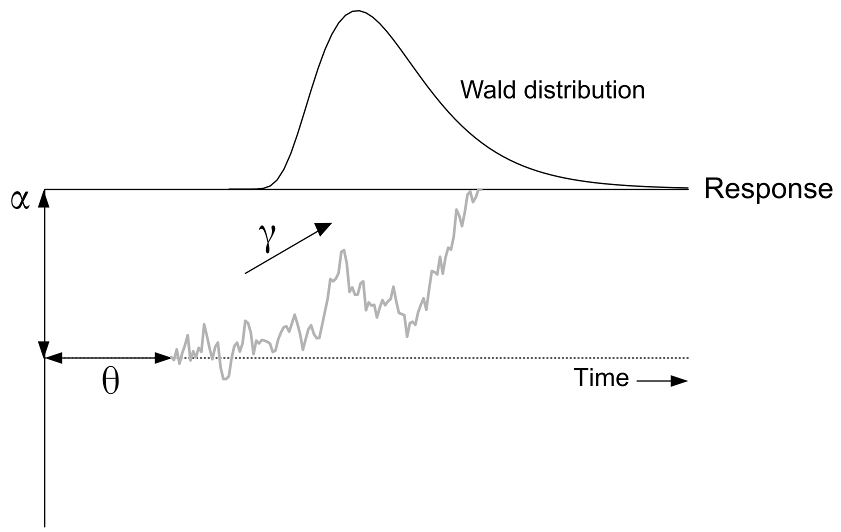


Figure 2.2 *The shifted Wald model of RT and its parameters. See text for details.*

and

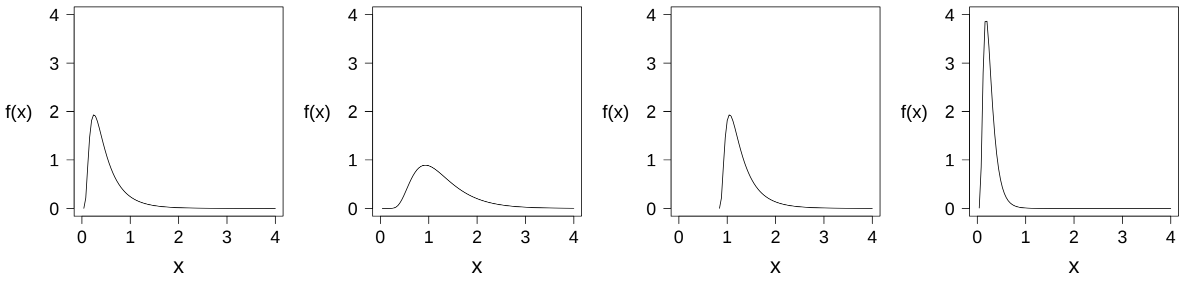
$$Var(x) = \alpha/\gamma^3. \tag{2.6}$$

For further discussion and applications of the Wald distribution, see Burbeck and Luce (1982), Luce (1986), and Emerson (1970). For discussion and application of a more general version of a single–boundary diffusion process, see P. L. Smith (1995).

The cognitive interpretation of the shifted Wald parameters is straightforward (see, e.g., Heathcote, 2004; Luce, 1986; W. Schwarz, 2001, 2002). Participants are assumed to accumulate noisy information until a predefined threshold amount is reached and a response is initiated. Drift rate γ quantifies task difficulty or subject ability, response criterion α quantifies response caution, and the shift parameter θ quantifies the time needed for non–decision processes.

Although the shifted Wald distribution has a sound theoretical basis, the cognitive interpretation of its parameters has rarely been subject to empirical validation. The shifted Wald model may be particularly suited for paradigms in which there is likely only a single response boundary. Such paradigms may include simple RT tasks (Luce, 1986, pp. 51–57), go/no–go tasks (Heathcote, 2004; W. Schwarz, 2001) or tasks that involve saccadic eye movements that result in very few errors (Carpenter & Williams, 1995). It is not clear whether the cognitive interpretation of the shifted Wald parameters still holds when the distribution is applied to data from a paradigm that clearly involves two response alternatives.

In summary, both the ex–Gaussian and the shifted Wald distributions provide excellent tools to summarize RT distributions. However, the cognitive interpretation of their parameters is unclear. The ex–Gaussian distribution lacks an adequate theoretical basis and the substantive interpretation of its parameters has been repeatedly questioned. Although the shifted Wald distribution is theoretically better justified, it is currently unclear whether the substantive interpretation of its parameters carry over from one–boundary paradigms to two–boundary paradigms.



(a) Default parameter set (b) Increasing α (c) Increasing θ (d) Increasing γ

Figure 2.3 *Changes in the shape of the shifted Wald distribution as a result of changes in the shifted Wald parameters α, θ, and γ.* The parameter sets used to generate the distributions are α=1, θ=0, γ=2 (panel *a*), α=2.5, θ=0, γ=2 (panel *b*), α=1, θ=0.8, γ=2 (panel *c*), and α=1, θ=0, γ=3.8 (panel *d*).

2.4 The Ratcliff Diffusion Model

The diffusion model (Ratcliff, 1978; for reviews see Ratcliff & McKoon, 2008; Wagenmakers, 2009) is a prominent cognitive process model of speeded two–choice decisions. The diffusion model assumes that noisy information is accumulated over time from a starting point toward one of two response boundaries (see Figure 2.4). A response is initiated when one of the two response boundaries is reached. The diffusion model has been successfully applied to a wide range of experimental paradigms, including brightness discrimination, letter identification, lexical decision, recognition memory and signal detection (e.g., Ratcliff, 1978, 2002; Ratcliff, Gomez, & McKoon, 2004; Ratcliff & Rouder, 2000; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2001, 2003, 2004; Thapar, Ratcliff, & McKoon, 2003; Wagenmakers, Ratcliff, et al., 2008). The diffusion model generally provides an excellent fit to all aspects of the observed RT data, including response accuracy and the RT distributions of both correct and error responses. As indicated by Ratcliff and McKoon (2008, p. 918), "(...) the class of diffusion models has as near to provided a solution to simple decision making as is possible in behavioral science."

One of the major strengths of the diffusion model is its ability to provide parameter estimates that can be interpreted in terms of the cognitive components underlying the decision process (e.g., Voss et al., 2004). The central parameters of the model are drift rate v, boundary separation a, starting point z, and nondecision time T_{er}. Drift rate v represents the mean rate of information accumulation and it is determined by the quality of information that is extracted from the stimulus. Drift rate can be influenced either by individual differences in the quality of information processing or by stimulus characteristics that reflect task difficulty. Boundary separation a quantifies the distance between the two response boundaries and represents response caution. Large values of a indicate that more information must be accumulated before a decision can be made. Boundary separation is usually manipulated via speed–accuracy instructions. Starting point z represents subjects' a priori bias for one of the two response alternatives. Starting point is usually manipulated either by varying the proportion of stimuli that is associated with the upper and the lower response boundaries or by payoff manipulations. Both boundary separation a and starting point z are assumed to be under the subjective control of participants. Nondecision time T_{er} quantifies the duration of processes that are unrelated to the decision process, including stimulus encoding and response execution.

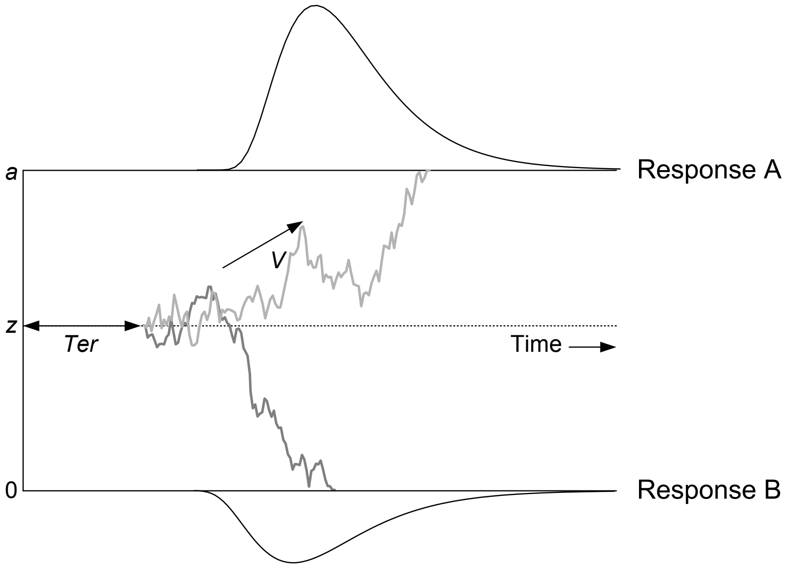


Figure 2.4 *The diffusion model and its central parameters. See text for details.*

In addition to these key parameters, the diffusion model also features parameters that describe how the values of these key parameters fluctuate from one trial to the next. Specifically, the model assumes across–trial variability in drift rate (according to a normal distribution with variance η), starting point (according to a uniform distribution with range s_z), and nondecision time (according to a uniform distribution with range s_t).[4]

To summarize, the diffusion model provides a general theoretical account of decision making in speeded two–choice tasks. Previous research has shown that the parameters of the model correspond to the psychological processes that they are assumed to represent (e.g., Ratcliff & McKoon, 2008; Voss et al., 2004; Wagenmakers, 2009). Therefore, the diffusion model can be used to judge whether the cognitive interpretation of the ex–Gaussian and shifted Wald parameters is warranted when these descriptive distributions are applied to data from speeded two–choice tasks. Ideally, the parameters of the ex–Gaussian and shifted Wald distributions would correspond uniquely with parameters of the diffusion model. For instance, one would hope that, say, a change in drift rate in the diffusion model would correspond to a change in the τ parameter in the ex–Gaussian and the γ parameter in the shifted Wald.

Links between the Ex–Gaussian and the Diffusion Model Parameters

Several attempts have been made to relate the ex–Gaussian parameters to those of the diffusion model. For instance, Schmiedek et al. (2007) showed that both the ex–Gaussian parameter τ and the diffusion model parameter v correlated strongly with people's higher cognitive functions such as working memory and reasoning. In addition, Schmiedek et al. demonstrated by simulations that the relation between τ and higher cognitive functions could be fully explained in terms of

[4] The model also features a parameter s that quantifies the amplitude of the noise in the information accumulation process. This scaling parameter is usually fixed at 0.1, a convention that we adhere to throughout the present article.

individual differences in drift rate v. Schmiedek et al. concluded that τ is associated with drift rate v.

Other researchers adopted a different approach and used simulations to examine changes in the ex–Gaussian parameters as a result of changes in diffusion model drift rate v and boundary separation a parameters. The results typically showed that (1) an increase in drift rate mainly causes a decrease in τ and (2) an increase in boundary separation mainly causes an increase in μ (e.g., Spieler, 2001; Spieler, Balota, & Faust, 2000; Yap et al., 2006).

Finally, Ratcliff (1978) used the ex–Gaussian μ and τ parameters to fit the diffusion model to data obtained from various experimental paradigms, such as the study–test paradigm (e.g., Ratcliff & Murdock, 1976), the Sternberg paradigm (e.g., Sternberg, 1966), and the continuous recognition memory paradigm (e.g., Okada, 1971). Contrary to the simulation results above, Ratcliff found that μ and τ are both sensitive to changes in drift rate and boundary separation. In particular, the results indicated that (1) increases in drift rate and starting point cause decreases in both μ and τ and (2) an increase in boundary separation causes increases in both μ and τ.

The above results are therefore far from conclusive. Some studies (e.g., Spieler, 2001) report that τ and μ are selectively influenced by drift rate v and boundary separation a, respectively. Other studies (e.g., Ratcliff, 1978), however, show that τ and μ are sensitive to changes in a variety of diffusion model parameters. In addition, previous work examined only a limited range of values for the diffusion model parameters. A comprehensive investigation will require that the diffusion model parameters be manipulated on a realistic and sufficiently large range.

Links between the Shifted Wald and the Diffusion Model Parameters

To the best of our knowledge, no one has yet attempted to relate the shifted Wald parameters to those of the diffusion model. Nevertheless, both the shifted Wald distribution and the diffusion model conceptualize the decision process as a gradual information accumulation process. In fact, the shifted Wald can be thought of as a single–boundary diffusion process (cf. Figures 2.2 and 2.4). On the basis of the conceptual similarities of the two models, one might expect that (1) an increase in drift rate v mainly causes an increase in γ, (2) an increase in boundary separation a mainly causes an increase in α, (3) an increase in starting point z mainly causes a decrease in α, and (4) an increase in nondecison time T_{er} mainly causes an increase in θ.

Empirical evidence for some of these relations has been reported by W. Schwarz (2001). W. Schwarz (2001) used a go/no–go digit comparison task and manipulated numerical distance (1 or 4) and the prior probability of go trials (0.5 or 0.75). Shifted Wald analyses by Heathcote (2004) confirmed that the manipulation of numerical distance selectively influenced the Wald drift rate γ and that the manipulation of prior probability selectively influenced the Wald response criterion α. As expected, the Wald nondecision time θ was not influenced by either of these two manipulations. These results are encouraging, but it remains unclear to what extent the parameters from the shifted Wald correspond to those from the diffusion model in case the data are obtained in a speeded task that clearly features two response alternatives.

2.5 Validation of the Ex–Gaussian and Shifted Wald Parameters Using Diffusion Model Simulations

In this section, we investigate the association between parameters from the ex–Gaussian and the shifted Wald and parameters from the diffusion model. To this end, we simulated data from

Table 2.3 Minimum, Maximum, and Mean Values of the Diffusion Model Parameters Used in the Simulations

Diffusion Model Parameter	Minimum	Maximum	Mean
Drift rate v	0.0	0.586	0.223
Boundary separation a	0.056	0.393	0.125
Starting point z	0.028	0.182	0.063
Nondecision time T_{er}	0.206	0.942	0.435
Trial–to–trial variability in drift rate η	0.0	0.329	0.133
Trial–to–trial variability in starting point s_z	0.0	0.169	0.037
Trial–to–trial variability in nondecision time s_t	0.0	0.630	0.183
Bias z/a	0.272	0.782	-
s_z/a	0.0	0.900	-

the diffusion model by systematically varying its parameter values. Next, we fitted both the ex–Gaussian and the shifted Wald distributions to the simulated data sets.

Diffusion Model Simulations and Model Fitting

In each simulation, we generated data by manipulating a particular diffusion model parameter from a minimum to a maximum value while keeping the other parameters constant on their average values. Realistic parameter values were based on an extensive literature survey that covered 23 diffusion model applications. Table 2.3 shows the minimum, maximum, and mean values of the parameters used in the simulations. Note that some of the estimates from which the minimum, maximum, and mean values are derived result from parameter estimation with theoretically motivated constraints on some of the diffusion model parameters. For the histograms of the diffusion model parameter values found in the literature, the reader is referred to Appendix A.1.[5]

For the manipulation of boundary separation a, starting point z was assumed to be equidistant from the two response boundaries, so that $z = a/2$. Further, the manipulation of starting point z was carried out with respect to the mean value of boundary separation a by using the minimum and maximum values of the z/a ratios, the so–called bias parameters, found in the literature. A z/a ratio of 0.5 indicates that starting point z is equidistant from the two response boundaries. Similarly, the manipulation of the trial–to–trial variability of starting point s_z parameter was carried out with respect to the mean value of a by using the minimum and maximum values of the s_z/a ratios. Each parameter was manipulated in 1,000 steps of equal size, resulting in 1,000 data sets per parameter. In order to obtain relatively noise–free parameter estimates, each data set contained 10,000 RTs. The simulations were carried out using the Diffusion Model Analysis Toolbox (DMAT; Vandekerckhove & Tuerlinckx, 2007, 2008).

Next, the ex–Gaussian and shifted Wald distributions were fitted to the simulated data sets using maximum likelihood estimation (e.g., Myung, 2003). Extreme parameter estimates (i.e., 15 ex–Gaussian and 8 shifted Wald estimates) were removed from the analyses. Note that the descriptive distributions were fitted to the RTs of correct responses only.

[5] The exact parameter values are available in the supplemental materials.

Simulation Results

Figure 2.5 and Figure 2.6 show the changes in the ex–Gaussian and shifted Wald parameters as a function of changes in the diffusion model parameters. Table 2.4 gives a summary of the associations between the two sets of parameters. In this section, we present only the results related to the manipulation of the key diffusion model parameters: drift rate v, boundary separation a, starting point z, and nondecision time T_{er}. Because the across–trial variability parameters cannot be interpreted in terms of cognitive processes, the results related to these parameters are presented in Appendix A.2.

Table 2.4 The Associations Between Parameters of the Ex–Gaussian and Shifted Wald Distributions and Parameters of the Diffusion Model.

		Diffusion model parameters			
		v	a	z	T_{er}
Ex–Gaussian	μ	−	++	− −	++
	σ	−	+	−	×
	τ	− −	++	−	×
Shifted Wald	α	++	− −	− −	+
	θ	−	++	−	++
	γ	++	− −	−/+	×

Note. ++, substantial positive association; +, weak positive association; − −, substantial negative association; −, weak negative association; ×, no association; v, drift rate; a, boundary separation; z, starting point; T_{er}, nondecision time.

Ex–Gaussian Parameters

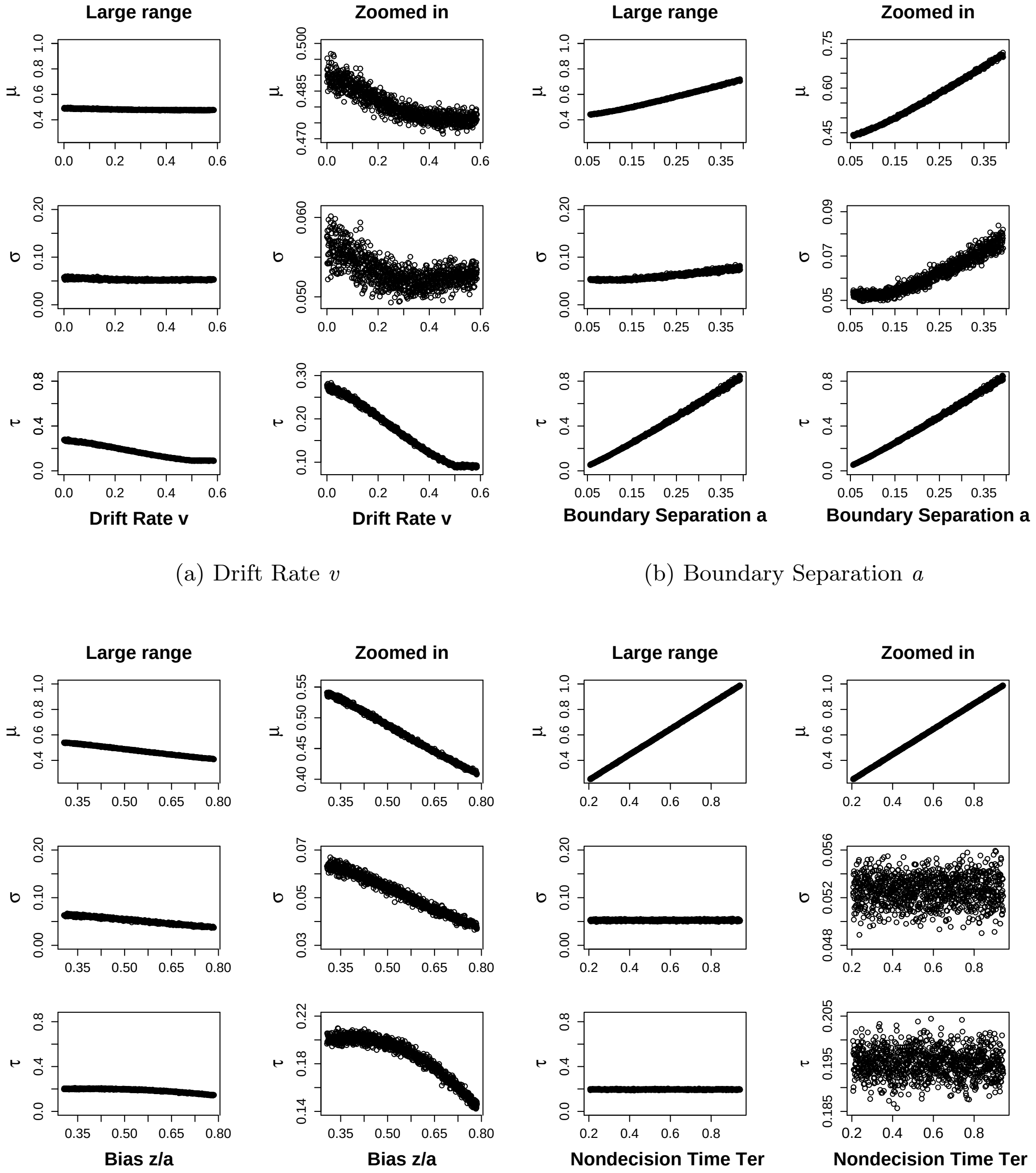
With respect to drift rate v, Figure 2.5a shows that the three ex–Gaussian parameters all decrease as v increases. The decrease in both μ and σ are, however, extremely small. In fact, changes in v are primarily reflected in τ. Also, τ continues to decrease until extreme values of v, whereas μ and σ level off already at intermediate values of v.

Turning to boundary separation a, Figure 2.5b shows that the three ex–Gaussian parameters all increase as a increases. Although τ increases more than either μ or σ, the increase in μ is also substantial. Note that τ changes substantially more as a function of a than as a function of any other diffusion model parameter.

With respect to starting point z, Figure 2.5c shows that the three ex–Gaussian parameters all decease as z increases. However, the decrease in both σ and τ are negligible. Also, τ seems relatively constant for low values of z. Changes in z are thus primarily reflected in μ.

Turning to nondecision time T_{er}, Figure 2.5d shows that μ increases as T_{er} increases. In contrast, both the τ and the σ parameters are unaffected by T_{er}. Note that μ changes substantially more as a function of T_{er} than as a function of any other diffusion model parameter.

To summarize, the results of the simulations indicate that the ex–Gaussian parameters do not correspond uniquely to parameters of the diffusion model. The μ parameter is substantially influenced by boundary separation a, starting point z, and nondecision time T_{er}. The σ parameter is not influenced substantially by any of the key diffusion model parameters, and τ is substantially influenced by both drift rate v and boundary separation a.



(a) Drift Rate v

(b) Boundary Separation a

(c) Starting Point z

(d) Nondecision Time T_{er}

Figure 2.5 *Changes in the ex–Gaussian parameters μ, σ, and τ as a function of systematic changes in the diffusion model parameters drift rate v (panel a), boundary separation a (panel b), starting point z (panel c), and nondecision time T_{er} (panel d).* The left-hand figures in each panel plot the results on scales ranging from the minimum to the maximum values of the ex–Gaussian parameters found across all simulations. The right-hand figures in each panel plot the same results on scales ranging from the minimum to the maximum values of the ex–Gaussian parameters found for the manipulation of the given diffusion model parameter.

Shifted Wald Parameters

With respect to drift rate v, Figure 2.6a shows that both α and γ increase as v increases. In contrast, θ seems to decrease with increasing v. However, the decrease in θ is extremely small. In fact, changes in drift rate v are primarily reflected in α and γ. Note that the three shifted Wald parameters all level off for high values of v and that α and θ are relatively constant for low values of v.

Turning to boundary separation a, Figure 2.6b shows that γ decreases and θ increases with increasing a. Quite unexpectedly, the response criterion α parameter *de*creases as a *in*creases. The changes in all three parameters are substantial. Note that the rates of decrease in both α and γ slow down as a increases. In fact, α levels off at intermediate values of a.

With respect to starting point z, Figure 2.6c shows that both α and θ decrease as z increases. In contrast, γ decreases for lower values and increases for higher values of z. The point of reversal in γ seems to correspond to the point where z is equidistant to the two response boundaries. However, the changes in both θ and γ are extremely small. Changes in z are thus primarily reflected in α.

Turning to nondecision time T_{er}, Figure 2.6d shows that γ is unresponsive to changes in T_{er}. In contrast, both α and θ increase as T_{er} increases. However, the increase in α is negligible and limited to high values of T_{er}. Changes in T_{er} are thus primarily reflected in θ. In fact, θ changes substantially more as a function of T_{er} than as a function of any other diffusion model parameter.

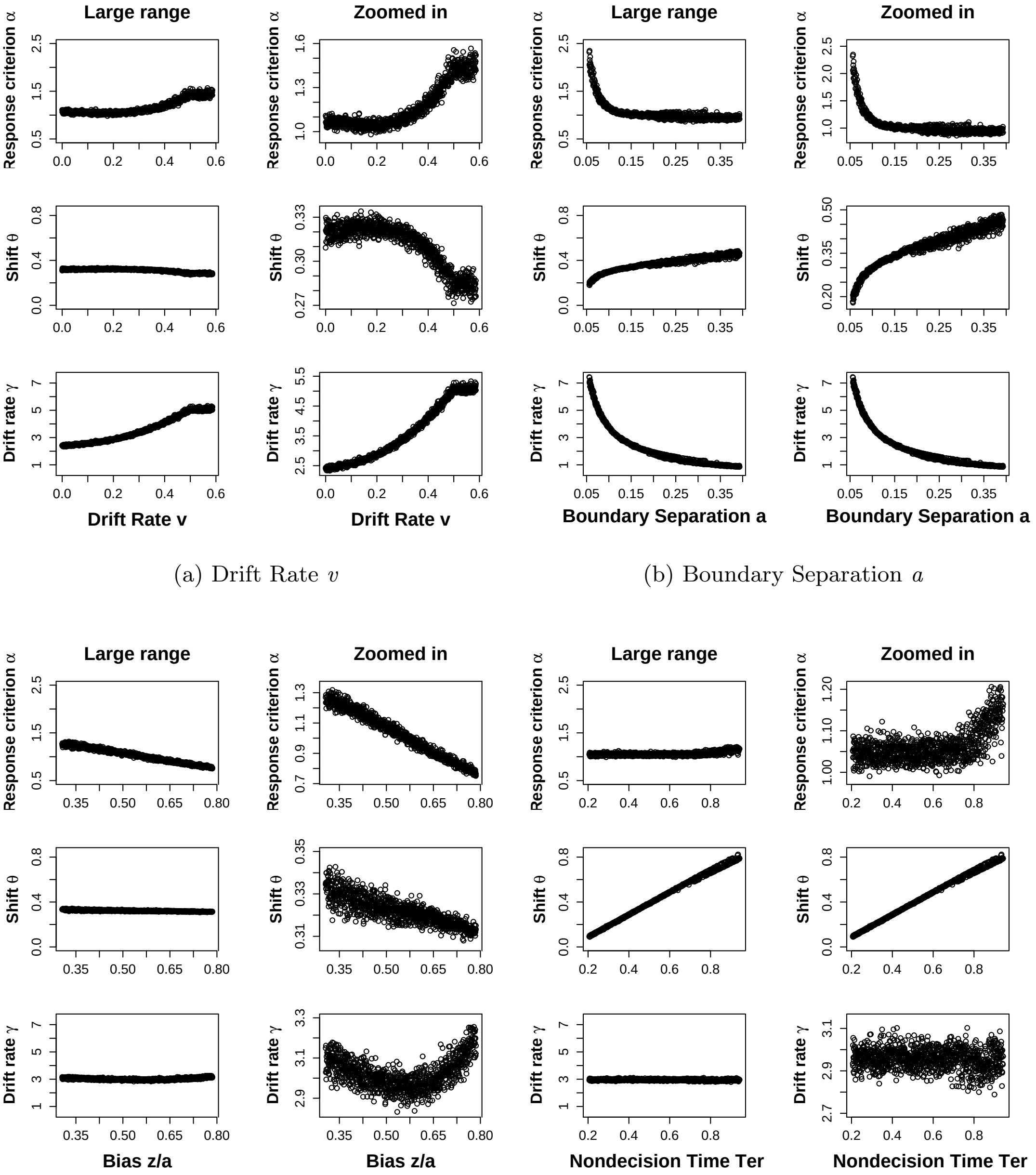
To summarize, the results of the simulations indicate that the shifted Wald parameters likewise do not correspond uniquely to parameters of the diffusion model. The α parameter is substantially influenced by drift rate v, boundary separation a, and starting point z. Remarkably, α decreases as a increases. The θ parameter is substantially influenced by both boundary separation a and nondecision time T_{er}. Finally, the γ parameter is substantially influenced by both drift rate v and boundary separation a.

2.6 Validation of the Ex–Gaussian and Shifted Wald Parameters Using Experimental Manipulations

In this section, we present a concrete empirical illustration of the simulation results reported above by examining how the parameters of the descriptive distributions relate to experimentally induced changes in the diffusion model parameters. Specifically, we investigate how the ex–Gaussian and shifted Wald parameters respond to experimental manipulations that selectively affect the key parameters of the diffusion model (i.e., drift rate v, boundary separation a, and starting point z). To this end, we fitted the ex–Gaussian and shifted Wald distributions to data sets obtained from two lexical decision experiments.

The Lexical Decision Data

To separately estimate the effects of lexical processing from the effects of strategic threshold adjustments, Wagenmakers, Ratcliff, et al. (2008) applied the diffusion model to the data of two lexical decision experiments. In the first experiment (N=15), task difficulty was manipulated on three levels by varying word frequency (high, low, and very low frequency), and response caution was manipulated on two levels by instructions and feedback that emphasized either response speed or response accuracy. The resulting 3 (word frequency) × 2 (speed–accuracy instruction) cells of the experimental design each contained 160 trials per participant. The diffusion model was able to account for the effects of the manipulations with only two parameters free to vary across conditions. The effects of word frequency were entirely accounted for by changes in drift rate v, with



(a) Drift Rate v

(b) Boundary Separation a

(c) Starting Point z

(d) Nondecision Time T_{er}

Figure 2.6 *Changes in the shifted Wald parameters α, θ, and γ as a function of systematic changes in the diffusion model parameters drift rate v (panel a), boundary separation a (panel b), starting point z (panel c), and nondecision time T_{er} (panel d).* The left-hand figures in each panel plot the results on scales ranging from the minimum to the maximum values of the shifted Wald parameters found across all simulations. The right-hand figures in each panel plot the same results on scales ranging from the minimum to the maximum values of the shifted Wald parameters found for the manipulation of the given diffusion model parameter.

higher word frequency associated with higher values of v. In contrast, the effects of the speed–accuracy instructions were entirely accounted for by changes in boundary separation a, with speed instructions associated with lower values of a.

In the second experiment (N=19), in addition to the task difficulty manipulation, participants' a priori bias was manipulated on two levels by varying the proportion of word versus nonword stimuli in a list (i.e., 75% words or 75% nonwords). The resulting 3 (word frequency) × 2 (word/nonword proportion) cells of the experimental design each contained 160 trials per participant. As in Experiment 1, the effects of word frequency were entirely accounted for by changes in drift rate v. In contrast, the effects of the proportion manipulation were accounted for by changes in starting point z, with the 75% word condition associated with higher values of z (in the modeling, the upper and lower boundaries were associated with the "word" and "nonword" responses, respectively).

We fitted this data set using the ex–Gaussian and shifted Wald distributions and examined how their parameters relate to the experimental manipulation of drift rate v, boundary separation a, and starting point z. We expected that the pattern of association between the two sets of parameters would largely follow the pattern found in our simulations. The empirical results are, however, unlikely to precisely mirror the results of the simulations. Although the effects of the experimental manipulations were adequately accounted for by changes in the above-mentioned diffusion model parameters, the manipulations might not have had *completely* selective influence on these parameters. Hence, changes in the ex–Gaussian and the shifted Wald parameters as a function of the experimental manipulations might reflect slight changes in diffusion model parameters other than the intended ones.

Hierarchical Bayesian Modeling

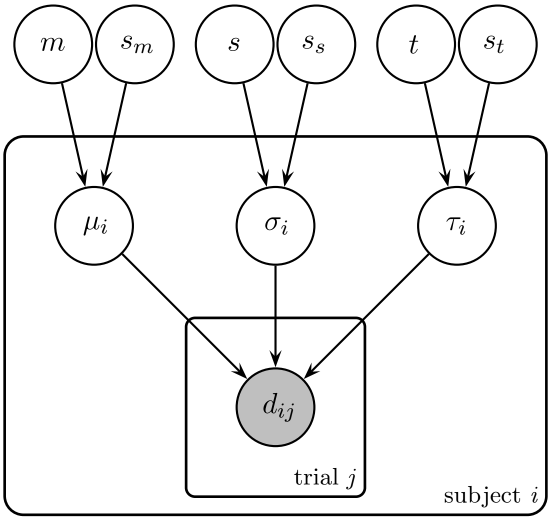
We used hierarchical Bayesian modeling (e.g., Farrell & Ludwig, 2008; Gelman & Hill, 2007; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder, Sun, Speckman, Lu, & Zhou, 2003; Shiffrin, Lee, Kim, & Wagenmakers, 2008) to fit the ex–Gaussian and shifted Wald distributions to the lexical decision data. We used a hierarchical Bayesian approach to fit the descriptive distributions because the individual subject data obtained from the lexical decision tasks were considerably noisier than the synthetic data used in the previous section. As shown by Farrell and Ludwig (2008) and Rouder et al. (2005), hierarchical Bayesian methods reduce the variability in the recovered parameters and produce more accurate individual parameter estimates than single–level maximum likelihood estimation.

The hierarchical Bayesian approach assumes that the parameters of individual participants are drawn from group–level distributions that specify how the individual parameters are distributed in the population. The group–level distributions define thus the between–subject variations of the parameters and can themselves be characterized by a set of parameters. For example, suppose that the RT data of each participant is assumed to come from an ex–Gaussian distribution, but with different values of μ, σ, and τ. The individual subject parameters μ_i, σ_i, and τ_i might in turn be assumed to come from normal distributions with means m, s, and t and with variances s_m^2, s_s^2, and s_t^2, respectively. The benefits of hierarchical modeling arise from using the group–level distributions as priors to adjust extreme individual parameter estimates to more moderate values. In summary, hierarchical Bayesian modeling involves a

> (...) tension between fitting each subject as well as possible (optimal choice of individual parameters) and fitting the group as a whole (... .) This tension results in a movement of the individual parameters toward the group mean, a desirable characteris-

> tic given that we do not desire to overfit the data, and fit the noise in each individual's data. (Shiffrin et al., 2008, p. 1261).

Figure 2.7 and Figure 2.8 show the graphical models for the hierarchical ex–Gaussian and shifted Wald analyses reported in this section. The nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. We use the convention of representing unobserved variables without shading and observed variables with shading (e.g., M. D. Lee, 2008). Figure 2.7 shows that the ex–Gaussian parameters μ_i, σ_i, and τ_i vary from participant to participant and are assumed to be drawn from group–level normal distributions with means m, s, and t, respectively. Similarly, Figure 2.8 shows that the shifted Wald parameters α_i, θ_i, and γ_i vary from participant to participant and are assumed to be drawn from group–level normal distributions with means a, h, and g, respectively.



$$m \sim \text{Uniform}\big(0.25, 0.99\big)$$
$$s_m \sim \text{Uniform}\big(0, 0.214\big)$$
$$s \sim \text{Uniform}\big(0.02, 0.18\big)$$
$$s_s \sim \text{Uniform}\big(0, 0.046\big)$$
$$t \sim \text{Uniform}\big(0.05, 0.85\big)$$
$$s_t \sim \text{Uniform}\big(0, 0.231\big)$$

$$\mu_i \sim \text{Gaussian}\big(m, s_m^2\big)$$
$$\sigma_i \sim \text{Gaussian}\big(s, s_s^2\big)$$
$$\tau_i \sim \text{Gaussian}\big(t, s_t^2\big)$$

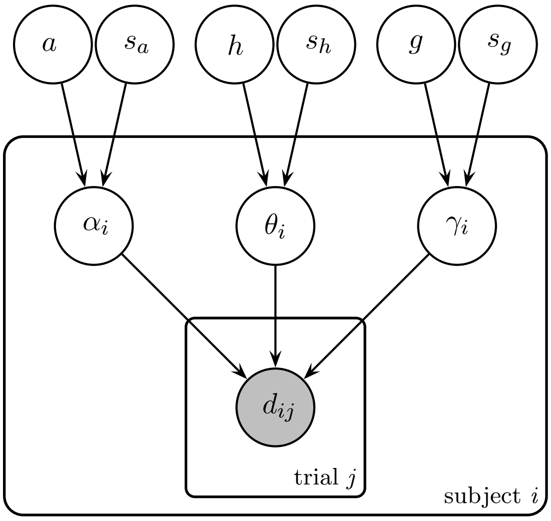
$$d_{ij} \sim \text{Ex} - \text{Gaussian}\big(\mu_i, \sigma_i, \tau_i\big)$$

Figure 2.7 *Graphical model for the hierarchical ex–Gaussian analysis.* Note that the ranges of the uniform prior distributions for the group means are based on the minimum and maximum values of the corresponding ex–Gaussian parameters found in the simulation study reported above.

The ranges of the uniform prior distributions for the group means are based on the minimum and maximum values of the corresponding ex–Gaussian and shifted Wald parameters found in the simulation study reported above. The uniform prior distributions for the group standard deviations range from 0 to the standard deviations of the uniform priors for the corresponding group means. For example, the uniform prior for the ex–Gaussian group mean m parameter ranges from 0.25 to 0.99; values of μ more extreme than this did not occur in our earlier diffusion model simulation study. The uniform prior for the associated group standard deviation s_m ranges from 0 to 0.214. The latter value is the maximum standard deviation for a unimodal distribution on m — that is, the standard deviation for a uniform distribution on m (i.e., $(0.99 - 0.25)/\sqrt{12} \approx 0.214$).

The starting values for the hierarchical Bayesian analysis were based on the individual parameter estimates.[6] At the beginning of each sampling run, the first 1,000 trials of the Markov chain

[6] We later confirmed our results by using overdispersed starting values and multiple chains to obtain an $\hat{R}$ statistic



$$
\begin{aligned}
a &\sim \text{Uniform}\big(0.67, 2.35\big)\\
s_a &\sim \text{Uniform}\big(0, 0.485\big)\\
h &\sim \text{Uniform}\big(0, 0.82\big)\\
s_h &\sim \text{Uniform}\big(0, 0.237\big)\\
g &\sim \text{Uniform}\big(0.85, 7.43\big)\\
s_g &\sim \text{Uniform}\big(0, 1.899\big)
\end{aligned}
$$

$$
\begin{aligned}
\alpha_i &\sim \text{Gaussian}\big(a, s_a^2\big)\\
\theta_i &\sim \text{Gaussian}\big(h, s_h^2\big)\\
\gamma_i &\sim \text{Gaussian}\big(g, s_g^2\big)
\end{aligned}
$$

$$
d_{ij} \sim \text{ShiftedWald}\big(\alpha_i, \theta_i, \gamma_i\big)
$$

Figure 2.8 *Graphical model for the hierarchical shifted Wald analysis.* Note that the ranges of the uniform prior distributions for the group means are based on the minimum and maximum values of the corresponding shifted Wald parameters found in the simulation study reported above.

Monte Carlo (MCMC) chains were discarded. Each analysis was based on 10,000 recorded samples. We used the WinBUGS program (Lunn, Thomas, Best, & Spiegelhalter, 2000) for parameter estimation.[7] Note that the descriptive distributions were fitted only to the RTs of correct responses. Similar to the procedure of Wagenmakers, Ratcliff, et al. (2008), we used only RTs that were slower than 300 ms and faster than 2,500 ms.

Results

Figure 2.9 shows the boxplots of the posterior distributions for the ex–Gaussian group mean parameters m, s, and t. Figure 2.10 shows the boxplots of the posterior distributions for the shifted Wald group mean parameters a, h, and g. Our discussion of the results is based on a visual inspection of the posterior distributions.

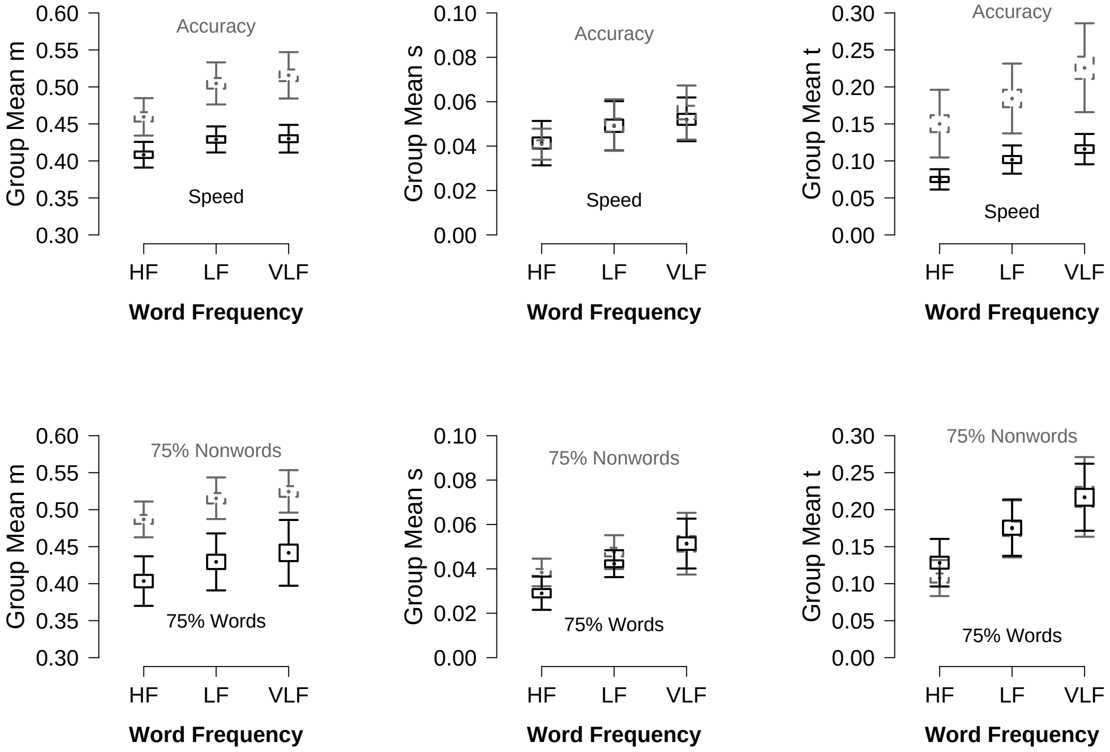
Ex–Gaussian Parameters

With respect to m (i.e., the group mean for the μ parameter), Figure 2.9a shows that m increases when instructions emphasize choice accuracy. Since the effects of the speed–accuracy manipulation can be accounted for by changes in boundary separation a, this result suggests that μ increases with increasing boundary separation. Similarly, m increases when stimuli consist of 75% nonwords. Since the effects of the proportion manipulation can be accounted for by changes in starting point z, this result suggests that μ increases with decreasing starting point. The effect of the word frequency manipulation is less clear—m increases from high-frequency words to low-frequency words, but does

of about 1, indicating that the chains have converged to the stationary distribution (Brooks & Gelman, 1998).

[7] The WinBUGS code and the lexical decision data are available in the supplemental materials.

not change considerably from low-frequency words to very-low-frequency words. Since the effects of the word frequency manipulation can be accounted for by changes in drift rate v, this result suggests that μ increases slightly with decreasing drift rate.



(a) Group Mean m (b) Group Mean s (c) Group Mean t

Figure 2.9 *Boxplots of the posterior distributions for the ex–Gaussian group means m, s, and t, derived separately for each condition of the two lexical decision experiments of Wagenmakers, Ratcliff, et al. (2008).* HF, high-frequency words; LF, low-frequency words; VLF, very-low-frequency words.

With respect to s (i.e., the group mean for the σ parameter), Figure 2.9b shows that s is influenced by the effects of neither the speed–accuracy instructions nor the proportion manipulation. These results suggest that σ is not influenced by either boundary separation a or starting point z. Again, the effect of the word frequency manipulation is less clear—s increases somewhat from high-frequency words to low-frequency words, but does not change considerably from low-frequency words to very-low-frequency words. This finding suggests that σ increases slightly with decreasing drift rate v. Note, however, that the increase in σ is relatively small.

Turning to t (i.e., the group mean for the τ parameter), Figure 2.9c shows that t increases with decreasing word frequency and when instructions emphasize choice accuracy. In contrast, t is unresponsive to the effects of the proportion manipulation. These results suggest that τ increases with increasing boundary separation a and with decreasing drift rate v but is unaffected by changes in starting point z.

In summary, the results above indicate that the ex–Gaussian parameters do not respond selectively to the effects of the word frequency, speed–accuracy, and proportion manipulations. Consistent with the diffusion model simulations reported above, these results suggest that the ex–Gaussian parameters do not correspond uniquely to the drift rate v, boundary separation a and starting point z parameters of the diffusion model. The μ parameter is sensitive to changes in all three diffusion model parameters. Although σ seems to be influenced only by drift rate v, this influence is relatively small. Finally, τ is responsive to changes in both drift rate v and boundary separation a. These results indicate that changes in the two most important ex–Gaussian parameters, μ and τ, can reflect changes in a variety of diffusion model parameters.

Shifted Wald Parameters

With respect to a (i.e., the group mean for the α parameter), Figure 2.10a shows that a *increases* when instructions emphasize fast responding. In contrast, a seems to be unresponsive to the effects of the proportion manipulation. These results suggest that α increases with decreasing boundary separation a and is unaffected by changes in starting point z. The effect of word frequency is less clear—a increases somewhat with decreasing word frequency when stimuli consists of 75% words but is relatively constant in the other conditions. This result suggests that, under certain conditions, α increases slightly with decreasing drift rate v.

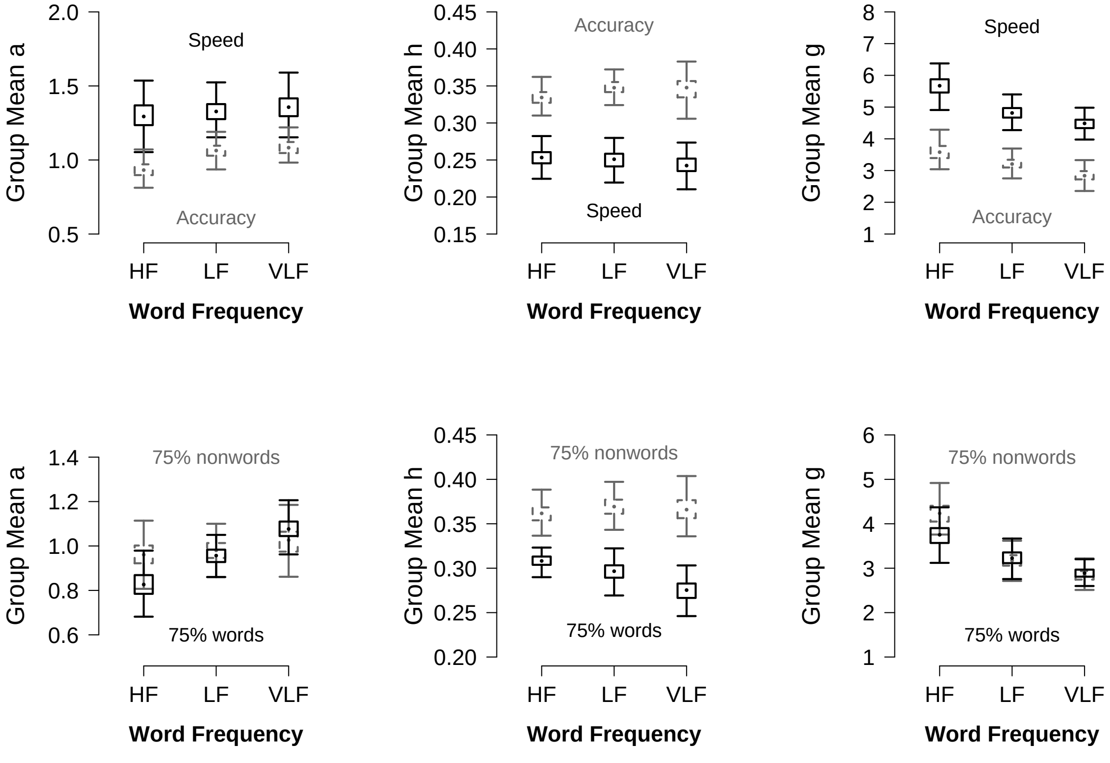
Turning to h (i.e., the group mean for the θ parameter), Figure 2.10b shows that h increases when instructions emphasize choice accuracy and when stimuli consist of 75% nonwords. These results suggest that θ increases with increasing boundary separation a and with decreasing starting point z. Again, the effect of the word frequency manipulation is less clear—h seems to decrease with decreasing word frequency when stimuli consist of 75% words but is relatively constant in the other conditions. This result suggests that, under certain conditions, θ decreases with decreasing drift rate v.

With respect to g (i.e., the group mean for the γ parameter), Figure 2.10c shows that g increases when instructions emphasize fast responding and decreases with decreasing word frequency. In contrast, it seems unresponsive to the effects of the proportion manipulation. These results suggest that γ increases with decreasing boundary separation a, decreases with decreasing drift rate v, and is unaffected by changes in starting point z.

To summarize, the above results indicate that the shifted Wald parameters also do not respond selectively to the effects of the word frequency, speed–accuracy, and proportion manipulations. Consistent with the diffusion model simulations reported above, these results indicate that the shifted Wald parameters do not correspond uniquely to the drift rate v, boundary separation a, and starting point z parameters of the diffusion model. The α and the γ parameters are responsive to changes in both boundary separation a and drift rate v, and θ is influenced by all three diffusion model parameters. These results indicate that changes in the shifted Wald parameter can reflect changes in a diversity of diffusion model parameters.

2.7 Discussion

The goal of this study was to examine the extent to which the ex–Gaussian and shifted Wald parameters could be associated with the kind of psychological processes that are hypothesized by the diffusion model, one of the most successful process models of speeded two–choice decision making. First, we generated synthetic data by systematically manipulating the parameters of the diffusion model, and examined the associated changes in the parameters of the ex–Gaussian



(a) Group Mean a (b) Group Mean h (c) Group Mean g

Figure 2.10 *Boxplots of the posterior distributions for the shifted Wald group means a, h, and g, derived separately for each condition of the two lexical decision experiments of Wagenmakers, Ratcliff, et al. (2008).* HF, high-frequency words; LF, low-frequency words; VLF, very-low-frequency words.

and shifted Wald distributions. Second, we investigated empirical data and studied how the ex–Gaussian and shifted Wald parameters relate to the experimental manipulation of the diffusion model processes drift rate v, boundary separation a, and starting point z. The results were clear–cut: In the context of a two–choice task, the ex–Gaussian and shifted Wald parameters cannot be associated uniquely with the parameters of the diffusion model.

The Ex–Gaussian Distribution

Similar to the results of Ratcliff (1978), our results demonstrated that the two most important ex–Gaussian parameters, μ and τ, were sensitive to changes in a variety of diffusion model parameters. Specifically, μ was influenced by boundary separation a, starting point z, and nondecision time T_{er}. The τ parameter was sensitive to changes in both drift rate v and boundary separation a. The results related to the experimental manipulation of the diffusion model parameters followed a similar pattern. The only discrepancy was that μ also appeared to be influenced by the effects of the word frequency manipulation, suggesting that it was also sensitive to changes in drift rate

v. This difference notwithstanding, the results indicate that the ex–Gaussian parameters do not correspond uniquely to those of the diffusion model.

The Shifted Wald Distribution

The results of the simulations indicated that the shifted Wald parameters also could not be uniquely associated with parameters of the diffusion model. Each of the shifted Wald parameters appeared to be sensitive to changes in a diversity of diffusion model parameters. The α parameter was sensitive to changes in drift rate v, boundary separation a, and starting point z. Surprisingly, α decreased as boundary separation a increased. The θ parameter was affected by both boundary separation a and nondecision time T_{er}. Finally, the γ parameter was substantially influenced by both drift rate v and boundary separation a. The results related to the experimental manipulation of the diffusion model parameters largely followed the same pattern. However, α was unresponsive to the effects of the proportion manipulation, suggesting that this parameter was unaffected by changes in starting point z. Also, α seemed to increase, rather than decrease, with decreasing word frequency, suggesting that it increased with decreasing drift rate v. Finally, θ was responsive to the effects of the word frequency and proportion manipulations, suggesting that it was also sensitive to changes in v and z. These differences notwithstanding, the results indicate that the shifted Wald parameters do not correspond uniquely to parameters of the diffusion model.

The finding that neither the simulation nor the experimental results support the interpretation of the shifted Wald parameters in terms of the psychological processes of subject ability/task difficulty, response caution, and nondecision time is disappointing and comes somewhat as a surprise. First, in view of the conceptual similarities of the shifted Wald distribution and the diffusion model, one might expect some correspondence between the two sets of parameters and their underlying cognitive processes. Yet none of our predictions derived from the theoretical similarities of the two models was supported by the results. Second, our results indicate that the differential sensitivity of the shifted Wald parameters found in the go/no–go task (Heathcote, 2004) does not generalize to tasks that involve two response boundaries. Our results strongly suggest that when the shifted Wald is applied to paradigms that involve more than a single response boundary, the cognitive interpretation of the shifted Wald parameters no longer holds. It must further be noted that the results of Gomez, Ratcliff, and Perea (2007) suggest that the cognitive interpretation of the shifted Wald parameters might be problematic even when the distribution is applied to one–choice tasks. In particular, Gomez et al. showed that an adequate model of the go/no–go task must feature two response boundaries: one associated with the *go* response and another associated with the implicit choice not to respond (i.e., *no–go* decision).

The Effects of Error Rate, Parameter Correlations, and Parameter Combinations

Our results strongly suggest that the ex–Gaussian and shifted Wald parameters should not be interpreted in terms of the cognitive processes assumed by the diffusion model. Nevertheless, some issues warrant further discussion.

First, our method of data generation resulted in error rates ranging from 10% to 15% across the simulations. However, the shifted Wald distribution may perhaps be appropriate for paradigms that result in very few errors, such as tasks that involve saccadic eye movements (Carpenter & Williams, 1995). We therefore investigated how the ex–Gaussian and shifted Wald parameters change as a function of the manipulation of the diffusion model parameters in data sets with lower (0.9% – 4%), as well as higher (19% – 28%) error rates. Regardless of whether the error rate was low or high, the results clearly indicated that the ex–Gaussian and shifted Wald parameters cannot be associated uniquely with the parameters of the diffusion model.[8]

Second, we generated data by manipulating each diffusion model parameter separately while holding the other parameters constant on their average values. Although this approach yields clear–cut and comprehensible results, it ignores the possible associations among the diffusion model parameters. We therefore investigated how changes in the ex–Gaussian and shifted Wald parameters relate to changes in the diffusion model parameters when we take into account the correlations between the latter parameters. The simulations indicated that using parameter sets with realistic parameter associations yields results that are noisier but qualitatively similar to those reported in the present article.

Finally, our results indicate that the individual ex-Gaussian and shifted Wald parameters cannot be mapped uniquely onto the parameters of the diffusion model. However, the parameters of the descriptive distributions need not be considered in isolation. Unlike the individual parameters, certain (nonlinear) combinations of the ex-Gaussian or shifted Wald parameters might map uniquely onto parameters of the diffusion model. This possibility awaits further investigation.

A Common Problem?

Neither the ex–Gaussian nor the shifted Wald parameters appear to correspond to the psychological processes hypothesized by the diffusion model. A possible reason for this unfortunate result may be that neither of the two distributions take into account response accuracy. Without any knowledge of response accuracy, it is very difficult to distinguish between effects of task difficulty (or subject ability) and effects of response caution. For example, does an decrease in RT come about because of an increase in drift rate v or a decrease in boundary separation a? It is evident that in this case, a change in response accuracy is highly diagnostic; an increase in drift rate leads to fewer errors (i.e., an overall improvement), whereas a decrease in boundary separation leads to more errors (i.e., the speed–accuracy trade–off; e.g., Schouten & Bekker, 1967; Wickelgren, 1977). Because performance in RT tasks reflects the combined effects of task difficulty and response caution, a model that cannot separate these influences is unlikely to capture the cognitive processes that determine performance (Wagenmakers, van der Maas, & Grasman, 2007).

Conclusion

The present results indicate that—in the context of speeded two–alternative tasks—the ex–Gaussian and shifted Wald parameters should not be interpreted in terms of the cognitive processes hypothesized by the diffusion model. This does not imply that the ex–Gaussian and shifted Wald distributions should no longer be used as purely descriptive tools to economically summarize RT data and to constrain model development. Such descriptive use of the ex–Gaussian and shifted Wald distributions is perfectly legitimate and highly encouraged. What our findings do imply is that it may be ill–advised to attribute changes in the ex–Gaussian and shifted Wald parameters to changes in specific components of cognitive processing.

[8] The results of these additional simulations are available in the supplemental materials.

Chapter 3

Bayesian Parametric Estimation of Stop-Signal Reaction Time Distributions

This chapter has been published as:
Dora Matzke, Conor V. Dolan, Gordon D. Logan, Scott D. Brown. and Eric-Jan Wagenmakers (2013).
Bayesian parametric estimation of stop-signal reaction time distributions.
Journal of Experimental Psychology: General, 142, 1047-1073, doi: 10.1037/a0030543.[1]

Abstract

The cognitive concept of response inhibition can be measured with the stop-signal paradigm. In this paradigm, participants perform a two-choice response time (RT) task where, on some of the trials, the primary task is interrupted by a stop signal that prompts participants to withhold their response. The dependent variable of interest is the latency of the unobservable stop response (stop-signal reaction time or SSRT). Based on the horse race model (Logan & Cowan, 1984), several methods have been developed to estimate SSRTs. None of these approaches allow for the accurate estimation of the entire distribution of SSRTs. Here we introduce a Bayesian parametric approach that addresses this limitation. Our method is based on the assumptions of the horse race model and rests on the concept of censored distributions. We treat response inhibition as a censoring mechanism, where the distribution of RTs on the primary task (go RTs) is censored by the distribution of SSRTs. The method assumes that go RTs and SSRTs are ex-Gaussian distributed and uses Markov chain Monte Carlo sampling to obtain posterior distributions for the model parameters. The method can be applied to individual as well as hierarchical data structures. We present the results of a number of parameter recovery and robustness studies and apply our approach to published data from a stop-signal experiment.

[1] This chapter may not exactly replicate the final version published in the *Journal of Experimental Psychology: General*. It is not the copy of record.

3.1 Introduction

The stop-signal task (Lappin & Eriksen, 1966; Logan & Cowan, 1984) is frequently used to investigate response inhibition. Response inhibition refers to the ability to stop an ongoing action that is no longer appropriate: for example, driving your car and rapidly hitting the break when you notice that the traffic light turned red. The stop-signal paradigm can be used to investigate the operation of such simple type of inhibitory control in a carefully controlled laboratory setting.

In the standard stop-signal paradigm, participants perform a two-choice response time (RT) task, such as responding to the orientation of the visually presented stimuli. On some of the trials, this primary task is interrupted by an auditory stop signal that prompts participants to withhold their response on that trial. One of the primary dependent variables is the time required to inhibit the ongoing response (stop-signal RT [SSRT]). However, unlike the latency of the overt primary response, SSRTs cannot be observed directly.

To formally account for performance in the stop-signal paradigm, Logan (1981) and Logan and Cowan (1984) proposed that response inhibition can be viewed as a horse race between two competing processes: a go process that is set into motion by the primary task and a stop process that is initiated by the stop signal. If the go process wins the race, the primary response is executed; if the stop process wins the race, the primary response is successfully inhibited.

Since its development, the horse race model (Logan, 1981; Logan & Cowan, 1984) has successfully accounted for stop signal data in various settings and has facilitated the interpretation of numerous stopping experiments. For instance, the stop signal task has been used extensively to investigate response inhibition in different age groups (e.g., Kramer, Humphrey, Larish, Logan, & Strayer, 1994; Ridderinkhof, Band, & Logan, 1999; Schachar & Logan, 1990; Williams, Ponesse, Schachar, Logan, & Tannock, 1999) and clinical populations, such as children with Attention Deficit Hyperactivity Disorder (ADHD; Oosterlaan, Logan, & Sergeant, 1998; Schachar & Logan, 1990; Schachar, Mota, Logan, Tannock, & Klim, 2000).

The horse race model owes its popularity to the ability to quantify the otherwise unobservable latency of stopping. Various methods are available to estimate SSRTs. The standard analysis methods for the horse race model only yield a summary measure of the latency of inhibition, such as the mean SSRT; they do not reveal the shape of the entire SSRT distribution. It is well known that important features of the data may be missed in focusing only on the mean (e.g., Heathcote et al., 1991). A growing number of researchers therefore rely on distributional models, like the ex-Gaussian distribution (e.g., Balota & Yap, 2011; Matzke & Wagenmakers, 2009) to estimate the shape of entire RT distributions. For instance, Leth-Steensen et al. (2000) reported that children with ADHD differed from age-matched controls only in the ex-Gaussian parameter that quantifies the tail (i.e., very long RTs) of the RT distribution. Similarly, the RT distribution of schizophrenia patients is more variable and follows a markedly different shape than the RT distribution of controls, without necessarily differing in the mean (Belin & Rubin, 1995).

In the context of the stop-signal paradigm, focusing only on mean SSRT may likewise mask crucial features of the data and result in erroneous conclusions about the nature of response inhibition. Consider, for instance, the two SSRT distributions shown in Figure 3.1. The distributions have the same mean, but have clearly different shapes. The distribution drawn in solid line is more peaked, whereas the distribution drawn in dashed line is more spread out, with a faster leading edge and a longer tail. Ignoring such differences in the shapes of SSRT distributions may lead to the incorrect conclusion that two clinical groups or experimental conditions do not differ in SSRT. Unfortunately, the existing methods for obtaining SSRT estimates do not enable researchers to accurately estimate and evaluate differences in the shape of SSRT distributions.

The goal of this article therefore is to introduce a method that allows for the estimation of

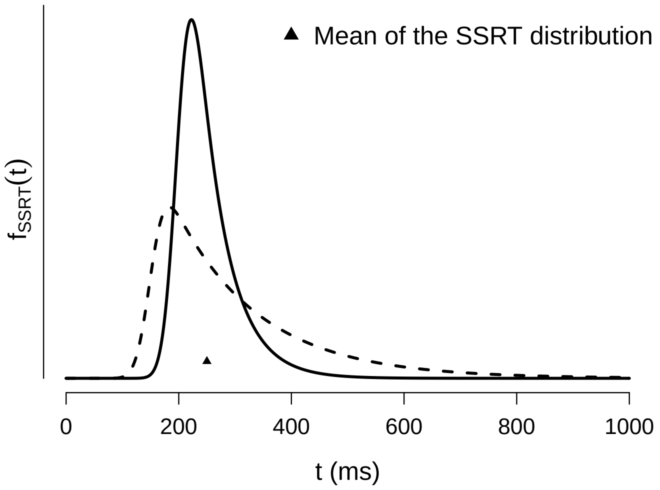


Figure 3.1 *Examples of stop-signal reaction time (SSRT) distributions with synthetic data.* The solid line shows an SSRT distribution with a slow leading edge and a short tail. The dashed line shows an SSRT distribution with a fast leading edge and a long tail. Despite the differences in their shapes, the means (i.e., black triangle) of the two distributions are equal.

the entire distribution of SSRTs, such as those shown in Figure 3.1. Our approach is based on the assumptions of the horse race model. The new method rests on the concept of censored distributions, where response inhibition is treated as a mechanism for censoring observed RTs. In order to quantify the shape of the distributions, the method assumes that the go RTs and SSRTs follow a parametric form, namely an ex-Gaussian distribution. Note, however, that our method does not hinge on this choice of parametric form; almost any other choice of distribution would do just as well. The ex-Gaussian distribution is purely used as a convenient choice to summarize the go RTs and the SSRTs. The ex-Gaussian is a commonly used distributional model, and it typically produces excellent fit to empirical RT distributions. (Heathcote et al., 1991; Hockley, 1982, 1984; Ratcliff, 1978, 1993; Ratcliff & Murdock, 1976). Our approach relies on Markov chain Monte Carlo (MCMC) sampling (Gamerman & Lopes, 2006; Gilks et al., 1996) and calculates posterior distributions for the model parameters.

An important advance of our Bayesian parametric method is that it makes it relatively easy to conduct both individual and hierarchical analyses. In individual analysis, the parameters of the SSRT distribution are estimated separately for each participant. In contrast, the hierarchical analysis (e.g., Gelman & Hill, 2007) recognizes that participants share some similarities and uses information available from the entire group to improve parameter estimation for the individual participants. The hierarchical approach has the potential to provide accurate parameter estimates with relatively few observations. Hierarchical modeling is therefore especially valuable in developmental and clinical stop-signal studies that typically use a very small number of trials per participant.

The outline of the article is as follows. In the first section, we describe the stop-signal paradigm in more detail and discuss existing methods for estimating SSRTs. In the second section, we introduce the individual and the hierarchical Bayesian parametric approach to the estimation of SSRT distributions. In the third section, we report the results of various parameter recovery studies and show that our method accurately recovered the parameters of the generating SSRT distributions. In the fourth section, we apply the Bayesian parametric approach to an existing stop-signal data set. The fifth section concludes our investigation.

3.2 The Stop-Signal Paradigm

In the standard stop-signal paradigm (Lappin & Eriksen, 1966; Logan & Cowan, 1984), participants perform a two-choice RT task (i.e., the go task), such as responding to the orientation of the visually presented stimuli (e.g., press the right button for a right-pointing arrow and press the left button for a left-pointing arrow). Occasionally, the go stimulus is followed by an auditory stop signal (e.g., a high-pitched tone) that prompts participants to withhold their response on that trial. Typically, the stop signal is presented on a random 25-30% of the trials. The probability of successful inhibition can be experimentally manipulated by varying the time interval between the onset of the go stimulus and the onset of the stop signal (i.e., stop-signal delay [SSD]). The shorter the SSD, the more likely participants are to inhibit their response to the go stimulus.

To facilitate the interpretation of stop-signal data, Logan (1981) and Logan and Cowan (1984) introduced the horse race model. The horse race model conceptualizes response inhibition as a horse race between a go and a stop process. If the go process finishes before the stop process, the response is an error of commission. If the stop process finishes before the go process, the response is successfully inhibited. According to the horse race model, response inhibition is thus determined by the relative finishing times of the go and the stop process. Figure 3.2 illustrates how the probability of responding to the go stimulus (i.e., gray area) and the probability of inhibiting the response to the go stimulus (i.e., white area) are determined by the SSD, the SSRT, and the go RT distribution. Go RTs that are longer than $SSD + SSRT$ are successfully inhibited. In contrast, go RTs that are shorter than $SSD + SSRT$ cannot be inhibited and result in signal-respond RTs.

The standard horse race model depicted in Figure 3.2 assumes that, conditional on SSD, SSRT is constant (Logan & Cowan, 1984). This assumption is implausible, as SSRTs are certainly variable. Also, estimated SSRTs tend to decrease as SSD increases, a common finding that is explained in terms of the variability in SSRT. At short SSDs, almost all SSRTs are fast enough to win the race against the go RTs. The estimated mean SSRT therefore closely approximates the mean of the entire SSRT distribution. At long SSDs, only very fast SSRTs can win the race against the go RTs. The estimated mean SSRT is therefore lower than the mean of the entire SSRT distribution. As a result, SSRT estimates are longer at short SSDs than at long SSDs (de Jong, Coles, Logan, & Gratton, 1990; Logan & Burkell, 1986; Logan & Cowan, 1984).

To account for variability in SSRT, Logan and Cowan (1984) introduced the complete version of the horse race model. The complete race model treats both go RTs and SSRTs as independent random variables. To formalize the model, Logan and Cowan made the following simplifying assumptions about the independence of the go and the stop process. According to the context independence assumption, the distribution of go RTs is the same for go trials and for stop-signal trials. According to the stochastic independence assumption, the finishing times of the go and the stop process are uncorrelated. These two independence assumptions allow one to treat the go RT distribution on go trials as the underlying distribution of go RTs on stop-signal trials.

The formulation of the complete race model is closely connected to the concept of inhibition functions: functions that describe the relationship between the $P(\text{respond} \mid \text{stop signal})$ and SSD.

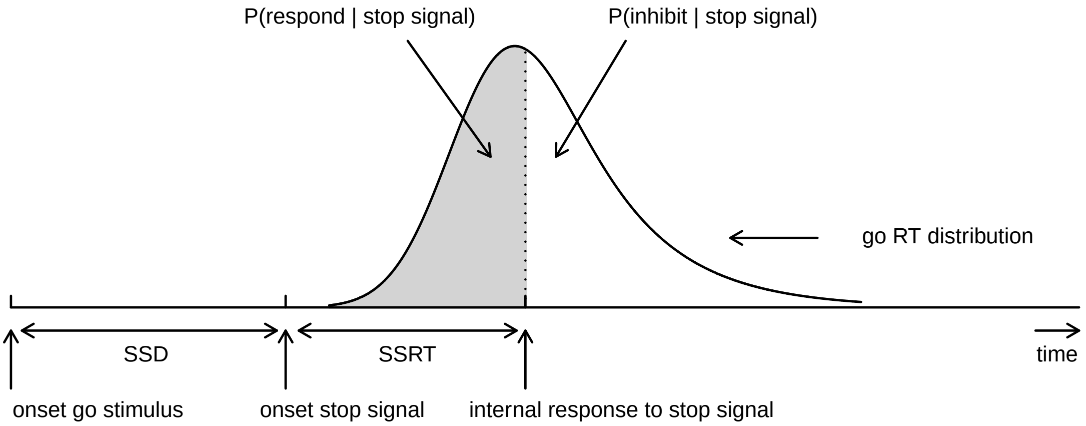


Figure 3.2 *Graphical representation of the horse race model.* RT = response time; SSD = stop-signal delay; SSRT = stop-signal reaction time.

As shown in Figure 3.3, the P(respond | stop signal) typically increases with increasing SSD. Logan and Cowan (1984) treated the inhibition function as a cumulative distribution and showed that its mean equals the difference between the mean go RT and the mean SSRT:

$$E(\mathrm{I}) = \mathrm{E(goRT)} - \mathrm{E(SSRT)}. \tag{3.1}$$

Further, they showed that the variance of the inhibition function equals the sum of the variances of the go RTs and the SSRTs:

$$\mathrm{Var(I)} = \mathrm{Var(goRT)} + \mathrm{Var(SSRT)}. \tag{3.2}$$

Note that the derivation of the complete horse race model is not based on any specific distribution shapes for the go RT and SSRT distributions.

Estimating SSRTs

One of the major advantages of the horse race model is that it allows for the estimation of the otherwise unobservable SSRT. Various methods are available for estimating SSRTs. The choice of method depends on the way SSDs are set in a particular experiment.

The SSD can be set according to the fixed-SSDs procedure or according to the staircase tracking procedure (e.g., Logan, 1994). The fixed-SSDs procedure requires a number of a priori chosen delays to be presented to the participants (e.g., SSDs of 80, 160, 240, 320, 400, and 480 ms; Logan & Burkell, 1986). Stop signals at the different SSDs are presented with equal frequencies at a random order. The challenge is to find a set of SSDs that span the entire range of the inhibition function. For the fixed-SSDs procedure, the integration method (Logan, 1981; Logan & Cowan, 1984) is the most popular approach to estimate SSRTs. The integration method assumes that SSRT is constant. SSRTs are estimated from the observed go RT distribution and the P(respond | stop

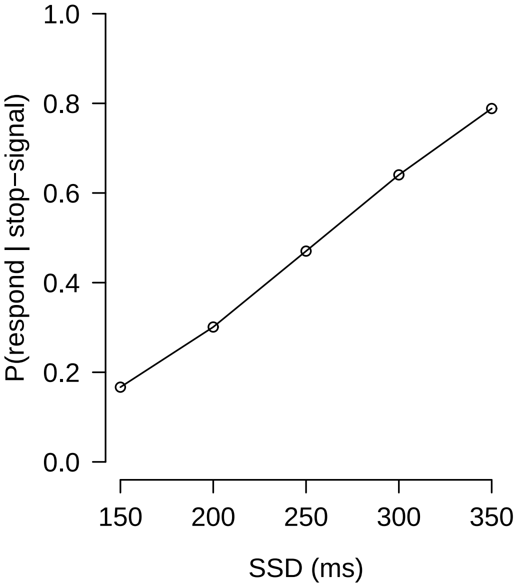


Figure 3.3 *Example of an inhibition function based on synthetic data from the recovery study for the individual Bayesian parametric approach.* The figure shows how the probability of responding on a stop-signal trial increases with increasing stop-signal delay (SSD).

signal) by finding the point (i.e., $SSRT + SSD$) at which the integral of the go RT distribution equals the P(respond | stop signal),

$$P(\text{respond} \mid \text{stop signal}) = \int_{-\infty}^{\text{SSRT+SSD}} f_{go}(t)\text{dt}. \tag{3.3}$$

In terms of Figure 3.2, the integration method involves deriving the time point at which the internal response to the stop signal occurs and subtracting SSD to obtain the SSRT. In practice, the following procedure is used: Go RTs are collapsed into a single distribution and are rank ordered. Subsequently, the nth go RT is selected, where n is obtained by multiplying the number of go RTs by the P(respond | stop signal) at a given SSD. Lastly, the SSD is subtracted to arrive at the SSRT. The integration method yields SSRT estimates for each SSD. As estimated SSRTs tend to decrease with increasing SSD (Logan & Burkell, 1986; Logan & Cowan, 1984), SSRTs at different SSDs are often averaged to yield a summary score for each participant.

The integration method has several drawbacks. It assumes that SSRT is constant, an assumption that is certainly incorrect. Moreover, the integration method requires a relatively large number of observations to produce accurate estimates of average SSRT. Researchers are advised to present participants with at least 900 go trials and 60 stop-signal trials on each of five different SSDs (Band, van der Molen, & Logan, 2003).

The second method for presenting SSDs, the staircase tracking procedure, sets SSDs dynamically, contingent on participants' performance. A typical staircase procedure will increase SSD by, say, 50 ms after successful inhibition, and decrease SSD by 50 ms after unsuccessful inhibition (see e.g., Bissett & Logan, 2011; Logan, Schachar, & Tannock, 1997; Osman, Kornblum, & Meyer, 1986; Verbruggen, Logan, & Stevens, 2008). This tracking procedure results in an overall $P(\text{respond} \mid \text{stop signal})$ of 0.50 for each participant.

For the staircase tracking procedure, the mean method is the easiest approach to estimate SSRTs. The mean method originates from Logan and Cowan's (1984) treatment of SSRT as a random variable and is based on the following relationship:

$$\text{E(SSRT)} = \text{E(goRT)} - \text{E(I)}. \tag{3.4}$$

Mean SSRT is thus given by the difference between the mean go RT and the mean of the inhibition function. Several approaches are available to compute the mean of the inhibition function (see, e.g., Logan, 1994; Logan & Cowan, 1984). The simplest way is to exploit the fact that when the staircase tracking procedure yields an overall $P(\text{respond} \mid \text{stop signal})$ of 0.50, the mean of the inhibition function equals the mean of the SSDs. As shown in Equation 3.4, the mean SSRT can be obtained by subtracting the mean SSD form the mean of the go RTs (Logan & Cowan, 1984; Logan et al., 1997).

The mean method can be used with a relatively small amount of data. Stop-signal experiments with healthy young adults typically include a total of 500-1,000 trials. Developmental and clinical studies generally include 250-500 trials, but investigations with as few as 100-250 trials are also common. Note, however, that contrary to the integration method, the mean method cannot be used to calculate SSRTs for each SSD separately.

Several variants of the integration and the mean method are available for the fixed-SSDs as well as the staircase tracking procedure (for a summary, see Verbruggen & Logan, 2009). Band et al. (2003) used simulations to show that SSRT estimates for which the $P(\text{respond} \mid \text{stop signal})$ equals 0.50, such as the mean method, are the most reliable. The mean method therefore has become the dominant method for estimating SSRTs.

Estimating Variability in SSRT

Logan and Cowan's (1984) treatment of SSRT as a random variable provides a method for estimating the variability in SSRT. Logan and Cowan showed that the variance of the inhibition function can be calculated from its slope at the median. Once the variance of the inhibition function is known, the variance of SSRTs can be obtained from Equation 3.2. Logan and Cowan's method is based on the observation that in a symmetrical distribution, the variance is proportional to the slope of the cumulative distribution at the median. If we treat the inhibition function as a cumulative distribution and assume a particular parametric form, say normal, the slope of the inhibition function at the median is given by

$$B_{0.5} = \frac{1}{\sqrt{2\pi} \times \text{SD(I)}}. \tag{3.5}$$

It then follows from Equation 3.2 that the variance of SSRTs can be obtained by

$$\text{Var(SSRT)} = \left(\frac{1}{B_{0.5}\sqrt{2\pi}}\right)^2 - \text{Var(goRT)}. \tag{3.6}$$

In contrast to the generality of the horse race model, the Logan and Cowan method for estimating SSRT variability assumes a particular parametric form of the inhibition function. Most importantly, Band et al. (2003) showed with simulations that the Logan and Cowan method overestimates the true variability in SSRT.

An Existing Method for Estimating SSRT Distributions

Up to now, the only existing approach for estimating the entire distribution of SSRTs was developed by Colonius (1990, see also de Jong et al., 1990, p. 181). Colonius showed that the survival distribution of SSRTs can be recovered using the distribution of go RTs, the distribution of signal-respond RTs, and the $P(\text{respond} \mid \text{stop signal})$ at a given SSD. Formally,

$$P(\text{SSRT+SSD} > t \mid \text{SSD}) = P(\text{respond} \mid \text{stop signal, SSD}) \times \frac{f_{SR}(t \mid \text{SSD})}{f_{go}(\text{t})}, \tag{3.7}$$

where $f_{go}(t)$ and $f_{SR}(t|SSD)$ are the probability density functions of the go RTs and the signal-respond RTs, respectively. Colonius' method does not depend on the specific parameterization of the go RT and the signal- respond RT distributions. The densities $f_{go}(t)$ and $f_{SR}(t|SSD)$ can be estimated with various nonparametric density estimation methods (e.g., Silverman, 1986). Once the survival distribution of SSRTs is obtained, measures of location (e.g., median) and dispersion (e.g., interquartile distance) can be calculated easily.

Although Colonius' (1990) method is straightforward and elegant, it requires a very large number of observations to perform adequately (Logan, 1994). Band et al. (2003) used simulations to show that the Colonius method underestimates SSRT and overestimates its variability. In our implementation, over 250,000 stop-signal trials per SSD were required to obtain relatively accurate estimates of SSRT distributions. Using a more realistic number of stop-signal trials (e.g., 200 per SSD) resulted in inaccurate estimates, especially in the tails of the SSRT distribution. These problems are typical of nonparametric methods that estimate distribution tails from data (Luce, 1986).

To summarize, the stop-signal paradigm offers various methods to estimate the otherwise unobservable latency of stopping. Most methods only provide a summary measure of SSRT and are unable to accurately estimate the variability in SSRT. The only existing method for estimating entire SSRT distributions requires an unrealistically large number of observations to produce accurate estimates, particularly in the tail of the SSRT distribution. In what follows, we present a novel approach that relies on a parametric assumption to quantify the shape of the go RT and the SSRT distributions. As a result, the new method can provide accurate estimates of SSRT distributions even with relatively few observations.

3.3 Bayesian Parametric Approach for the Estimation of SSRT Distributions

Here we introduce a novel approach that allows for the estimation of the entire distribution of SSRTs. The method assumes that the go RTs and SSRTs follow an ex-Gaussian distribution. The ex-Gaussian distribution is purely used as a convenient choice to describe the go RTs and SSRTs. The ex-Gaussian is a frequently used distributional model that typically produces excellent fit to empirical RT distributions (Heathcote et al., 1991; Hockley, 1982, 1984; Ratcliff, 1978, 1993; Ratcliff & Murdock, 1976). The new approach may be applied to individual as well as hierarchical data structures and relies on MCMC sampling to obtain estimates of the parameters of the ex-Gaussian SSRT distribution.

We first introduce the rationale behind the Bayesian parametric approach (BPA), with special focus on the ex-Gaussian distribution and the assumptions of the method. We then introduce the basic concepts of Bayesian parameter estimation. Lastly, we present the individual and hierarchical BPA models for estimating SSRT distributions.

Introducing the Bayesian Parametric Approach

Rationale

The BPA rests on the concept of right-censored distributions. In right-censored distributions, observations to the right of a cutoff point (i.e., the censoring point) are omitted, but the number of censored observations is known. Censoring is a type of missing data problem that is frequently encountered in survival analysis (e.g., Elandt-Johnson & Johnson, 1980). In most applications, the censoring point is known and the focus is on estimating the parameters of the censored distribution. For instance, imposed censoring has been considered as a method to accommodate outliers in estimating the parameters of RT distributions (Ulrich & Miller, 1994).

As shown is Figure 3.2, the estimation of SSRT using the standard horse race-model with constant SSRT can be viewed as a censoring problem. Specifically, the signal-respond RT distribution (i.e., gray area) can be treated as a right-censored go RT distribution with a constant censoring point that is given by the finishing time of the stop process (i.e., $SSD + SSRT$). On a given SSD, go RTs that are shorter than the finishing time of the stop process are observed. In contrast, go RTs that are longer than the finishing time of the stop process are successfully inhibited and therefore cannot be observed. Note that contrary to typical censoring problems, the censoring point of the go RT distribution is unknown. The estimation of SSRT therefore involves estimating the censoring point of the go RT distribution.

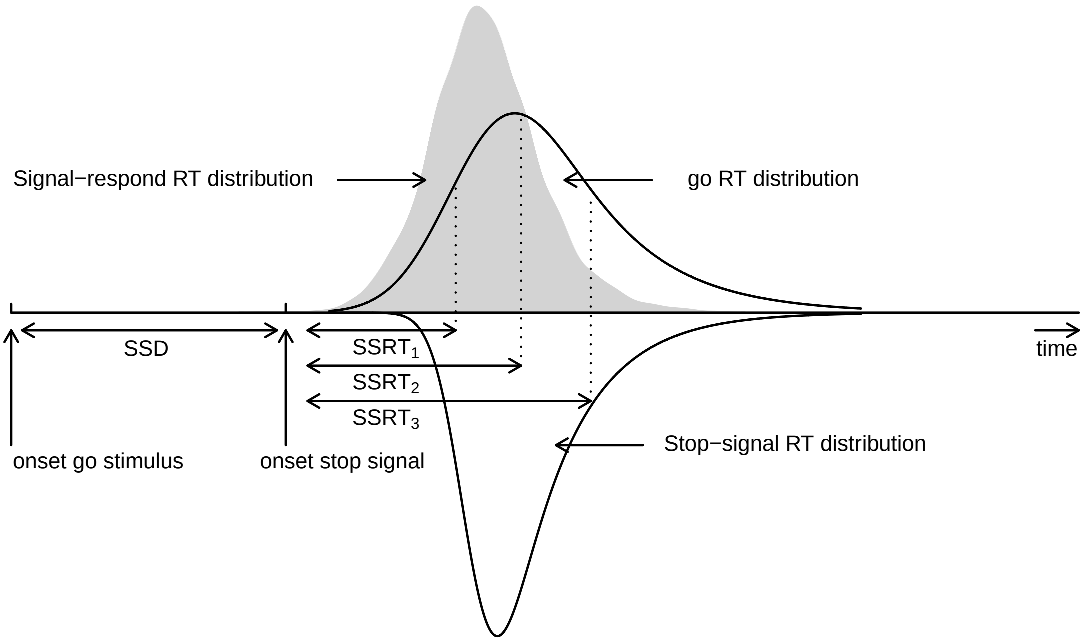


Figure 3.4 *Graphical representation of the complete horse race model.* RT = response time; SSD = stop-signal delay; SSRT = stop-signal reaction time.

The same reasoning can be extended to the estimation of the entire SSRT distribution using the complete horse race model. The censoring problem is, however, complicated by the fact that both go RTs and SSRTs are treated as random variables. As shown in Figure 3.4, the censoring point on a given SSD takes on different values on each stop-signal trial (i.e., $SSD + SSRT_1$, $SDD + SSRT_2$, and $SSD + SSRT_3$). The signal-respond RT distribution (i.e., gray area) can be viewed as a censored go RT distribution with censoring points drawn from the SSRT distribution that is shifted with the SSD on the time axis to longer RTs. The estimation of the SSRT distribution therefore involves estimating the finishing time distribution of the stop process that censors the go RT distribution.

The BPA is a parametric approach and as such involves choosing a parametric form for the go RT and the SSRT distribution. In what follows, we assume that go RTs and SSRTs —and therefore the finishing times of the stop process— are ex-Gaussian distributed and focus on simultaneously estimating the parameters of the two distributions.

Ex-Gaussian Distribution

The BPA assumes that the go RTs and the SSRTs are ex-Gaussian distributed. The ex-Gaussian distribution is given by the convolution of a Gaussian and an exponential distribution. The ex-Gaussian has three parameters. The μ and σ parameters give the mean and the standard deviation of the Gaussian component and reflect the leading edge and mode of the distribution. The τ parameter gives the mean of the exponential component and reflects the tail of the distribution.

The ex-Gaussian distribution has a positively skewed unimodal shape that typically fits empirical RT distributions well (Heathcote et al., 1991; Hockley, 1982, 1984; Ratcliff, 1978, 1993; Ratcliff & Murdock, 1976). Figure 3.5 shows changes in the ex-Gaussian distribution as a result of changes in the μ, σ and τ parameters. Increasing the μ parameter shifts the entire distribution to longer RTs and increases only the mean. Increasing σ influences the shape of the distribution and increases only the variance. Lastly, increasing τ influences both the location and the shape of the distribution and therefore increases both the mean and the variance (see Equation 3.11 and Equation 3.12).

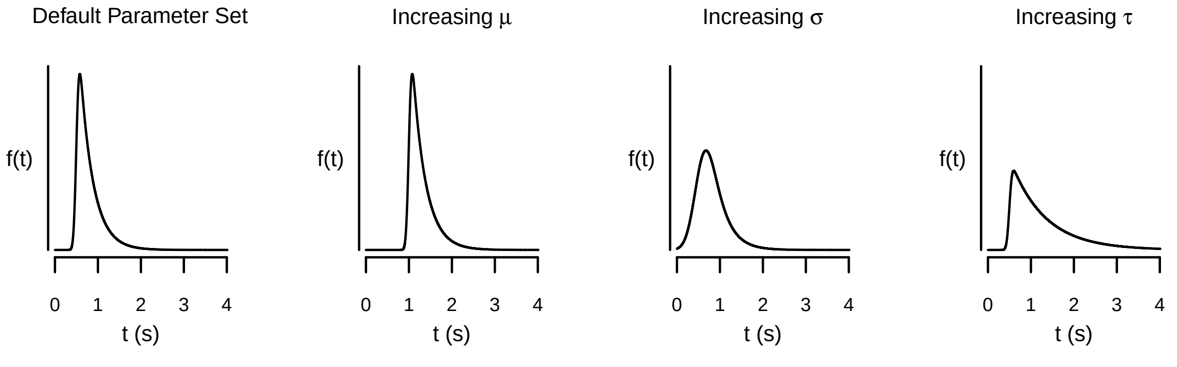


Figure 3.5 *Changes in the shape of the ex-Gaussian distribution as a result of changes in the ex-Gaussian parameters μ, σ, and τ.* The parameter sets used to generate the distributions are $\mu = 0.5$, $\sigma = 0.05$, $\tau = 0.3$ (Panel 1); $\mu = 1$, $\sigma = 0.05$, $\tau = 0.3$ (Panel 2); $\mu = 0.5$, $\sigma = 0.2$, $\tau = 0.3$ (Panel 3); and $\mu = 0.5$, $\sigma = 0.05$, $\tau = 0.8$ (Panel 4).

The probability density function of the ex-Gaussian is

$$f(t;\mu,\sigma,\tau)=\frac{1}{\tau}\exp\left(\frac{\mu-t}{\tau}+\frac{\sigma^2}{2\tau^2}\right)\Phi\left(\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}\right)\text{ for }\sigma>0,\ \tau>0, \tag{3.8}$$

where Φ is the standard normal distribution function, given by

$$\Phi\left(\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}}\exp\left(\frac{-y^2}{2}\right)dy. \tag{3.9}$$

The distribution function of the ex-Gaussian is

$$F(t;\mu,\sigma,\tau)=\Phi\left(\frac{t-\mu}{\sigma}\right)-\exp\left(\frac{\sigma^2}{2\tau^2}-\frac{t-\mu}{\tau}\right)\Phi\left(\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}\right), \tag{3.10}$$

and its mean and variance equal

$$\mathrm{E(t)}=\mu+\tau \tag{3.11}$$

and

$$\mathrm{Var(t)}=\sigma^2+\tau^2, \tag{3.12}$$

respectively. Equation 3.11 and Equation 3.12 show how two SSRT distributions with the same mean or variance may have very different shapes, as illustrated in Figure 3.1.

We use the ex-Gaussian distribution purely as a descriptive tool to summarize the go RT and the SSRT distributions (see also Band et al., 2003; Heathcote et al., 1991; Ratcliff, 1978; Wagenmakers, van der Maas, et al., 2008). We do not assume that changes in the ex-Gaussian parameters map onto changes in specific cognitive processes (Matzke & Wagenmakers, 2009). Nevertheless, the ex-Gaussian can excellently accommodate the shape of RT distributions and is easy to fit to data. Moreover, as will be discussed later, sensitivity analyses indicated that the ex-Gaussian based BPA is robust to misspecification of the parametric form of the go RT and SSRT distributions. Note that other distributional assumptions can easily be made within our method.

Assumptions of the BPA

Similar to the complete horse race model, the BPA assumes that go RTs and SSRTs are independent random variables. The independence of the go and the stop process allows one to treat the go RT distribution on go trials as the underlying distribution of go RTs on stop-signal trials. The BPA assumes that the go RTs and the SSRTs follow ex-Gaussian distributions, with parameters μ_{go}, σ_{go}, and τ_{go} for the go RT distribution, and μ_{stop}, σ_{stop}, and τ_{stop} for the SSRT distribution. The log-likelihood of the $g=1,...,G$ go RTs is given by

$$\ln L(\mu_{go},\sigma_{go},\tau_{go})_{go}=\sum_{g=1}^{G}\ln f_{go}(t_g;\mu_{go},\sigma_{go},\tau_{go}), \tag{3.13}$$

where $f_{go}(t;\mu_{go},\sigma_{go},\tau_{go})$ is the probability density of the ex-Gaussian go RT distribution given in Equation 3.8.

The log-likelihood of the data on the $s=1,...,S$ stop-signal trials on a given SSD consists of the sum of the log-likelihoods of the $r=1,...,R$ signal-respond RTs and the $i=1,...,I$ successful inhibitions. According to the race model, signal-respond RTs are obtained on stop-signal trials where the finishing time of the go process is shorter than the finishing time of the stop process (i.e., go RT $< SSD + SSRT$). The log-likelihood of a given signal-respond RT, t_r, can therefore be computed by (1) evaluating the probability density function of the go RT distribution at t_r and (2) evaluating the probability of obtaining an $SSD + SSRT$ that is longer than t_r with the distribution function of the finishing time distribution of the stop process, that is, the distribution function of the SSRTs shifted with the SSD.

Similar reasoning can be extended to the log-likelihood of the successful inhibitions on signal-inhibit trials. According to the race model, successful inhibitions are obtained on stop-signal trials where the finishing time of the go process is longer than the finishing time of the stop process (i.e., go RT $> SSRT + SSD$). The log-likelihood of a given $SSD + SSRT$, t_i, can be computed by (1) evaluating the probability of obtaining a signal-respond RT that is longer than t_i with the distribution function of the go RT distribution and (2) evaluating the probability density function of the finishing time distribution of the stop process (i.e., SSRT distribution shifted with SSD) at t_i. Note, however, that SSRTs are by definition unobservable. Obtaining the log-likelihood on signal-inhibit trials therefore involves integrating out t_i from the go RT and the stop process finishing time distributions. Formally,

$$\begin{aligned}
&\ln L(\mu_{go}, \sigma_{go}, \tau_{go}, \mu_{stop}, \sigma_{stop}, \tau_{stop})_{stop} = \\
&\quad = \sum_{r=1}^{R} \Big\{ \ln f_{go}(t_r; \mu_{go}, \sigma_{go}, \tau_{go}) + \ln \Big[1 - F_{stop}(t_r; \mu_{stop}, \sigma_{stop}, \tau_{stop}, SSD) \Big] \Big\} \\
&\quad + \sum_{i=1}^{I} \ln \int_{-\infty}^{\infty} \Big[1 - F_{go}(t_i; \mu_{go}, \sigma_{go}, \tau_{go}) \Big] \times f_{stop}(t_i; \mu_{stop}, \sigma_{stop}, \tau_{stop}, SSD) dt, \qquad (3.14)
\end{aligned}$$

where $f_{go}(t; \mu_{go}, \sigma_{go}, \tau_{go})$ and $F_{go}(t; \mu_{go}, \sigma_{go}, \tau_{go})$ are the probability density and the distribution function of the ex-Gaussian go RT distribution given in Equation 3.8 and Equation 3.10, respectively. Similarly, $f_{stop}(t; \mu_{stop}, \sigma_{stop}, \tau_{stop}, SSD)$ and $F_{stop}(t; \mu_{stop}, \sigma_{stop}, \tau_{stop}, SSD)$ are the probability density and the distribution function of the ex-Gaussian finishing time distribution of the stop process, that is, the SSRT distribution shifted with the SSD.

The goal is to simultaneously estimate the μ_{go}, σ_{go}, and τ_{go} parameters of the go RT distribution and the μ_{stop}, σ_{stop}, and τ_{stop} parameters of the SSRT distribution. Parameter estimation may proceed by means of standard maximum likelihood estimation (Dolan, van der Maas, & Molenaar, 2002; Myung, 2003). However, the BPA is intended to handle individual as well as hierarchical data structures. Maximum likelihood estimation can become practically difficult for hierarchical problems, so we chose to use Bayesian parameter estimation instead. This also confers the typical benefits of Bayesian estimation, such as a coherent inferential framework.

Bayesian Parameter Estimation

In Bayesian parameter estimation, we start with a prior probability distribution for the parameter of interest. The prior distribution quantifies the existing knowledge about the parameter. The prior distribution is then updated by the incoming data (i.e., likelihood) to yield a posterior probability distribution under Bayes' rule:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \qquad (3.15)$$

The marginal likelihood is the probability of the observed data and does not involve the parameters of interest. Equation 3.15 can hence be expressed as

$$\text{posterior} \propto \text{likelihood} \times \text{prior}. \tag{3.16}$$

The top panels of Figure 3.6 illustrate the basic concepts of Bayesian estimation for the parameters of the SSRT distribution with simulated data from a synthetic participant. For each parameter, we start with a uniform prior distribution reflecting the assumption that all values of the parameter within some wide range are equally likely a priori. The prior distributions are then updated by the data to yield the posterior distributions. The posterior distributions quantify all the available information about the parameters. The central tendency of the posterior distribution can be expressed by its mean, median or mode. The central tendency of the posterior is often used as a point estimate of the parameter (e.g., with a uniform prior, the mode corresponds to the maximum-likelihood estimator). The dispersion of the posterior distribution can be quantified by the standard deviation or the percentiles. The dispersion of the posterior conveys important information about the precision of the parameter estimates: The larger the posterior standard deviation, the greater the uncertainty of the estimated parameter.

In many applications, the posterior distribution cannot be derived analytically. Fortunately, the posterior can be approximated using numerical sampling techniques such as MCMC sampling (Gamerman & Lopes, 2006; Gilks et al., 1996). The BPA currently relies on WinBUGS (Bayesian inference Using Gibbs Sampling for Windows; Lunn et al., 2000; see Kruschke, 2010b for an introduction) to obtain the posterior distributions of the model parameters. WinBUGS is a general-purpose statistical software for Bayesian analysis that uses MCMC techniques to sample from the posterior distribution of the model parameters.

Figure 3.6 gives a simple illustration of Bayesian parameter estimation with MCMC sampling. The bottom panels show sequences of values (i.e., MCMC chains) sampled from the posterior distribution of the parameters of the SSRT distribution. More accurate sampling from the posterior distribution can be obtained by running multiple chains and discarding the beginning of each chain as burn-in. For each parameter, we ran three chains, each with different starting values (i.e., overdispersed starting values). The starting values were randomly generated from uniform distributions covering a wide range of possible parameter values. Per chain, we collected $2,000$ iterations, resulting in a total of $6,000$ samples from the posterior distributions. The chains converged successfully from the starting values to their stationary distributions; the individual chains look like "hairy caterpillars" and they seem identical to one another. Formal diagnostic measures of convergence are available. For instance, $\hat{R}$ (Gelman & Rubin, 1992) compares the between-chain variability to the within-chain variability. As a rule of thumb, $\hat{R}$ should be lower than 1.1 if the chains have properly converged. For the present example, $\hat{R}$ was lower than 1.05 for all of the parameters.

The top panels of Figure 3.6 show histograms and density estimates of the posterior samples of the stop parameters. The histograms were plotted by collecting the sampled values across the three chains and projecting them onto the x-axis of the top panel figures. The median of the posterior distribution equals 186.80 for μ_{stop}, 32.76 for σ_{stop}, and 57.43 for τ_{stop}. The region extending from the 2.5^{th} to the 97.5^{th} percentile of the posterior distribution gives the so-called 95% Bayesian confidence interval. For example, the 95% Bayesian confidence interval for μ_{stop} ranges from 178.30 to 195.70, indicating that we can be 95% confident that the true value of μ_{stop} lies within this range. The Bayesian confidence interval is the narrowest for the μ_{stop} parameter, indicating that μ_{stop} is estimated the most precisely among the stop parameters.

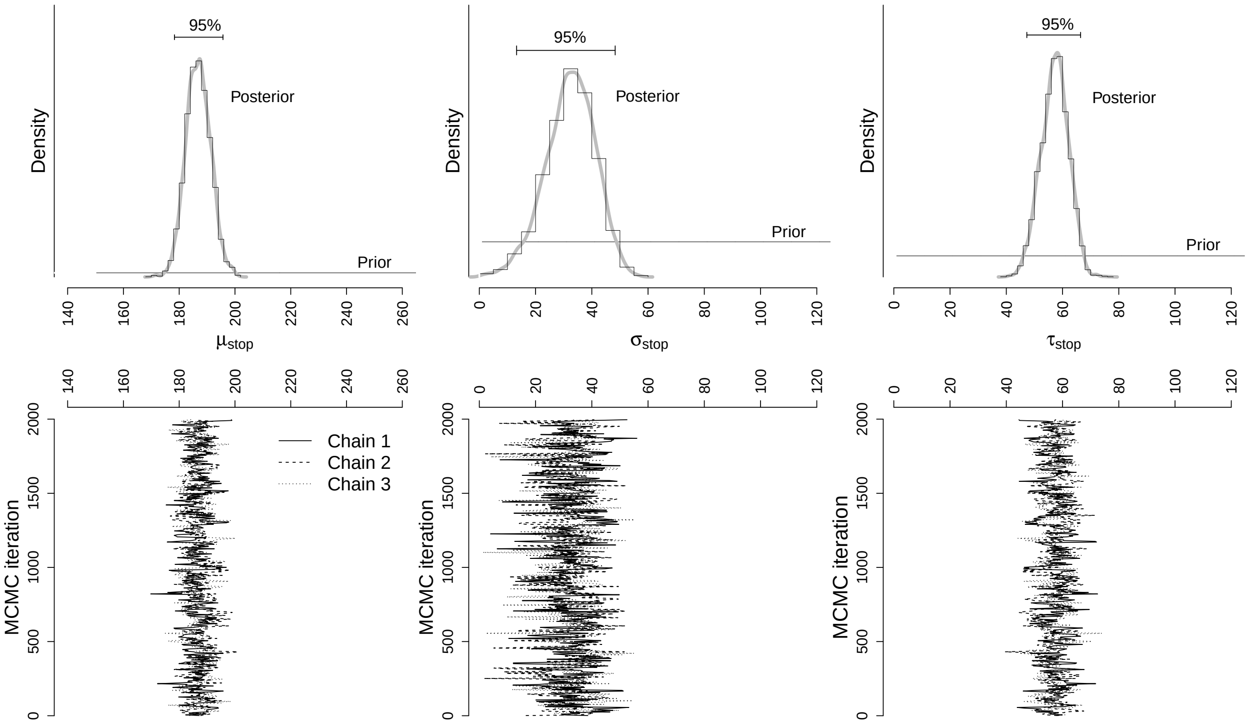


Figure 3.6 *Illustration of Markov chain Monte Carlo (MCMC)-based Bayesian estimation for the ex-Gaussian parameters μ_{stop}, σ_{stop}, and τ_{stop} for a synthetic data set with the individual Bayesian parametric approach.* The histograms in the top panels show the posterior distribution of the parameters. The corresponding thick gray lines indicate the fit of a nonparametric density estimator to the posterior samples. The horizontal black lines at the bottom show the prior distribution of the parameters. The horizontal black lines at the top show the 95% Bayesian confidence interval. The solid, dashed and dotted lines in the bottom panels represent the different sequences of values (i.e., MCMC chains) sampled from the posterior distribution of the parameters.

The Bayesian approach can be applied to hierarchical as well as individual data. In individual estimation, the parameters of the SSRT distribution are estimated separately for each participant. In the hierarchical approach (e.g., Gelman & Hill, 2007; M. D. Lee, 2011; Lindley & Smith, 1972; Rouder et al., 2005, 2003), the estimation of the individual stop parameters is supported by information from the entire group. In the next section, we introduce the individual and the hierarchical BPA models for estimating SSRT distributions.

Individual BPA

Figure 3.7 shows the graphical model for the individual BPA. Observed variables are represented by shaded nodes and unobserved variables are represented by unshaded nodes. The graph structure indicates dependencies between the nodes, and the plates represent independent replications of the different types of trials (e.g., M. D. Lee, 2008).

The individual BPA assumes that the $g = 1, ..., G$ go RTs come from an ex-Gaussian distribution, with parameters μ_{go}, σ_{go}, and τ_{go} (see Equation 3.13). On the $s = 1, ..., S$ stop-signal trials, the $r = 1, ..., R$ signal-respond RTs (i.e., SR-RT) and the $i = 1, ..., I$ successful inhibitions (i.e., NA)

Model parameters:

$\mu_{go} \sim \text{Uniform}(1, 1000)$

$\sigma_{go} \sim \text{Uniform}(1, 300)$

$\tau_{go} \sim \text{Uniform}(1, 300)$

$\mu_{stop} \sim \text{Uniform}(1, 600)$

$\sigma_{stop} \sim \text{Uniform}(1, 250)$

$\tau_{stop} \sim \text{Uniform}(1, 250)$

Data:

$goRT_g \sim \text{ExGaussian}(\mu_{go}, \sigma_{go}, \tau_{go})$

$SR - RT_r \sim \text{CensoredExGaussian} - \text{SR}(\mu_{go}, \sigma_{go}, \tau_{go}, \mu_{stop}, \sigma_{stop}, \tau_{stop}, SSD_s)$

$NA_i \sim \text{CensoredExGaussian} - \text{I}(\mu_{go}, \sigma_{go}, \tau_{go}, \mu_{stop}, \sigma_{stop}, \tau_{stop}, SSD_s)$

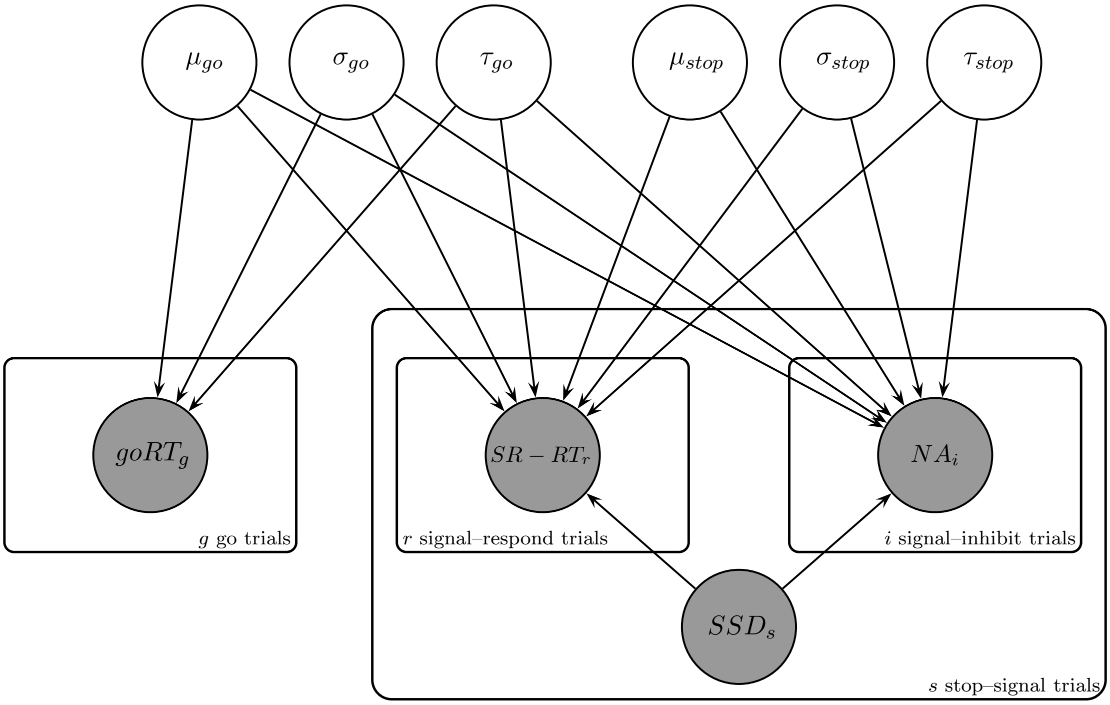


Figure 3.7 *Graphical model for the individual Bayesian parametric approach.* Observed variables are represented by shaded nodes and unobserved variables are represented by unshaded nodes. The plates represent independent replications of the different types of trials. The go response times (RTs) come from an ex-Gaussian distribution, with parameters μ_{go}, σ_{go}, and τ_{go}. The signal-respond RTs (i.e., SR-RT) and the successful inhibitions (i.e., NA) come from censored ex-Gaussian distributions, with parameters μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, τ_{stop}, and stop-signal delay (SSD). The priors for the model parameters are uniform distributions.

come from censored ex-Gaussian distributions, with parameters μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, τ_{stop}, and SSD$_s$ (see Equation 3.14). The priors for the model parameters are uniform distributions, spanning a plausible but wide range of parameter values. The range of the uniform prior distributions is loosely based on the results of a life-span study of stop-signal performance reported in Williams et al. (1999) and the corresponding ex-Gaussian parameter values used in the simulation studies of Band et al. (2003).

The individual BPA makes no connections between participants; it assumes that they are completely independent. The goal is to estimate the μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, and τ_{stop} parameters for each participant separately. To this end, we use WinBUGS to sample from the posterior distribution of the model parameters. The WinBUGS script for the individual BPA is available in Appendix B.1. The median of the posterior distributions can be used as a point estimate of the model parameters. SSRT distributions, such as those shown in Figure 3.1, can be obtained by evaluating the ex-Gaussian probability density function (Equation 3.8) with the posterior median of the parameters. The mean and the variance of the SSRT distribution can be computed from Equation 3.11 and Equation 3.12 with the posterior median of the parameters. Alternatively, we can quantify the uncertainty of the estimated SSRT distribution by drawing random parameter vectors from the joint posterior of the stop parameters and evaluating the ex-Gaussian probability density function using the chosen parameter vectors.

Hierarchical BPA

A particularly useful application of the Bayesian hierarchical approach (Farrell & Ludwig, 2008; Gelman & Hill, 2007; M. D. Lee, 2011; Nilsson, Rieskamp, & Wagenmakers, 2011; Rouder & Lu, 2005; Rouder et al., 2005, 2003; Shiffrin et al., 2008; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) explicitly models individual differences in the parameter values, but at the same time recognizes that participants share some similarities. Hierarchical modeling is a compromise between the assumption that participants are completely independent (i.e., individual BPA) and the assumption that all participant are identical (Gelman & Hill, 2007). Rather than estimating the parameters separately for each individual, hierarchical modeling assumes that the individual parameters are drawn from group-level distributions. The group-level distributions specify how the individual parameters are distributed in the population and thus define the between-subject variability in the model parameters. The goal is to obtain individual parameter estimates as well as estimates for the parameters of the group-level distributions.

Hierarchical methods have the potential to provide more accurate and less variable parameter estimates than individual Bayesian and maximum likelihood estimation (Farrell & Ludwig, 2008; Rouder et al., 2005). The advantages of hierarchical modeling are the most pronounced in situations with only moderate between-subject variability and a small number of observations per participant (Gelman & Hill, 2007). The benefits of hierarchical modeling arise from using information available from the whole group to improve parameter estimation for the individual participants. Hierarchical modeling uses the group-level distributions as priors to adjust poorly estimated extreme parameter values to more moderate ones. As a result, outlying individual estimates —especially the ones that are estimated with a great degree of uncertainty— are "shrunk" towards the group mean. The hierarchical approach is especially valuable in situations with relatively few observations per participant, as is often the case in stop-signal experiments.

Figure 3.8 shows the graphical model for the hierarchical BPA. The hierarchical BPA assumes that the $g = 1, ..., G$ go RTs of each participant, $j = 1, ..., J$, come from ex-Gaussian distributions, but with different values of μ_{go}, σ_{go}, and τ_{go}. On the $s = 1, ..., S$, stop-signal trials, the $r = 1, ..., R$ signal-respond RTs (i.e., SR-RT) and the $i = 1, ..., I$ successful inhibitions (i.e., NA) of each participant come from censored ex-Gaussian distributions, but again with different values of μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, τ_{stop}, and SSD_s. The individual μ_{go_j}, σ_{go_j}, τ_{go_j}, μ_{stop_j}, σ_{stop_j}, and τ_{stop_j} parameters are in turn assumed to come from truncated normal group-level distributions that are characterized by group-level parameters. For example, the μ_{stop} parameters codetermine the location of the individual SSRT distributions. As SSRTs are by definition positive, the μ_{stop} parameters must be positive as well. The μ_{stop} parameters are therefore assumed to come from a normal group-level distribution truncated at 0 ms, with mean $\mu_{\mu_{stop}}$ and standard deviation

$\sigma_{\mu_{stop}}$. Similarly, the σ_{stop} parameters are the standard deviations of the Gaussian component of the individual SSRT distributions and are by definition positive (see Equation 3.8). The σ_{stop} parameters are assumed to come from a normal group-level distribution truncated at one ms, with mean $\mu_{\sigma_{stop}}$ and standard deviation $\sigma_{\sigma_{stop}}$.[2]

The use of normal group-level distributions is a common choice in Bayesian hierarchical modeling (e.g., Gelman & Hill, 2007; M. D. Lee & Wagenmakers, 2013). In real data, however, the distribution of individual parameters may deviate from normality, especially in clinical populations. As will be described later, sensitivity analyses indicated that the hierarchical BPA is robust to misspecification of the group-level distribution of the individual go parameters. In contrast, the BPA with misspecified group-level distributions results in biased stop parameter estimates. Fortunately, the bias decreases substantially as the number of participants and especially as the number of trials increase. The reader is referred to the Discussion for some suggestions on examining the validity of the hierarchical assumptions of the BPA.

The priors for the mean and standard deviation of the group-level distributions are normal and uniform distributions, respectively. For example, the $\mu_{\sigma_{stop}}$ parameter is the mean of the group-level distribution of the individual σ_{stop} parameters and as such it must be positive. The group mean $\mu_{\sigma_{stop}}$ parameter is assumed to come from a normal distribution censored to be positive, with mean 40 and standard deviation $1/\sqrt{0.001}$. The group standard deviation $\sigma_{\sigma_{stop}}$ parameter is assumed to be uniformly distributed between 0 and 100. The parameters of the priors for the group-level means and standard deviations are loosely based on the results reported in Williams et al. (1999) and the corresponding ex-Gaussian parameter values used in Band et al. (2003).

In the hierarchical BPA, the goal is to estimate the group-level means and standard deviations as well as the individual go and stop parameters. The WinBUGS script for the hierarchical BPA is available in Appendix B.1. The median of the posterior distributions can be used as a point estimate for the parameters. The SSRT distribution of each participant can be obtained by evaluating the ex-Gaussian probability density function with the posterior median of the individual parameters. The mean and the variance of the individual SSRT distributions can be computed from Equation 3.11 and Equation 3.12 with the posterior median of the individual parameters. Also, we can quantify the uncertainty of the individual SSRT distributions by drawing random parameter vectors from the joint posterior of the individual stop parameters and evaluating the ex-Gaussian probability density function using the chosen parameter vectors.

3.4 Parameter Recovery Studies

Individual BPA

We conducted two simulation studies to investigate the ability of the individual BPA to recover underlying true parameter values. The first recovery study examined the asymptotic properties of the parameter estimates. The second recovery study investigated the number of stop-signal trials necessary to obtain accurate parameter estimates.

Methods

We generated stop-signal data from the horse race model, where the go RTs and the SSRTs were drawn from ex-Gaussian distributions, with parameters $\mu_{go} = 440$, $\sigma_{go} = 80$, $\tau_{go} = 60$, $\mu_{stop} = 190$,

[2] For computational reasons, the truncated normal group-level distributions of the σ_{go}, τ_{go}, σ_{stop}, and τ_{stop} parameters are truncated at 1 ms instead of 0.

Group–level parameters:

$\mu_{\mu_{go}} \sim \text{Normal}(500, 0.0001)I[0, \infty]$

$\sigma_{\mu_{go}} \sim \text{Uniform}(0, 300)$

$\mu_{\sigma_{go}} \sim \text{Normal}(100, 0.001)I[0, \infty]$

$\sigma_{\sigma_{go}} \sim \text{Uniform}(0, 200)$

$\mu_{\tau_{go}} \sim \text{Normal}(80, 0.001)I[0, \infty]$

$\sigma_{\tau_{go}} \sim \text{Uniform}(0, 200)$

$\mu_{\mu_{stop}} \sim \text{Normal}(200, 0.0001)I[0, \infty]$

$\sigma_{\mu_{stop}} \sim \text{Uniform}(0, 200)$

$\mu_{\sigma_{stop}} \sim \text{Normal}(40, 0.001)I[0, \infty]$

$\sigma_{\sigma_{stop}} \sim \text{Uniform}(0, 100)$

$\mu_{\tau_{stop}} \sim \text{Normal}(30, 0.001)I[0, \infty]$

$\sigma_{\tau_{stop}} \sim \text{Uniform}(0, 100)$

Individual parameters:

$\mu_{go_j} \sim \text{TruncatedNormal}(\mu_{\mu_{go}}, \frac{1}{\sigma^2_{\mu_{go}}}, 0)$

$\sigma_{go_j} \sim \text{TruncatedNormal}(\mu_{\sigma_{go}}, \frac{1}{\sigma^2_{\sigma_{go}}}, 1)$

$\tau_{go_j} \sim \text{TruncatedNormal}(\mu_{\tau_{go}}, \frac{1}{\sigma^2_{\tau_{go}}}, 1)$

$\mu_{stop_j} \sim \text{TruncatedNormal}(\mu_{\mu_{stop}}, \frac{1}{\sigma^2_{\mu_{stop}}}, 0)$

$\sigma_{stop_j} \sim \text{TruncatedNormal}(\mu_{\sigma_{stop}}, \frac{1}{\sigma^2_{\sigma_{stop}}}, 1)$

$\tau_{stop_j} \sim \text{TruncatedNormal}(\mu_{\tau_{stop}}, \frac{1}{\sigma^2_{\tau_{stop}}}, 1)$

Data:

$goRT_{gj} \sim \text{ExGaussian}(\mu_{go_j}, \sigma_{go_j}, \tau_{go_j})$

$SR - RT_{rj} \sim \text{CensoredExGaussian} - \text{SR}(\mu_{go_j}, \sigma_{go_j}, \tau_{go_j}, \mu_{stop_j}, \sigma_{stop_j}, \tau_{stop_j}, SSD_{sj})$

$NA_{ij} \sim \text{CensoredExGaussian} - \text{I}(\mu_{go_j}, \sigma_{go_j}, \tau_{go_j}, \mu_{stop_j}, \sigma_{stop_j}, \tau_{stop_j}, SSD_{sj})$

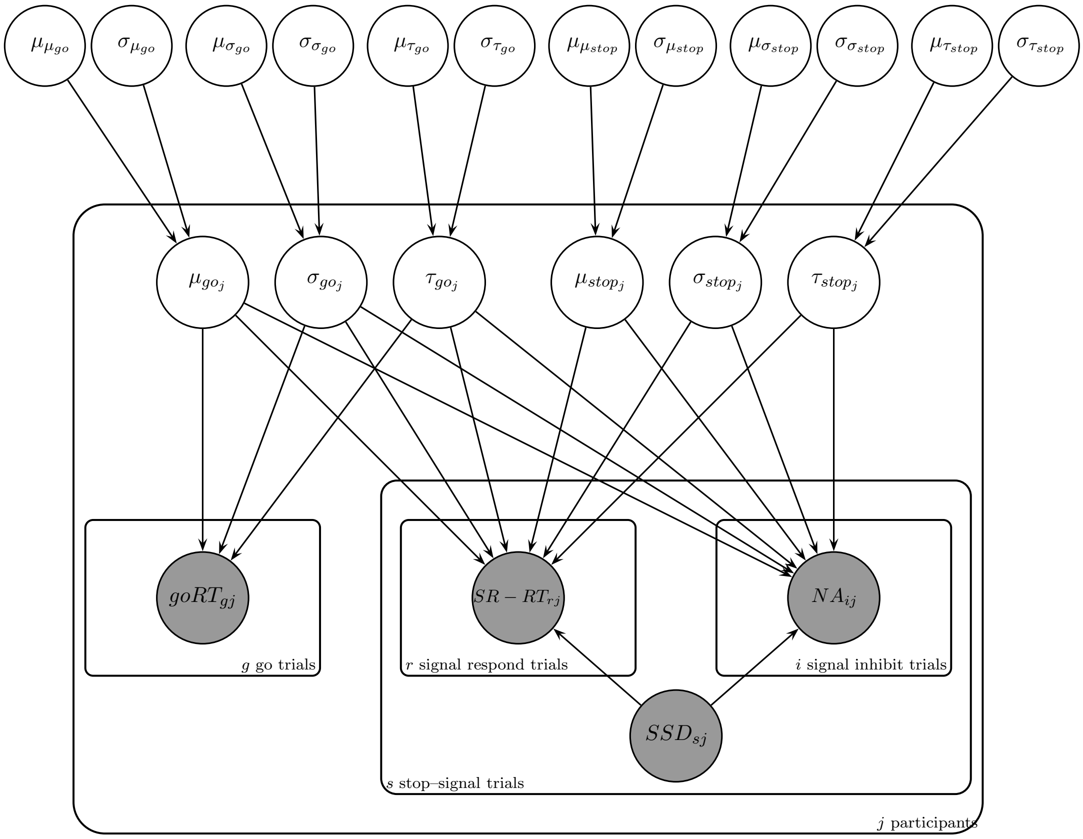


Figure 3.8 *Graphical model for the hierarchical Bayesian parametric approach.* The go response times (RTs) of each participant come from ex-Gaussian distributions, with different values of μ_{go}, σ_{go}, and τ_{go}. The signal-respond RTs (i.e., SR-RT) and the successful inhibitions (i.e., NA) of each participant come from censored ex-Gaussian distributions, with different values of μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, τ_{stop}, and stop-signal delay (SSD). The individual go and stop parameters come from truncated normal group-level distributions that are characterized by group-level parameters. In order to maintain consistency with the WinBUGS syntax, the group-level normal and truncated normal distributions are parameterized in terms of their precision (i.e., inverse variance) rather than their variance. The $I[0, \infty]$ construct denotes distributional censoring with lower bound equal to 0 and upper bound equal to infinity.

$\sigma_{stop} = 40$, and $\tau_{stop} = 50$.[3] These parameters made the mean and the standard deviation of the go RT distribution 500 and 100 ms, respectively, and the mean and the standard deviation of the SSRT distribution 240 and 64 ms, respectively. The SSDs were set to 150, 200, 250, 300, and 350 ms. The above parameter vales and SSDs resulted in $P(\text{respond} \mid \text{stop signal, SSD} = 150) = 0.17$, $P(\text{respond} \mid \text{stop signal, SSD} = 200) = 0.30$, $P(\text{respond} \mid \text{stop signal, SSD} = 250) = 0.47$, $P(\text{respond} \mid \text{stop signal, SSD} = 300) = 0.65$, and $P(\text{respond} \mid \text{stop signal, SSD} = 350) = 0.79$. The resulting inhibition function for a randomly chosen data set is shown in Figure 3.3.

In the first recovery study, we generated a single data set containing 200,000 go trials and 5 (SSD) × 100,000 stop-signal trials. The estimated μ_{stop}, σ_{stop}, and τ_{stop} parameters were free to vary across the five SSDs. In the second recovery study, we conducted four sets of simulations that varied the number of trials, with 100 data sets for each set. For the first set, each data set contained 4,500 go trials and 5 (SSD) × 300 stop-signal trials. For the second set, each data set contained 2,250 go trials and 5 × 150 stop-signal trials. For the third set, each data set contained 750 go trials and 5 × 50 stop-signal trials. For the fourth set, each data set contained 375 go trials and 5 × 25 stop-signal trials. In contrast to the first recovery study, the estimated μ_{stop}, σ_{stop}, and τ_{stop} parameters were constrained to be equal across the five SSDs.

We fit the data sets with the individual BPA using WinBUGS. We ran three MCMC chains and used overdispersed starting values to confirm that the chains have converged to the stationary distribution ($\hat{R} \approx 1$). The first 500 samples of each MCMC chain were discarded. The reported parameter estimates are based on 3 × 4,000 recorded samples.

Results

The parameters of the go RT distribution were excellently recovered in both recovery studies. As the go RT distribution is of little theoretical interest, the remainder of this section focuses exclusively on results related to the SSRT distribution.

The results of the first recovery study are shown in Figure 3.9. For all SSDs, the posterior median recovered the generating parameter values, and the mean and the standard deviation of the true SSRT distributions very well. Across the five SSDs, the posterior standard deviations ranged from 0.92 to 1.84 for μ_{stop}, from 1.50 to 2.2 for σ_{stop}, and from 0.91 to 2.09 for τ_{stop}. The posterior standard deviations were small, indicating that the parameters were estimated precisely. In contrast to the integration method, the mean SSRT estimated with the BPA did not decrease with increasing SSD. Theoretically, one may obtain accurate estimates for the stop parameters using stop-signal data on a single SSD.

The results of the second recovery study are shown in Figure 3.10 and Figure 3.11. Figure 3.10 shows the mean of the posterior medians of the stop parameters across the 100 replications and the mean and standard deviation of the SSRT distribution. Figure 3.11 shows the estimated SSRT distributions based on the posterior medians for the 100 replications. As shown in the figures, the BPA recovered the generating parameter values and the shape of the true SSRT distribution with little bias even with relatively few (i.e., 25) stop-signal trials per SSD. Naturally, as the number of trials increased, the bias, the standard error, and the posterior standard deviation of the estimates decreased. The mean posterior standard deviation of μ_{stop} across the 100 replications decreased from 27.40 for the simulation set with 25 stop-signal trials per SSD to 9.51 for the set with 300 stop-signal trials per SSD. The mean posterior standard deviation of σ_{stop} decreased from 26.05 to 12.51. The mean posterior standard deviation of τ_{stop} decreased from 26.64 to 9.98. Note also that the BPA parameter estimates as well as the average of the integration method estimates across the

[3] We conducted several recovery studies using alternative true parameter values. The results were essentially the same as the ones reported here.

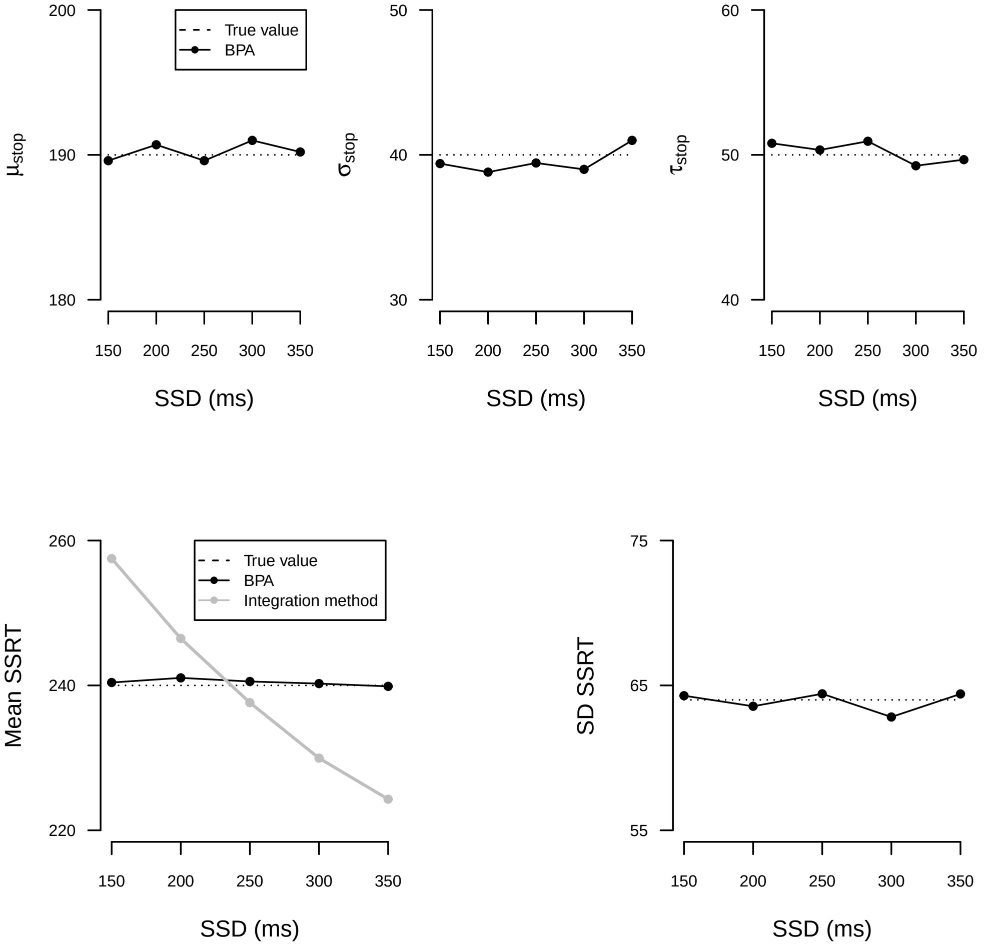


Figure 3.9 *Posterior medians of the stop parameters and the mean and standard deviation of the stop-signal reaction time (SSRT) distribution from the first recovery study for the individual Bayesian parametric approach.* The results are based on one data set containing 200,000 go trials and 100,000 stop-signal trials per stop-signal delay (SSD). The estimated μ_{stop}, σ_{stop}, and τ_{stop} parameters were free to vary across the five SSDs. The dashed lines give the true value of the stop parameters and the true mean and standard deviation of the SSRT distribution. In the top panels, the black bullets show the posterior median of the estimated stop parameters. In the bottom panels, the black bullets show the estimated mean and standard deviation of the SSRT distribution computed with the the posterior median of the stop parameters. The gray bullets show SSRT estimates computed with the traditional integration method.

SSDs recovered the mean of the generating SSRT distribution very accurately even with only 25 stop-signal trials per SSD.

To summarize, the results of the two simulation studies indicated that the individual BPA accurately recovered the parameters of the generating SSRT distribution. Also, similar to the integration method, the BPA recovered the mean of the generating SSRT distribution very accurately. As the number of stop-signal trials increased, the stop parameters and the mean SSRT were estimated more precisely. Nevertheless, the individual BPA was able to provide reasonable estimates even with relatively scarce data (i.e., 5 × 25 stop-signal trials) often encountered in stop-signal studies.

Hierarchical BPA

Methods

We generated 100 data sets from the horse race model, each containing the stop-signal data of $j = 1, ..., 25$ participants. The individual parameters μ_{go_j}, σ_{go_j}, τ_{go_j}, μ_{stop_j}, σ_{stop_j}, and τ_{stop_j} were drawn from truncated normal distributions. The individual parameters were then used to generate 300 go trials and 100 stop-signal trials for each participant, using the ex-Gaussian distribution. The generating parameter values are shown in Figure 3.12. For computational efficiency, we used a single SSD per participant that produced a $P(\text{respond} \mid \text{stop signal})$ equal to 0.50.

We fit the 100 data sets with the hierarchical BPA using WinBUGS. We ran three MCMC chains and used overdispersed starting values. The first 3,000 samples of each MCMC chain were discarded. The reported parameter estimates are based on 3 × 7,750 recorded samples.

Results

In this section, we focus exclusively on results related to the group-level parameters of the go RT and the SSRT distribution. The individual parameter estimates from the hierarchical BPA is discussed in the next section with experimental data.

Figure 3.12 shows the posterior median of the group-level parameters averaged over the 100 replications. The hierarchical BPA recovered the group-level parameters quite accurately. The posterior standard deviations and the standard errors are typically larger for the stop parameters than for the go parameters. This result is not surprising because the go parameters are estimated based on the go RTs as well as the signal-respond RTs. The go parameters are therefore better constrained by the data than the stop parameters.

In sum, the results of the simulation studies indicate that the individual and the hierarchical BPA accurately recovered the true individual stop parameters and the generating group-level parameters, respectively. In contrast to Colonius' (1990) method, the BPA resulted in accurate estimates with a reasonable amount of data. The individual BPA provided accurate estimates for the stop parameters with only 125 stop-signal trials per participant. The hierarchical BPA yielded precise group-level stop parameters with a modest sample size of only 25 participants, each performing as few as 100 stop-signal trials.

3.5 Fitting Experimental Data

The aim of this section is to illustrate the application of the BPA using the stop-signal data set reported by Bissett and Logan (2011). Bissett and Logan presented participants with two sessions

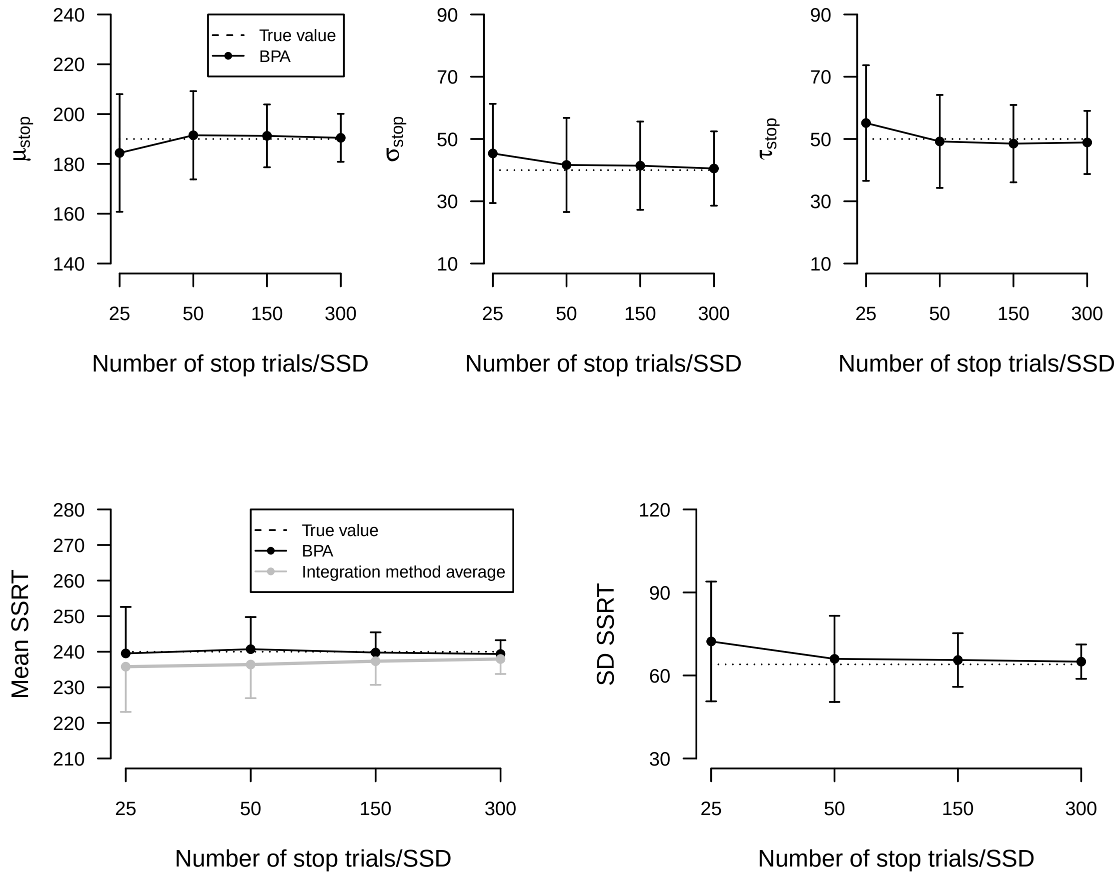


Figure 3.10 *Posterior medians of the stop parameters and the mean and standard deviation of the stop-signal reaction time (SSRT) distribution from the second recovery study for the individual Bayesian parametric approach.* We conducted four sets of simulations that varied the the number of go and stop-signal trials, with 100 data sets for each set. The estimated μ_{stop}, σ_{stop}, and τ_{stop} parameters were constrained to be equal across the five stop-signal delays (SSDs). The dashed lines give the true value of the stop parameters and the true mean and standard deviation of the SSRT distribution. In the top panels, the black bullets show the mean of the posterior medians of the estimated stop parameters across the 100 replications. In the bottom panels, the black bullets show the mean of the estimated mean and standard deviation of the SSRT distribution computed with the posterior median of the stop parameters across the 100 replications. The gray bullets show SSRT estimates computed by averaging the integration method SSRT estimates over the five SSDs. The vertical lines indicate the size of the standard error across the 100 replications.

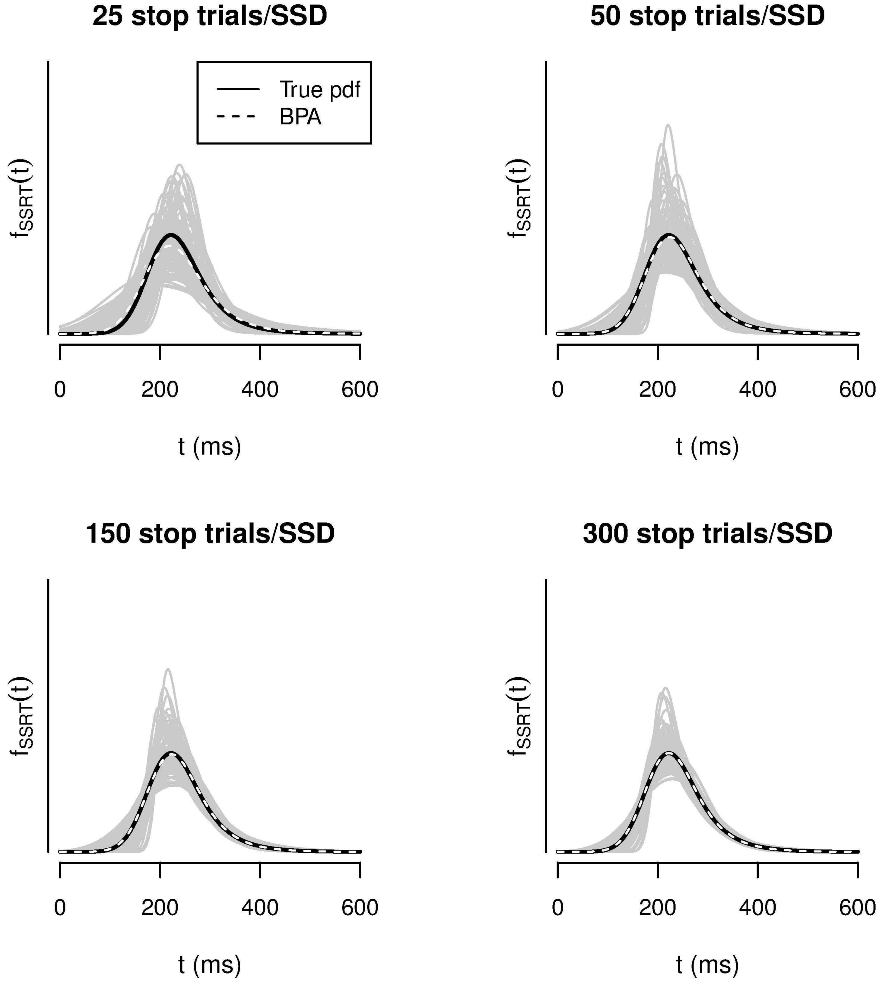


Figure 3.11 *Estimated stop-signal reaction time (SSRT) distributions from the second recovery study for the individual Bayesian parametric approach.* We conducted four sets of simulations that varied the number of go and stop-signal trials, with 100 data sets for each set. The estimated μ_{stop}, σ_{stop}, and τ_{stop} parameters were constrained to be equal across the five stop-signal delays (SSDs). The solid black line shows the true SSRT distribution. The gray lines show the SSRT distributions based on the posterior medians of the 100 replications. The dashed white line shows the SSRT distribution based on the mean of the posterior medians of the stop parameters across the 100 replications. pdf = probability density function.

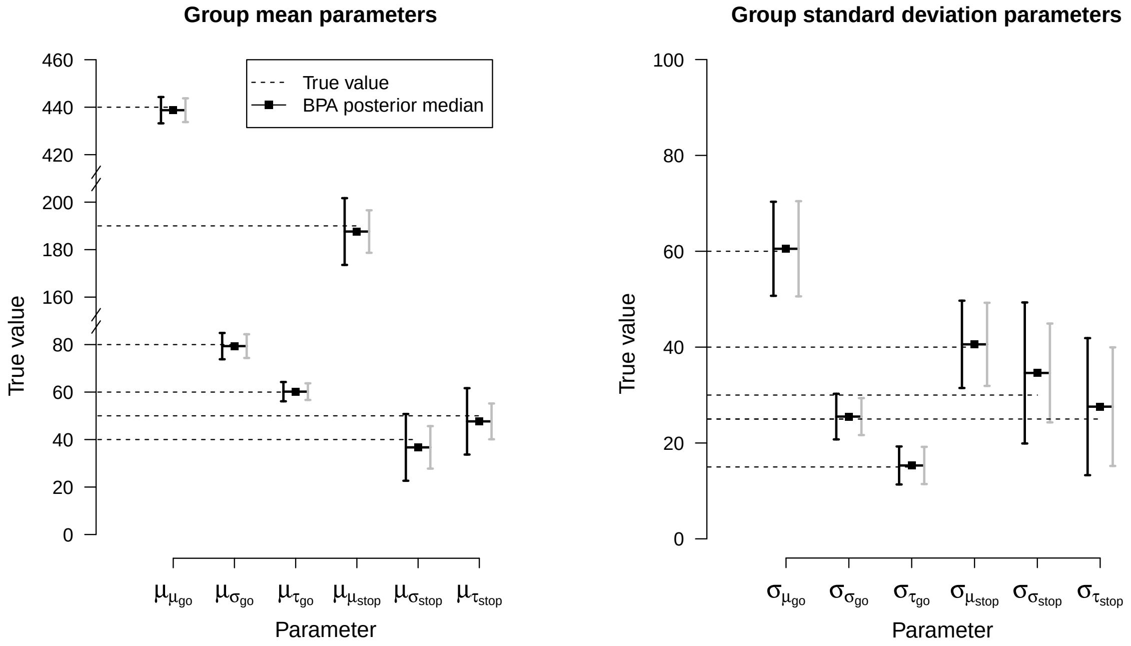


Figure 3.12 *Posterior medians of the group-level parameters from the hierarchical parameter recovery study.* We generated 100 data sets from the horse race model, each containing the stop-signal data of 25 participants responding to 300 go trials and 100 stop-signal trials. The dashed lines give the true value of the parameters. The black bullets indicate the mean of the posterior medians across the 100 replications. The black vertical lines show the size of the posterior standard deviations averaged across the 100 replications. The gray vertical lines indicate the size of the standard error. BPA = Bayesian parametric approach.

of the stop-signal task in order to investigate the adjustment of speed and caution in a dual-task environment. Here we focus on the first experiment of the Bissett and Logan study that manipulated the percentage of stop-signal trials across two sessions. The authors concluded that the two experimental sessions did not differ significantly in mean SSRT.

The Data Set

The go task required the 24 participants to respond to the shape of the presented stimuli. For instance, participants responded by pressing the “1” key on the computer keyboard when presented with a triangle or a circle, and by pressing the “0” key when presented with a square or a diamond. Each participant performed two sessions of the task. The first session featured 960 go trials and 240 stop-signal trials, resulting in 20% stop-signal trials. The second session featured 720 go trials and 480 stop-signal trials, resulting in 40% stop-signal trials. The SSD was set using the staircase tracking procedure; SSD was lengthened by 50 ms after successful inhibitions and SSD was shortened by 50 ms after incorrect responses, yielding 50% inhibition for each participant.

Incorrect RTs and RTs shorter than 200 ms and longer than 1,850 ms were excluded from all subsequent analyses (see Bissett & Logan, 2011). As the ex-Gaussian distribution is sensitive to outliers, we also removed RTs that were slower or faster than a given participant's mean RT plus or minus 2 × the standard deviation. For comparison, we report the results of fitting the raw data with the individual BPA and the results of fitting the data without the outliers. Moreover, we excluded four participants with erratic stop-signal performance, such as extremely long and variable go RTs and a very large number of SSDs.

Individual BPA

The individual BPA was fit to the Bissett and Logan (2011) data set with WinBUGS. The estimated μ_{stop}, σ_{stop}, and τ_{stop} parameters of each participant were constrained to be equal across the different SSDs. We ran three MCMC chains and used overdispersed starting values. The analyses were based on 3 × 5,000 recorded samples.

The parameters of the go RT distribution were estimated precisely for all of the 20 participants. The remainder of this section focuses exclusively on the parameters of the SSRT distribution. With the exception of four participants in the first session, the posterior distributions of the stop parameters were estimated well. For the four exceptions, we obtained unrealistically large posterior medians for τ_{stop} and/or very large posterior standard deviations for σ_{stop} and τ_{stop}.[4] In these cases, the stop parameters were thus estimated with great uncertainty, resulting in uninformative SSRT distributions and uninterpretable mean SSRT estimates.

The results from fitting the Bissett and Logan (2011) data set with the individual BPA are shown in Figure 3.13 and Figure 3.14. Figure 3.13 compares the mean SSRT of each participant computed with the mean method to the mean SSRT computed with the BPA posterior medians of the stop parameters. The BPA produced mean SSRT estimates very similar to those obtained by the mean method. The correspondence between the two methods further improved after the outliers were removed; this is not surprising because the two methods are affected to different degrees by the presence of outliers. Also, the agreement between the two sets of estimates is better for the second session than for the first session. Again, this is to be expected because the second session featured twice as many stop-signal trials than the first session, resulting in more accurate estimates for both methods.

The circles in the left panels of Figure 3.13 mark the three data points with the largest discrepancy between the two methods. The three estimates are clustered together at very high values of BPA mean SSRT. Note that these mean SSRTs belonged to three of the participants with uninformative posterior distributions with high medians and very large standard deviations for τ_{stop}. The high posterior median for τ_{stop} resulted in unusually high BPA mean SSRTs (see Equation 3.11). However, due to the large posterior uncertainty of τ_{stop}, the resulting mean SSRT estimates are uninterpretable. Lastly, consider the data point marked with "A" in the bottom right panel of Figure 3.13. For this mean SSRT, the mean method resulted in an unrealistic estimate of 118 ms. The BPA, however, yielded a more reasonable estimate of 209 ms.

The first published estimates of entire SSRT distributions are shown in Figure 3.14. The gray SSRT distributions are based on the posterior medians of the individual stop parameters. There is considerable between-participant variability in the shape of the SSRT distributions. Some distributions are very peaked, whereas others are more spread out indicating substantial within-participant variability in SSRT. Note the few extremely flat distributions with very large variance and long tail in the left panels of Figure 3.14. These flat distributions belonged to the four

[4] For these four participants, we used a uniform prior distribution ranging from 1 to 450 for τ_{stop} to accommodate the extreme parameter estimates. Note also that these participants are not the same as the four participants who were previously excluded from the analyses.

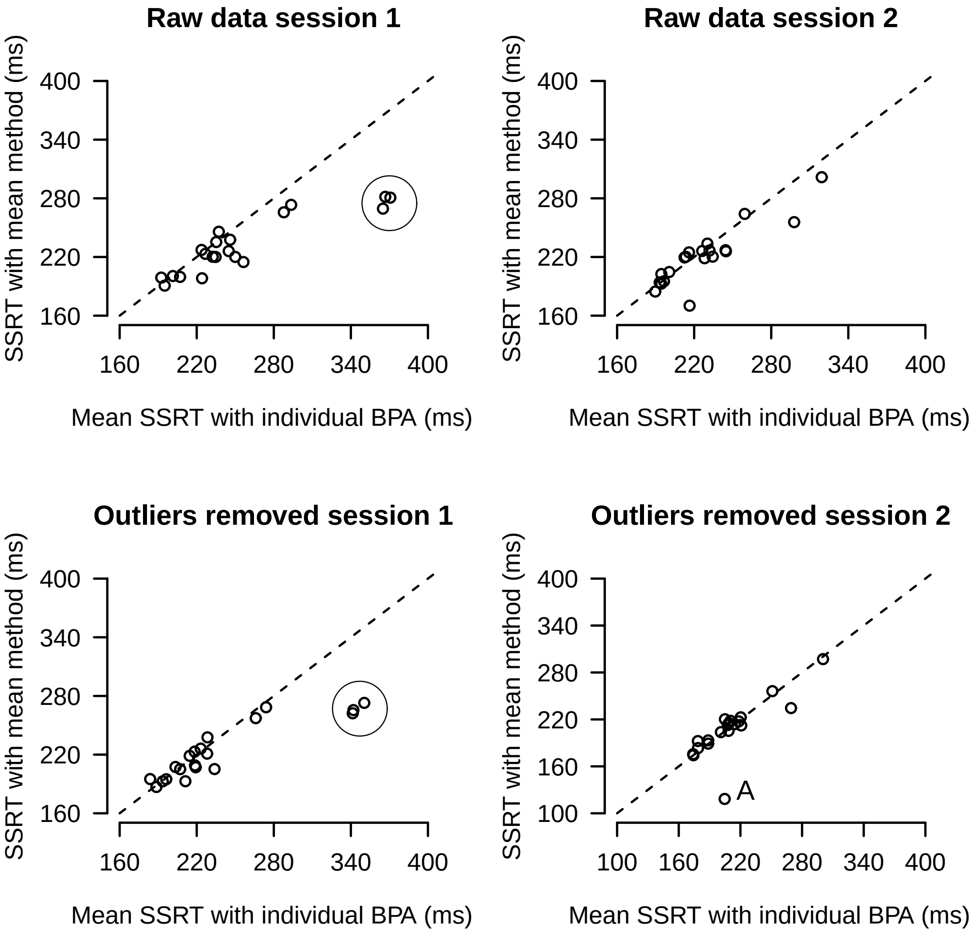


Figure 3.13 *Comparison of mean stop-signal reaction times (SSRTs) computed using the mean method and the individual Bayesian parametric approach (BPA) posterior medians in the Bissett and Logan (2011) data set.* The data points marked with circles represent mean SSRTs that are based on the imprecise and therefore uninterpretable posterior distributions. The data point marked with “A” in the bottom right panel represents a mean SSRT estimate for which the individual BPA resulted in a more reasonable estimate than the mean method.

.

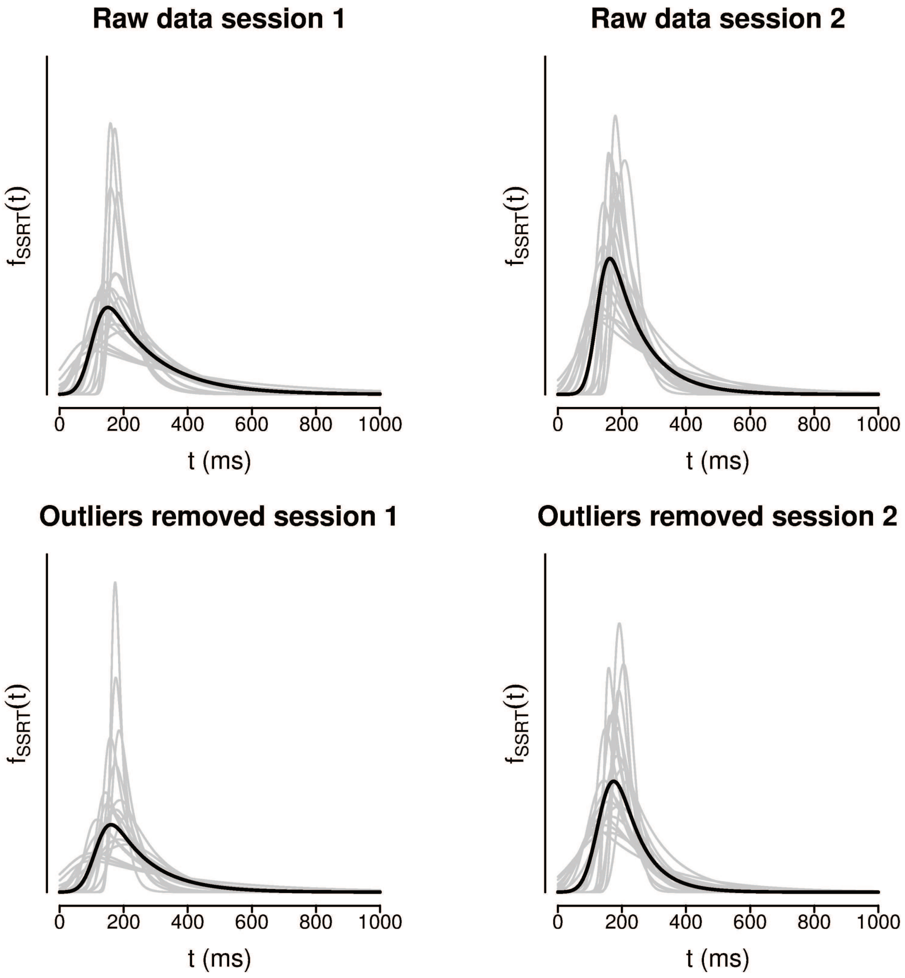


Figure 3.14 *Estimated stop-signal reaction times (SSRT) distributions for the Bissett and Logan (2011) data set.* The gray lines show the SSRT distributions based on the posterior medians of the stop parameters of each individual participant. The black line shows the SSRT distribution based on the mean of the posterior medians of the stop parameters across the 20 participants.

participants with the uninformative posterior distributions. The resulting SSRT distributions are therefore also uninformative.

The solid black line in Figure 3.14 shows the average SSRT distribution created using the mean of the posterior medians of the individual μ_{stop}, σ_{stop} and τ_{stop} parameters across participants. There are substantial differences between the shape of the average SSRT distributions in the two sessions of the experiment. The average SSRT distribution for the first session is spread out and has a fast leading edge and a long tail. In contrast, the average SSRT distribution for the second session is more peaked, with a slower leading edge and a shorter tail. Despite these differences, consistent with the results of Bissett and Logan (2011), the means of the two distributions are roughly equal. Similar to the example shown in Figure 3.1, ignoring the differences in the shape of these SSRT distributions would lead to the incorrect conclusion that the two experimental sessions do not differ with respect to SSRT.

In conclusion, the individual BPA provided well-behaved posterior distributions for most participants. Moreover, the mean SSRTs computed with the BPA posterior medians accurately approximated the mean SSRTs obtained by the traditional mean method. In the next section, we formally investigate whether the individual BPA adequately described the observed data.

Assessing Model Fit Using Posterior Predictive Model Checks

We used posterior predictive model checks to determine whether the individual BPA produced parameter estimates that adequately describe the Bissett and Logan (2011) data. Posterior predictive model checks are frequently used procedures in Bayesian inference to assess the absolute goodness of fit of a proposed model (e.g., Gelman & Hill, 2007; Gelman, Meng, & Stern, 1996). In posterior predictive checks, we assess the adequacy of the model by generating new data (i.e., predictions) using the posterior distributions of the parameters obtained from fitting the model. If the model adequately describes the data, the predictions based on the model parameters should closely approximate the observed data.

We formalized the model checks by computing posterior predictive p values (e.g., Gelman & Hill, 2007; Gelman et al., 1996). We first defined a test statistic T, and computed its value for the observed data: $T(data)$. For each of the $i = 1, ..., N$ draws from the posterior distribution of the parameters, we sampled new stop-signal data, $data^* = (data^*_1, data^*_2, ..., data^*_N)$, using the ex-Gaussian assumption. Lastly, we calculated the test statistic T for each $data^*_i$: $T(data^*_i)$. The posterior predictive p value is given by the fraction of times that $T(data^*)$ is greater than $T(data)$. The posterior predictive p value compares thus the observed value of the test statistic to its sampling distribution under the assumptions of the BPA. Extreme p values close to 0 or 1 (e.g., lower than 0.05 or higher than 0.95) indicate that the BPA does not describe the observed data adequately. For each participant we conducted two posterior predictive analyses using different test statistics.

In the first posterior predictive analysis, we compared the observed signal-respond RT distribution to the signal-respond RTs predicted by the posterior distribution of the model parameters. The model check was performed only for the SSD with the highest number of observed signal-respond RTs in order to obtain stable observed and predicted signal-respond RT distributions. For each participant, we randomly selected $N = 1,000$ parameter vectors from the joint posterior of μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, and τ_{stop}. Then, we generated 1,000 stop-signal data sets using the 1,000 parameter vectors, the chosen SSD and the corresponding number of stop-signal trials. We used the median of the signal-respond RTs of the observed and the predicted distributions as test statistic. For each participant, the 1,000 predicted signal-respond RT distributions were compared to the observed signal-respond RT distribution, using posterior predictive p values and visual inspection of the distributions.

Figure 3.15 shows the observed go RT and signal-respond RT distributions, and 100 randomly chosen predicted signal-respond RT distributions for six participants with satisfactory model fit. The predicted signal-respond RT distributions (i.e., gray lines) adequately followed the shape of the observed signal-respond RT distribution. Also, the predicted signal-respond RTs were generally faster than the observed go RTs (i.e., dashed line), a common finding that follows from the architecture of the horse race model (Logan & Cowan, 1984). Lastly, the average of the medians of the predicted signal-respond distributions closely matched the observed median. This result is also evident from the posterior predictive p values listed in the second column of Table 3.1. The posterior predictive p values for these six participants are well within the 0.05-0.95 range, indicating that the BPA adequately accounted for the median of the observed signal-respond RTs.

Figure 3.16 shows the observed go RT and signal-respond RT distributions, and 100 randomly chosen predicted signal-respond RT distributions for three participants with unsatisfactory model fit. Note that Participant 16 and Participant 20 were among the few cases that produced un-

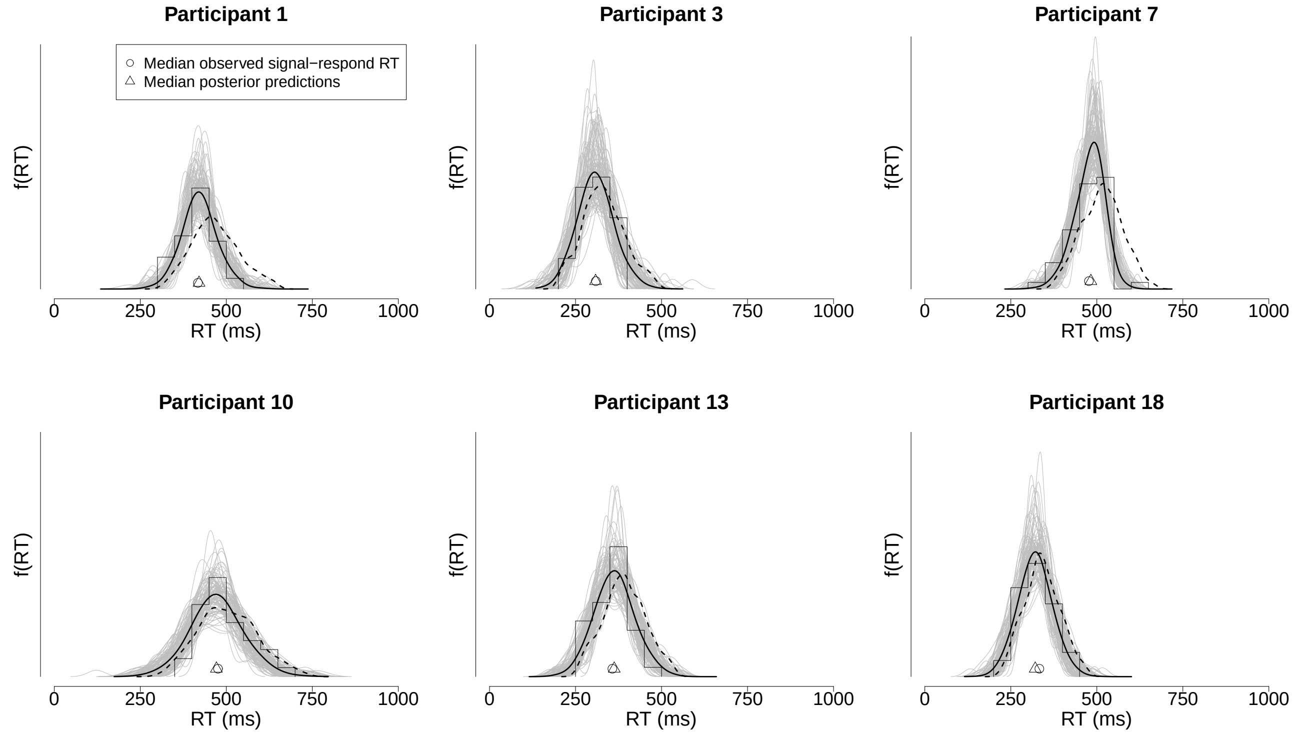


Figure 3.15 *Examples of satisfactory model fit with the individual Bayesian parametric approach (BPA): Predicted and observed signal-respond response time (RT) distributions for the first session of the Bissett and Logan (2011) experiment.* See text for a detailed description of the posterior predictive analyses. The histogram shows the observed signal-respond RT distribution. The gray lines show 100 randomly chosen predicted signal-respond RT distributions. The solid black line gives a predicted signal-respond RT distributions based on the mean of the posterior predictions. The circle indicates the median of the observed signal-respond RTs. The triangle indicates the median of the predicted signal-respond RTs. The median of the predicted signal-respond RTs is computed as the mean of the medians of the predicted signal-respond RT distributions. The dashed line shows the observed go RT distribution.

informative posterior distributions. The location and shape of the observed signal-respond RT distributions were not well approximated by the predicted signal-respond RT distributions. For Participant 8 and Participant 16, the predicted signal-respond RTs are shifted to the right. For Participant 20, the predicted signal-respond RT distribution fails to capture the bimodality of the observed signal-respond RTs. For Participant 16 and 20, the predicted signal-respond RTs are less variable than the observed signal-respond RTs. Moreover, the predicted signal-respond RTs are not substantially faster than the observed go RTs. For Participant 8 and Participant 16, the median of the predicted signal-respond RTs overestimated the median of the observed signal-respond RTs. In contrast, for Participant 20, the median of the predicted signal-respond RTs underestimated the observed median. These latter results are also shown in Table 3.1. The posterior predictive p values for these three participants are very close to or are equal to 0 or 1, indicating that the BPA failed to account for the median of the observed signal-respond RTs.

In the second posterior predictive analysis, we compared the observed response rates to the

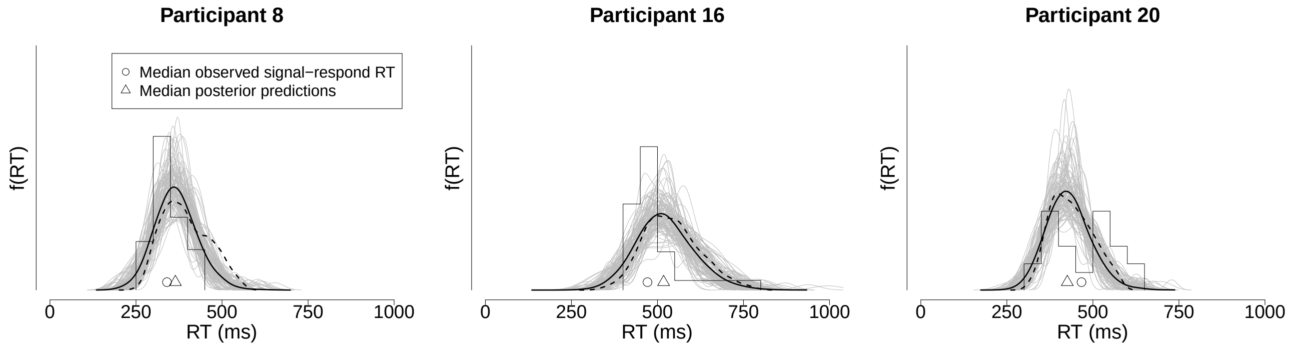


Figure 3.16 *Examples of unsatisfactory model fit with the individual Bayesian parametric approach (BPA): Predicted and observed signal-respond response time (RT) distributions for the first session of the Bissett and Logan (2011) experiment.* See Figure 3.15 for details.

Table 3.1 Posterior Predictive p Values for the Median of the Signal-Respond Response Time Distribution and the Response Rate for the First Session of the Bissett and Logan (2011) Experiment Computed From the Parameter Estimates From the Individual Bayesian Parametric Approach.

Participant	p value median	Minimum p value RR	Maximum p value RR
1	0.64	0.11	0.91
3	0.44	0.10	0.79
7	0.77	0.13	0.70
8	**0.98**	**0.03**	0.95
10	0.41	0.28	0.86
13	0.69	0.08	0.92
16	**1.00**	0.30	**0.96**
18	0.11	0.25	0.61
20	**0.00**	0.33	**0.96**

Note. Posterior predictive p values that indicate unsatisfactory model fit are shown in bold. p value median = posterior predictive p value for the median of the signal-respond response time distribution on the stop-signal delay (SSD) with the highest number of observed signal-respond trials; minimum p value RR = the lowest posterior predictive p value for the response rate (RR) computed for the SSDs that contained at least 10% of the trials; maximum p value RR = the highest posterior predictive p value for the RR computed for the SSDs that contained at least 10% of the trials.

response rates predicted by the posterior distribution of the model parameters. The model check was performed for the SSDs that featured at least 10% of the total number of stop-signal trials. For each participant, we generated 1,000 stop-signal data sets using the 1,000 parameter vectors selected for the first posterior predictive analysis, the chosen SSDs and the corresponding number of stop-signal trials. We computed posterior predictive p values for each participant on each SSD separately, where we used the observed and predicted response rates as test statistic.

Table 3.1 shows the minimum and the maximum of the posterior predictive p values for the response rates across the various SSDs of nine participants. For Participant 1, 3, 7, 10, 13 and 18, the minimum and maximum of the p values all lie between 0.05 and 0.95, corroborating our previous conclusion of satisfactory model fit with the median of the signal-respond RTs. In contrast, for Participant 8, 16, and 20, the minimum or maximum of the p values are very close to 0 or 1, supporting our previous finding that the BPA failed to account for the data of these participants.

In sum, the results of the posterior predictive model checks indicated that for most participants the BPA provided plausible parameter estimates that adequately describe the observed data. Additionally, the posterior predictions supported our earlier conclusion that the BPA resulted in uninterpretable parameter estimates for participants with imprecise posterior distributions. The results of the posterior predictive model checks for the remaining participants in the first as well as the second session of the Bissett and Logan (2011) experiment are available in the supplemental materials at `http://dora.erbe-matzke.com/publications.html`.

Hierarchical BPA

The hierarchical approach has the potential to provide accurate parameter estimates with relatively few observations per participant. We therefore did not analyze the complete Bissett and Logan (2011) data set, but used only a subsample of the available go RTs and signal-respond RTs from the second experimental session. Per participant, we fit a randomly selected 90 go RTs, 30 signal-respond RTs, and 30 successful inhibitions after removing the outliers.

The hierarchical BPA was fit to the data with WinBUGS. The estimated individual μ_{stop}, σ_{stop}, and τ_{stop} parameters were constrained to be equal across the different SSDs. We ran three MCMC chains and used overdispersed starting values. The hierarchical analysis was based on 3 × 23,750 samples.

The hierarchical BPA resulted in informative posterior distributions for the group-level parameters of the go RT as well as the SSRT distribution. As before, this section focuses exclusively on the parameters of the SSRT distribution. Figure 3.17 shows the prior and the posterior distributions of the group-level stop parameters. As for the hierarchical recovery study, the $\mu_{\mu_{stop}}$ and $\sigma_{\mu_{stop}}$ parameters were estimated the most precisely as indicated by the small posterior standard deviations. Also in line with the simulations, the $\sigma_{\sigma_{stop}}$ parameter was estimated with the largest posterior uncertainty. Nevertheless, the group-level stop parameters were estimated relatively well given the scarce data and the small sample size.

For most participants, the individual μ_{stop}, σ_{stop}, and τ_{stop} parameters were estimated adequately as evidenced by the well-behaved posterior distributions. For the same subsample of the data, the posterior distributions estimated with the hierarchical BPA were less variable than the posteriors estimated with the individual BPA. In fact, the 60 stop-signal trials were occasionally insufficient to obtain informative posterior distributions with the individual BPA. Figure 3.18 illustrates the benefits of hierarchical modeling for a representative participant. For the same subsample of the data, the 95% Bayesian confidence intervals are smaller for the posterior distributions estimated with the hierarchical BPA (i.e., gray line) than for the posteriors estimated with the individual BPA (i.e., black line). Also, the posterior medians from the hierarchical analysis

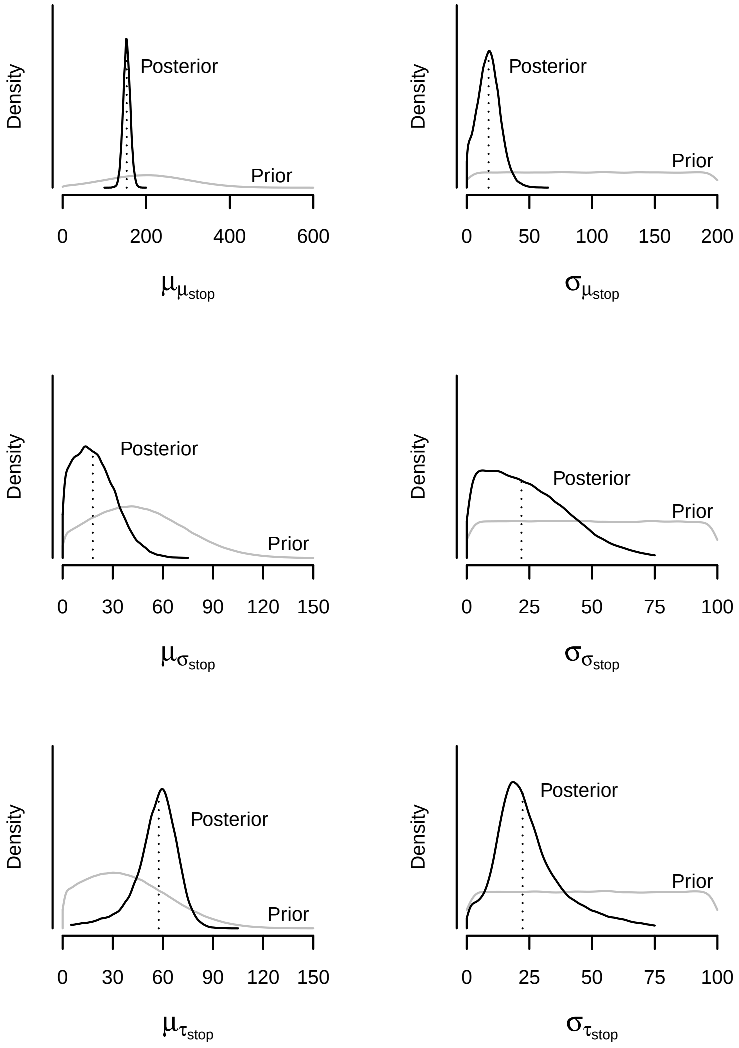


Figure 3.17 *Posterior distributions for the group-level stop parameters from a subsample of the Bissett and Logan (2011) data set.* The black lines show the posterior distributions and the gray lines show the prior distributions of the group-level parameters. The dashed lines give the posterior medians.

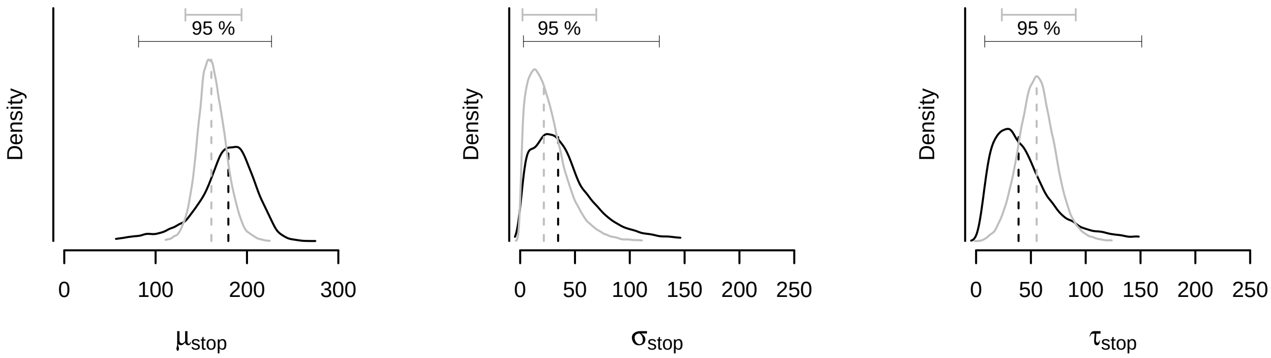


Figure 3.18 *Posterior distribution of the stop parameters for Participant 1 from the Bissett and Logan (2011) data set estimated using the individual and the hierarchical Bayesian parametric approach (BPA).* The solid black and gray lines show the posterior distribution of the stop parameters and the corresponding 95% Bayesian confidence intervals obtained with the individual and the hierarchical BPA, respectively. The dashed black and gray lines show the median of the posterior distributions obtained with the individual and the hierarchical BPA, respectively.

are slightly pulled towards their corresponding group mean, a typical consequence of hierarchical modeling.

As pointed out above, hierarchical Bayesian estimation has the potential to reduce the variability in the estimated parameters compared to individual-level estimation. Figure 3.19 compares the mean SSRTs computed with the traditional mean method, the posterior medians from the individual BPA, and the posterior medians from the hierarchical BPA with the same subsample of the data. As shown in Figure 3.19a, the individual BPA provided mean SSRTs that are slightly less variable than the mean SSRTs obtained using the mean method. More importantly, Figure 3.19b and Figure 3.19c show that the hierarchical BPA resulted in mean SSRTs that are substantially less variable than the mean SSRTs obtained with either the mean method or the individual BPA.

Assessing Model Fit Using Posterior Predictive Model Checks

We used posterior predictive model checks to determine whether the individual parameter estimates from the hierarchical BPA adequately describe the Bissett and Logan (2011) data. As the data of most participants featured fewer than 10 observed signal-respond RTs even on the SSD with the highest number of observations, posterior predictive model checks using the signal-respond RT distributions are not sensible for the present data set. For each participant, we compared the observed response rates to the response rates predicted by the posterior distribution of the model parameters. The posterior predictive analyses followed the procedure described for the individual BPA, using $N = 1,000$ parameter vectors from the joint posterior of the individual μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, and τ_{stop} parameters.

Table 3.2 shows the minimum and the maximum of the posterior predictive p values for the response rates across the various SSDs. With the exception of Participant 12 and 19, the minimum and maximum value of the posterior predictive p values all lie well within the 0.05 - 0.95 range,

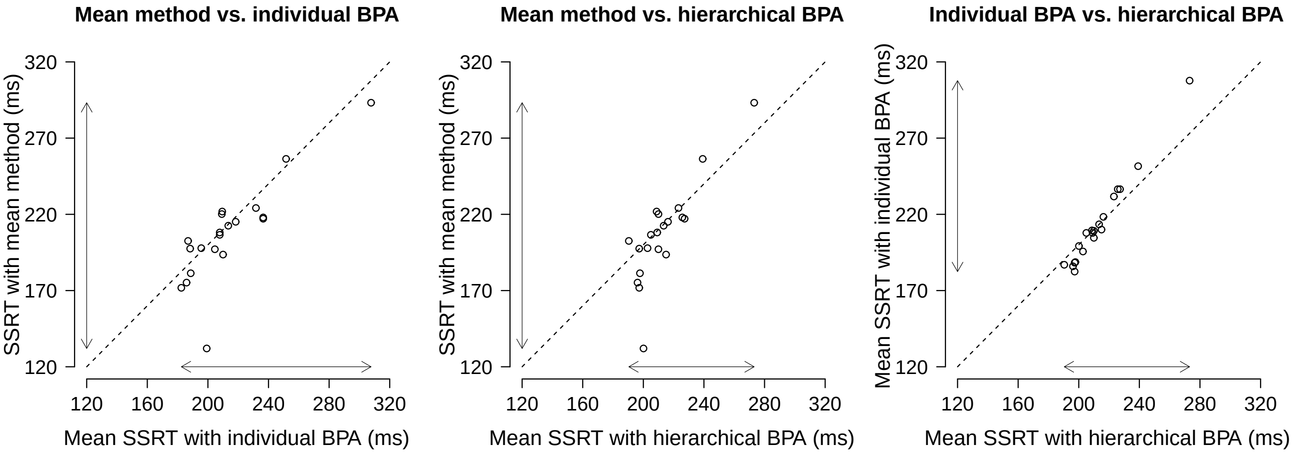


Figure 3.19 *Comparison of mean stop-signal reaction times (SSRTs) obtained with the mean method, with the posterior medians of the stop parameters from the individual Bayesian parametric approach (BPA), and with the posterior medians of the individual stop parameters from the hierarchical BPA in the Bissett and Logan (2011) data set.* The arrows indicate the range of the estimates. Note that the range is smallest for the hierarchical BPA.

indicating that the BPA adequately accounted for the response rates of most participants.

The above section illustrated that the hierarchical BPA provided sensible group-level parameters with relatively well-calibrated posterior distributions even with a small sample size and only 60 stop-signal trials per participant. Posterior predictive model checks indicated that for most participants the hierarchical BPA provided plausible individual parameter estimates that adequately describe the observed data. Moreover, the individual parameter estimates yielded sound mean SSRTs and demonstrated the characteristic benefits of hierarchical modeling.

3.6 Discussion

The stop-signal task is a frequently used experimental measure of response inhibition. Over the past 30 years, the horse race model (Logan, 1981; Logan & Cowan, 1984) has successfully accounted for stop-signal data in different settings and has facilitated the interpretation of stopping experiments with various age groups and clinical populations (e.g., Kramer et al., 1994; Oosterlaan et al., 1998; Ridderinkhof et al., 1999; Schachar & Logan, 1990; Schachar et al., 2000; Williams et al., 1999). The horse race model offers numerous methods to estimate the otherwise unobservable latency of stopping.

The existing methods to estimate SSRT are unable to adequately estimate the shape of entire SSRT distributions. Ignoring the shape of SSRT distributions, and focusing only on the mean SSRT, may mask crucial features of the data and result in erroneous conclusions about the nature of response inhibition. The goal of this paper was therefore to introduce a novel method —a Bayesian parametric approach— that enables researchers to estimate the entire distribution of SSRTs. The BPA is based on the assumptions of the horse race model and treats response inhibition as a censoring mechanism. The method assumes that go RTs and SSRTs are ex-Gaussian distributed and relies on MCMC sampling to obtain posterior distributions for the model parameters. Note

Table 3.2 Posterior Predictive p Values for the Response Rate for the Second Session of the Bissett and Logan (2011) Experiment Computed From the Individual Parameter Estimates From the Hierarchical Bayesian Parametric Approach.

Participant	Minimum p value RR	Maximum p value RR
1	0.31	0.51
2	0.40	0.61
3	0.14	0.56
4	0.36	0.64
5	0.17	0.54
6	0.16	0.40
7	0.14	0.69
8	0.15	0.35
9	0.34	0.65
10	0.27	0.60
12	**0.01**	0.30
13	0.38	0.59
15	0.16	0.52
16	0.33	0.72
18	0.10	0.32
19	**0.04**	0.64
20	0.28	0.55
21	0.23	0.67
22	0.19	0.30
23	0.26	0.36

Note. Posterior predictive p values that indicate unsatisfactory model fit are shown in bold. Minimum p value RR = the lowest posterior predictive p value for the response rate (RR) computed for the stop-signal delays (SSDs) that contained at least 10% of the trials; maximum p value RR = the highest posterior predictive p value for the RR computed for the SSDs that contained at least 10% of the trials.

that we could have carried out parameter estimation by means of standard maximum likelihood estimation (Dolan et al., 2002; Myung, 2003). However, our goal was to develop a method for estimating SSRT distributions that may be applied to individual as well as hierarchical data structures. As maximum likelihood estimation can become practically difficult for hierarchical models, we chose to use Bayesian parameter estimation instead. This also brings along the typical benefits of Bayesian estimation, such as easy-to-use estimation software (e.g., WinBUGS) and a coherent inferential framework.

We demonstrated using simulations that the BPA adequately recovers the parameters of the generating SSRT distributions in individual and hierarchical data structures. We showed that the individual BPA can provide accurate estimates of SSRT distributions in experimental stop-signal data featuring a realistic number of trials. Similarly, we demonstrated using real data that the hierarchical BPA resulted in interpretable individual and group-level stop parameters with a small sample size and as few as 60 stop-signal trials per participant.

The BPA enables researchers to evaluate differences in the shape of SSRT distributions between clinical or experimental groups. SSRT distributions obtained from the individual BPA can be compared visually, as illustrated in the introduction and with the Bissett and Logan (2011) data set. Similarly, the group-level go and stop parameters obtained from the hierarchical BPA may be compared visually by inspecting the overlap —or the lack of overlap— between the posterior distribution of the parameters in the different groups. Alternatively, differences in the ex-Gaussian individual go and stop parameters between clinical groups or experimental conditions may be evaluated formally using Bayesian hypothesis testing. Various user-friendly options are now available to perform, for instance, Bayesian t tests (Rouder et al., 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009) and analyses of variance (Masson, 2011; Wetzels et al., 2012) using Bayes factors (e.g., Berger & Pericchi, 1996; Dickey, 1971; Gamerman & Lopes, 2006; Kass & Raftery, 1995; Klugkist, Laudy, & Hoijtink, 2005; O'Hagan & Forster, 2004). Moreover, easy-to-use Bayesian hypothesis tests for correlations (Wetzels & Wagenmakers, 2012) and regression analyses (Liang, Paulo, Molina, Clyde, & Berger, 2008)[5] are also available.

Although the BPA offers the advantages of estimating entire SSRT distributions, it also comes with a number costs. One drawback is related to the amount of data that is required to obtain precise stop parameter estimates. The individual BPA may fail to provide precise estimates with the very small amount of data that is sometimes collected in developmental and clinical stop-signal studies. For a number of participants, the sampled 150 trials (i.e., 90 go and 60 stop-signal trials) of the Bissett and Logan (2011) data set were in fact insufficient to obtain informative posterior distributions for the stop parameters using the individual BPA. Nevertheless, we illustrated that the hierarchical BPA may provide a solution in such situations.

Another drawback of the BPA is related to the present implementation in WinBUGS. First, the BPA requires some basic programming skills to obtain the necessary data format for the WinBUGS script. Second, the fitting algorithm is time-consuming. Running on a fast personal computer, WinBUGS required an average of about 5 hr to reach convergence per participant in the Bissett and Logan (2011) experiment. Likewise, it took several days to fit the hierarchical BPA to the subsample of the Bissett and Logan data set. We are currently working on a user-friendly implementation of the BPA that will increase the speed of the fitting routine.

Parametric Assumptions

In contrast to the Colonius method (1990, see also de Jong et al., 1990) for estimating SSRT distributions, the BPA requires a parametric form to describe the go RTs and the SSRTs. In the abstract sense, as the Colonius method does not require any assumptions about the distribution of the go RTs and the SSRTs, it may be preferable to the BPA. In the practical sense, however, the applicability of the Colonius method is limited by the amount of data that is available per participant. In contrast to the BPA, the Colonius method requires an unrealistically large amount of data to perform adequately. For the analysis of experimental stop-signal data, the BPA is therefore preferable to the Colonius method. Of course, the practical advantage of the BPA comes at a price: We need parametric assumptions to quantify the shape of the go RT and the SSRT distributions.

Here we assumed that the go RTs and the SSRTs are ex-Gaussian distributed. Note, however, that the BPA does not hinge on the particular parametric form used to summarize the distributions. The ex-Gaussian distribution is used as a convenient choice to quantify the go RTs and the SSRTs. The ex-Gaussian is a frequently used distributional model that can excellently accommodate the shape of RT distributions and is easy to fit to data (Heathcote et al., 1991; Hockley, 1982, 1984; Ratcliff, 1978, 1993; Ratcliff & Murdock, 1976). Moreover, sensitivity analyses indicated that the BPA is robust to misspecification of the parametric form of the go RT and SSRT distributions. Even when the go RTs and the SSRTs were drawn from shifted log-normal distributions, the ex-Gaussian based BPA excellently approximated the shape of their distribution.

[5] For software implementation, see `http://pcl.missouri.edu/bf-reg`. Retrieved May 23, 2012.

The above result is not surprising because the ex-Gaussian distribution is flexible enough to accommodate a wide range of distributions observed in RT data. Unless the go RTs and SSRTs are left skewed or bimodal —an unlikely scenario for RT distributions— the ex-Gaussian is likely to provide adequate description of their distributions. The interested reader is referred to the supplemental materials for the detailed results of the sensitivity analyses for the individual BPA. Note also that the posterior predictive model checks indicated that the ex-Gaussian distribution provided an excellent description of the go RTs and SSRTs in experimental stop-signal data.

Nevertheless, the ex-Gaussian distribution comes also with some disadvantages. Specifically, the ex-Gaussian has a number of characteristics that are atypical of empirical RT data. First, the ex-Gaussian has a monotonically increasing hazard function, while empirical hazard functions are typically peaked (e.g., W. Schwarz, 2001; Van Zandt, 2002). Second, the ex-Gaussian distribution assigns probability to unrealistically short and negative RTs. As an alternative, one may use "shifted" RT distributions with a parameter-dependent lower bound. For example, the ex-Gaussian distribution may be replaced by the shifted Wald, the shifted Weibull or the shifted log-normal distributions (e.g., Heathcote, 2004; Heathcote, Brown, & Cousineau, 2004). However, shifted distributions are notoriously difficult to fit. Moreover, in our implementation the shifted log-normal distribution resulted in somewhat less accurate estimates than the ex-Gaussian. Another alternative is to use the ex-Wald distribution (W. Schwarz, 2001) to describe the go RTs and the SSRTs. Heathcote (2004) showed, however, that the ex-Wald requires at least 400 observations to produce adequate parameter estimates, a requirement that is often not satisfied in stop-signal experiments.

Process Models

Process models of response inhibition provide further possibilities to model performance in the stop-signal paradigm. A prominent alternative to the BPA is the interactive race model (Boucher, Palmeri, Logan, & Schall, 2007), a neurally plausible instantiation of the horse race model. The interactive race model conceptualizes the go and the stop process as two noisy accumulators that race towards a fixed response threshold and may interact via inhibitory links. The interactive race model assumes constant rates of rise to the threshold and noise terms with standard deviations σ_{go} and σ_{stop} that reflect the amount of noise added in each step of the rise. Boucher et al. (2007) showed that the go and the stop process are for the most part independent. The inhibitory effect of the stop process on the go process is very brief and is much stronger than the inhibitory effect of the go process on the stop process. Note that the interactive race model applies specifically to saccadic inhibition (Verbruggen et al., 2008).

Another alternative is the Hanes-Carpenter model (Hanes & Carpenter, 1999; Hanes & Schall, 1995; Hanes, Patterson, & Schall, 1998) for saccade inhibition. The Hanes-Carpenter model is based on the Linear Approach to Threshold with Ergodic Rate (LATER; Carpenter, 1981; Carpenter & Williams, 1995). The model assumes that the competing go and the stop process rise in a linear fashion to a fixed response threshold. If the stop process reaches the threshold before the go process, the response is inhibited. If the go process reaches the threshold before the stop process, the response is executed.

The Hanes-Carpenter model is equivalent to the horse race model with specific distributional assumptions about the rate of information accumulation of the go and the stop process (Colonius, Özyurt, & Arndt, 2001). Specifically, the Hanes-Carpenter model assumes that the rates of rise are normally distributed, resulting in the following parameters: the means and the standard deviations of the rates of rise of the go and the stop process, μ_{go}, σ_{go}, μ_{stop}, and σ_{stop}, respectively. The model parameters can be estimated using Monte Carlo simulations (e.g., Asrress & Carpenter, 2001; Colonius et al., 2001; Hanes & Carpenter, 1999) or maximum likelihood estimation (e.g., Corneil & Elsley, 2005; Kornylo, Dill, Saenz, & Krauzlis, 2003). Using the properties of the stop process accumulator, one can obtain the distribution of the finishing times of the stop process (SSRT distribution). Note, however, that in typical applications of LATER, the goal is to use the estimated rate parameters to test and compare the predictions of competing models of response inhibition and not to explicitly estimate SSRT distributions. For yet another alternative to model inhibitory control in the stop-signal task, see Shenoy, Rao, and Yu (2010) and Shenoy and Yu (2011).

LATER and the BPA constitute different perspectives on modeling response inhibition. LATER is focused on the nature of the (neural) processes underlying response inhibition and thereby makes particular assumptions of the shape for the finishing time distribution of the stop processes. In contrast, the BPA constitutes a more statistical approach. The BPA is not concerned with the nature of the underlying go and stop process; it rather focuses on how the SSRT distribution can be estimated irrespective of the particular parametric choice —be it ex-Gaussian or shifted Wald — used to quantify its shape.

Prior Distributions

The BPA uses Bayesian parameter estimation and therefore involves choosing prior distributions for the ex-Gaussian go and stop parameters. With respect to the individual BPA, the priors for the go and stop parameters are informative in the sense that they cover a wide but realistic range of values informed by results from the stop-signal literature (Williams et al., 1999; Band et al., 2003). We feel that using informative priors is justified since there is a large body of past research that provides valuable information about the plausible range of parameter values. Also, with increasing opportunity to apply the BPA to empirical data sets, we will be able to make even better informed choices about the prior distribution of the parameters. Note also that as long as sufficiently informative data are available, the data readily overwhelm the prior (e.g., M. D. Lee & Wagenmakers, 2013). Whereas Bayesian parameter estimation can be robust to changes in priors, Bayesian hypothesis testing using Bayes factors (e.g., Berger & Pericchi, 1996; Dickey, 1971; Gamerman & Lopes, 2006; Kass & Raftery, 1995; Klugkist et al., 2005; O'Hagan & Forster, 2004) can be relatively sensitive to prior inputs. The shape of the prior distribution can greatly influence the Bayes factor and the resulting inferences (e.g., Bartlett, 1957; Liu & Aitkin, 2008; but see Vanpaemel, 2010). Fortunately, various user-friendly approaches to Bayesian hypothesis testing are now available that rely on principled choices of prior distributions (e.g., Rouder et al., 2009; Wetzels et al., 2009).

With respect to the hierarchical approach, the BPA assumes that the individual go and stop parameters come from truncated normal group-level distributions. The use of normal group-level distributions is a common choice in Bayesian hierarchical modeling (e.g., Gelman & Hill, 2007; M. D. Lee & Wagenmakers, 2013). Also, sensitivity analyses indicated that the hierarchical BPA is relatively robust to misspecification of the group-level distribution of the individual go parameters. Even when the true go parameters were drawn from uniform or bimodal group-level distributions, the hierarchical BPA with truncated normal group-level distributions provided accurate individual go parameter estimates. Unfortunately, the hierarchical BPA is less robust to misspecification of the group-level distribution of the individual stop parameters. When the true stop parameters were drawn from uniform or bimodal group-level distributions, the hierarchical BPA with truncated normal group-level distributions resulted in biased parameter estimates, particularly for the σ_{stop} and τ_{stop} parameters. Fortunately, the bias decreased as the number of participants and especially as the number of trials increased. The finding that the go parameters are more robust to misspecification of the group-level distributions is not surprising. The go parameters are estimated based on the go RTs as well as the signal-respond RTs. Also, the sensitivity analyses —similar to typical stop-signal studies— featured three times as many go trials as stop-signal trials. As a result, the go parameters are more strongly constrained by the data and are less strongly influenced by their group-level distribution than the stop parameters.

Moreover, the sensitivity analyses indicated that misspecification of the group-level distributions often results in convergence problems. We therefore recommend researchers to carefully monitor the convergence of the individual parameter estimates. If there are reasons to suspect that the hierarchical assumptions are violated, we advise users to inspect the distribution of the individual go and stop parameters obtained either from the individual BPA or from the hierarchical BPA with very weak priors for the group-level parameters. If these preliminary analyses indicate that the distribution of the individual parameters substantially deviates from normality, one may use the unconstrained individual go and stop parameters. Alternatively, if substantive knowledge of the form of the group-level distributions is available, the hierarchical BPA may be adapted to accommodate the desired (mixture) distribution. The reader is referred to the supplemental materials for the detailed results of the sensitivity analyses for the hierarchical BPA.

Conclusion

Here we introduced a novel Bayesian parametric method that provides for the estimation of entire distribution of SSRTs. The new method enables researchers to evaluate differences in the shapes of SSRT distributions between various clinical populations or experimental groups. In doing so, our Bayesian parametric approach aids the interpretation of stop-signal data and may reveal some hitherto unknown aspects of response inhibition.

Chapter 4

Release the BEESTS: Bayesian Estimation of Ex-Gaussian Stop-Signal Reaction Time Distributions

This chapter has been published as:
Dora Matzke, Jonathon Love, Thomas Wiecki, Scott D. Brown, Gordon D. Logan, and Eric-Jan Wagenmakers (2013).
Release the BEESTS: Bayesian estimation of ex-Gaussian stop-signal reaction time distributions.
Frontiers in Quantitative Psychology and Measurement, 4:918, doi: 10.3389/fpsyg.2013.00918.

Abstract

The stop-signal paradigm is frequently used to study response inhibition. In this paradigm, participants perform a two-choice response time task where the primary task is occasionally interrupted by a stop-signal that prompts participants to withhold their response. The primary goal is to estimate the latency of the unobservable stop response (stop signal reaction time or SSRT). Recently, Matzke, Dolan, Logan, Brown, and Wagenmakers (2013) have developed a Bayesian parametric approach that allows for the estimation of the entire distribution of SSRTs. The Bayesian parametric approach assumes that SSRTs are ex-Gaussian distributed and uses Markov chain Monte Carlo sampling to estimate the parameters of the SSRT distribution. Here we present an efficient and user-friendly software implementation of the Bayesian parametric approach —BEESTS— that can be applied to individual as well as hierarchical stop-signal data. BEESTS comes with an easy-to-use graphical user interface and provides users with summary statistics of the posterior distribution of the parameters as well various diagnostic tools to assess the quality of the parameter estimates. The software is open source and runs on Windows and OS X operating systems. In sum, BEESTS allows experimental and clinical psychologists to estimate entire distributions of SSRTs and hence facilitates the more rigorous analysis of stop-signal data.

Response inhibition —the ability to stop an ongoing response— is frequently studied using the stop-signal paradigm. In the stop-signal paradigm (Lappin & Eriksen, 1966; Logan & Cowan, 1984), participants perform a two-choice visual response time (RT) task, such as responding to the color or the shape of the stimuli. This primary task is occasionally interrupted by a stop-signal that instructs participants not to respond on that trial. The goal is to estimate the latency of the unobservable stop response (stop-signal RT; SSRT).

Based on the independent horse-race model (Logan, 1981; Logan & Cowan, 1984), various methods are available to estimate SSRTs (e.g., Logan, 1994; Verbruggen & Logan, 2009; Verbruggen, Chambers, & Logan, 2009). Over the past decades, the horse-race model has been extensively used to estimate stopping latencies and compare the efficiency of response inhibition between different age groups (e.g., Kramer et al., 1994; Ridderinkhof et al., 1999; Schachar & Logan, 1990; Williams et al., 1999) and clinical populations (Oosterlaan et al., 1998; Schachar & Logan, 1990; Schachar et al., 2000). Unfortunately, most standard methods to estimate SSRTs only provide a summary measure of the latency of the stop process, such as the mean or the median SSRT.

Several researchers have argued, however, that the adequate analysis of RT data should not only focus on mean RT, but should consider the shape of the entire RT distribution (e.g., Heathcote et al., 1991; Matzke & Wagenmakers, 2009). The shape of SSRT distributions may, for example, differ between different clinical populations, without necessary differences in mean SSRT. Ignoring the shape of SSRT distributions may thus lead to incorrect conclusions about differences in response inhibition between groups.

To allow for a more thorough analysis of stop-signal data, Matzke et al. (2013) have recently developed a Bayesian parametric approach (BPA) that enables researchers to estimate the entire distribution of SSRTs (see Logan, Van Zandt, Verbruggen, and Wagenmakers, 2014; for an alternative approach). The BPA assumes that SSRTs follow an ex-Gaussian distribution and uses Bayesian parameter estimation to obtain posterior distributions for the model parameters. The BPA allows researchers to compare and evaluate differences in the ex-Gaussian stop parameters between experimental and clinical groups. By doing so, the BPA has the potential to facilitate the interpretation of stop-signal data and contribute to new insights on the nature of response inhibition.

Parameter estimation in the BPA currently relies on the popular Bayesian statistical package WinBUGS (Bayesian inference Using Gibbs Sampling for Windows; Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012). The practical usefulness of the BPA is, however, severely limited by the disadvantages of the present implementation. The WinBUGS routine is extremely time consuming and rather user-unfriendly. For instance, WinBUGS requires several hours to produce reliable parameter estimates for a single participant and it requires several days to fit a hierarchical data set. It is therefore all but impossible for experimental and clinical psychologists to take advantage of the theoretical progress offered by the BPA.

In order to overcome this obstacle and promote the broader application of the Bayesian analysis of stop-signal data, we introduce a relatively fast, user-friendly software that allows for the estimation of entire SSRT distributions. BEESTS (**B**ayesian **E**x-Gaussian **E**stimation of **ST**op-**S**ignal RT distributions) can be applied to individual and hierarchical stop-signal data and comes with an easy-to-use graphical user interface. BEESTS provides users with summary statistics of the posterior distribution of the parameters as well as various diagnostic tools to assess the quality of the parameter estimates.

The outline of the paper is as follows. First, we describe the Bayesian parametric approach in more detail. Second, we introduce BEESTS, present the installation instructions, and describe the various analysis and output options provided by the software. Third, we illustrate the use of BEESTS with experimental stop-signal data. The last section concludes.

4.1 The Bayesian Parametric Approach

Rationale and Assumptions

According to the standard horse-race model (Logan, 1981; Logan & Cowan, 1984), performance in the stop-signal paradigm can be conceptualized as a horse-race between two independent processes that compete against each other: a go-process that is initiated by the primary task "go" stimulus and a stop-process that is generated by the stop-signal. As shown in Figure 4.1, if the go-process finishes before the stop-process, the primary response is executed; if the stop-process finishes before the go-process, the primary response is inhibited. The shorter the time interval between the onset of the go-stimulus and the onset of the stop-signal (i.e., stop-signal delay; SSD), the more likely participants are to inhibit their response on the primary task (see also Matzke et al., 2013).

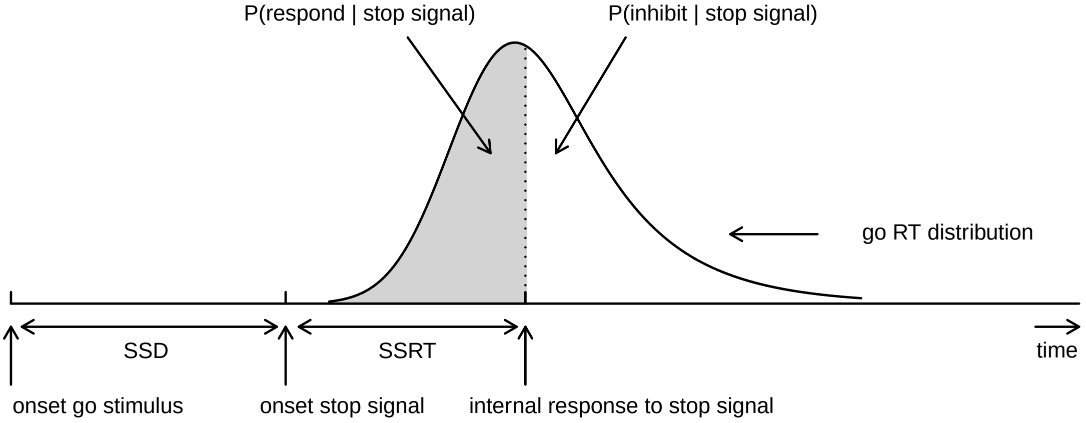


Figure 4.1 *Graphical representation of the independent horse-race model.* The success of response inhibition is determined by the relative finishing times of the go and the stop process. Primary task "go" RTs that are longer than SSD + SSRT are successfully inhibited (i.e., white area); go RTs that are shorter than SSD + SSRT escape inhibition and result in signal-respond RTs (i.e., gray area; see also Matzke et al., 2013). Constant SSRT is assumed.

The Bayesian parametric approach (BPA) is based on the rationale of the standard horse-race model, but it assumes that primary task "go" RTs and SSRTs are both independent random variables (i.e., complete horse-race model). As shown in Figure 4.2, the BPA assumes that the distribution of RTs that escape inhibition (i.e., signal-respond RTs) can be viewed as a censored go RT distribution. The censoring point is assumed to be drawn from the SSRT distribution and can take on a different value on each stop-signal trial (e.g., $SSD + SSRT_1$, $SSD + SSRT_2$, and $SSD + SSRT_3$). The estimation of the SSRT distribution therefore involves simultaneously estimating the parameters of the go RT distribution and its censoring distribution (see also Matzke et al., 2013).

The BPA assumes that the go RTs and SSRTs are ex-Gaussian distributed (Heathcote et al., 1991; Hockley, 1982, 1984; Matzke & Wagenmakers, 2009; Ratcliff, 1978, 1993; Ratcliff & Murdock,

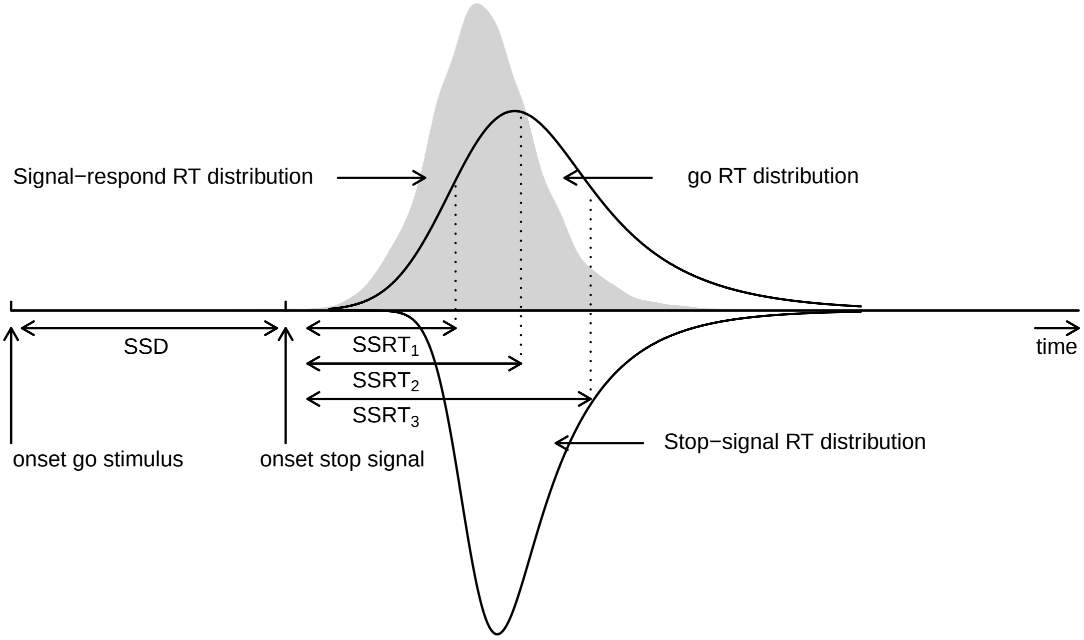


Figure 4.2 *Assumptions of the Bayesian parametric approach.* The BPA treats the distribution of signal-respond RTs (i.e., gray area) as a go RT distribution that is censored by the SSRT distribution. The censoring point can take on a different value on each stop-signal trial (e.g., $SSD + SSRT_1$, $SSD + SSRT_2$, and $SSD + SSRT_3$). If the go RT on a given trial is longer than $SSD + SSRT$, the go RT is successfully inhibited. In contrast, if the go RT on a given trial is shorter than $SSD+SSRT$, the go RT cannot be inhibited and results in a signal-respond RT. See Matzke et al. (2013) for details.

1976). The ex-Gaussian is a three-parameter distribution that is given by the convolution of a Gaussian and an exponential distribution. The μ and σ parameters are the mean and the standard deviation of the Gaussian component, respectively, and τ is the mean of the exponential component. The μ and σ parameters reflect the leading edge and mode of the distribution, whereas τ reflects the tail of the distribution. As shown in Figure 4.3, the ex-Gaussian is a positively skewed unimodal distribution that can excellently accommodate the shape of empirical RT data.

The probability density function of the ex-Gaussian is

$$f(t;\mu,\sigma,\tau) = \frac{1}{\tau}\exp\left(\frac{\mu-t}{\tau}+\frac{\sigma^2}{2\tau^2}\right)\Phi\left(\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}\right) \text{for } \sigma>0,\ \tau>0, \tag{4.1}$$

where Φ is the standard normal distribution function, given by

$$\Phi\left(\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{t-\mu}{\sigma}-\frac{\sigma}{\tau}} \exp\left(\frac{-y^2}{2}\right)dy. \tag{4.2}$$

The mean and variance of the ex-Gaussian distribution equals

$$\mathrm{E(t)} = \mu+\tau \tag{4.3}$$

and

$$\mathrm{Var(t)} = \sigma^2 + \tau^2, \tag{4.4}$$

respectively. Note that the BPA does not assume that the ex-Gaussian parameters correspond to specific cognitive processes (Matzke & Wagenmakers, 2009); the ex-Gaussian distribution is used as a convenient descriptive model to summarize the distribution of go RTs and SSRTs. As an alternative, one may use, for instance, the ex-Wald distribution (W. Schwarz, 2001), or "shifted" RT distributions with a parameter–dependent lower bound, such as the shifted Wald, the shifted Weibull or the shifted log normal distribution (e.g., Heathcote, 2004; Heathcote et al., 2004; Rouder, 2005; Rouder et al., 2005; see also Luce, 1986 for alternatives.)

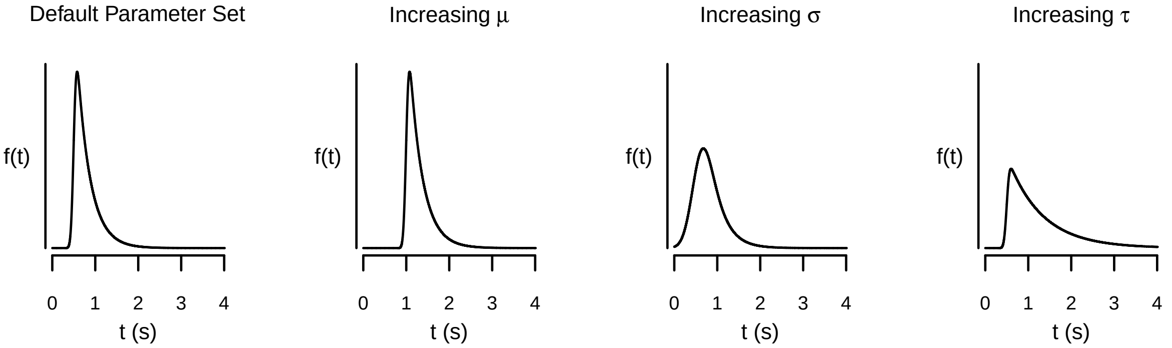


Figure 4.3 *The shape of the ex-Gaussian distribution as a function of the μ, σ, and τ parameters.* The distributions were generated with the following parameter sets: $\mu = 0.5$, $\sigma = 0.05$, $\tau = 0.3$ (Panel 1); $\mu = 1$, $\sigma = 0.05$, $\tau = 0.3$ (Panel 2); $\mu = 0.5$, $\sigma = 0.2$, $\tau = 0.3$ (Panel 3); and $\mu = 0.5$, $\sigma = 0.05$, $\tau = 0.8$ (Panel 4).

Bayesian Parameter Estimation and Priors

As explained in Matzke et al. (2013), the BPA simultaneously estimates the μ_{go}, σ_{go}, and τ_{go} parameters of the go RT distribution and the μ_{stop}, σ_{stop}, and τ_{stop} parameters of the SSRT distribution. The BPA relies on Bayesian parameter estimation and therefore involves specifying the prior distribution of the model parameters. BEESTS uses slightly different priors than the WinBUGS implementation of the BPA. Note however that Bayesian parameter estimation is insensitive to the choice of the prior as long as sufficiently diagnostic data are available (e.g., Edwards et al., 1963; Gill, 2002; M. D. Lee & Wagenmakers, 2013). The prior distributions of the model parameters for the BEESTS implementation are listed in the Appendix C.1. The ability of BEESTS to recover underlying true parameter values with the present prior setting has been validated in a series of simulation studies. See the supplemental materials at `http://dora.erbe-matzke.com/publications.html` for a summary of the results of the parameter recoveries.

The BPA relies on Markov chain Monte Carlo sampling (MCMC; Gamerman & Lopes, 2006; Gilks et al., 1996) to obtain posterior distributions for the go and stop parameters. Figure 4.4 illustrates the basic concepts of Bayesian parameter estimation using MCMC sampling. The bottom panel of Figure 4.4 shows sequences of values (i.e., MCMC chains) sampled from the posterior distribution of the τ_{stop} parameter. The accuracy of the sampling process can be increased by running multiple chains, discarding the beginning of each chain as burn-in, and by thinning the chains to decrease autocorrelation. In the present illustration, we ran three chains, each with different starting values and retained $2,000$ iterations per chain, resulting in a total of $6,000$ samples from the posterior distribution (see also Matzke et al., 2013).

The top panel of Figure 4.4 shows the prior and posterior distribution of the τ_{stop} parameter. The horizontal gray line at the bottom of the figure shows the prior distribution of τ_{stop}. The prior is updated by the incoming data to yield the posterior distribution. The histogram and the gray density plot show the distribution of the samples drawn from the posterior distribution of τ_{stop} collapsed over the three MCMC chains. The posterior distribution quantifies the uncertainty about the estimate of τ_{stop}. The central tendency of the posterior, such as the median, is often used as a point estimate of the parameter. The dispersion of the posterior, such as the standard deviation or the percentiles, quantifies the precision of the parameter estimate; the larger the dispersion, the greater the uncertainty in the estimated parameter. For example, the horizontal line at the top of Figure 4.4 ranges from the 2.5^{th} to the 97.5^{th} percentile of the posterior (i.e., 95% Bayesian credible interval), indicating that we can be 95% confident that the true value of τ_{stop} lies within this range (see also Matzke et al., 2013).

Before interpreting the parameter estimates, it is crucial to ensure that the chains have converged from their starting values to their stationary distribution. First, we verify that the posterior distributions of the model parameters are unimodal. Second, we run multiple MCMC chains and ascertain that the chains have mixed well. At convergence, the individual MCMC chains should look like "hairy caterpillars" and should be indistinguishable from one another. Lastly, we compute the $\hat{R}$ (Gelman & Rubin, 1992) convergence diagnostic measure for each model parameter. $\hat{R}$ compares the between-chain variability to the within-chain variability. As a rule of thumb, $\hat{R}$ should be lower than 1.1 if the chains have properly converged. In case of convergence problems, we recommend that users increase the number of samples, the length of the burn-in period, and the degree of thinning.

The BPA can be applied to individual as well as hierarchical stop-signal data. See Matzke et al. (2013) for the graphical representation of the individual and hierarchical BPA models. For the individual analysis, the goal is to estimate the ex-Gaussian go and stop parameters for each participant separately. In contrast, for the hierarchical analysis (e.g., Farrell & Ludwig, 2008; Gelman & Hill, 2007; M. D. Lee, 2011; Matzke & Wagenmakers, 2009; Rouder et al., 2005), the BPA assumes that the participant-level go and stop parameters are drawn from group-level distributions. The group-level distributions specify the between-subject variability of the participant-level parameters. The group-level distributions are themselves characterized by a set of group-level parameters. The goal is to simultaneously estimate the group-level parameters as well as the participant-level go and stop parameters. As explained in Matzke et al. (2013), hierarchical modeling is particularly valuable in situations with only a small number of observations per participant and moderate between-subject variability in parameter values (Gelman & Hill, 2007). In such situations, Bayesian hierarchical modeling typically yields less variable and more accurate estimates than single-level parameter estimation (Farrell & Ludwig, 2008; Rouder et al., 2005).The advantages of the hierarchical approach are less pronounced in situations with a large number of observations per participant. Similarly, in settings with only a few participants —a typical scenario in psychophysical experiments— the group-level parameters cannot be estimated precisely, a problem that diminishes the benefits of hierarchical modeling. In these cases, the individual approach may perform similarly well as the hierarchical approach.

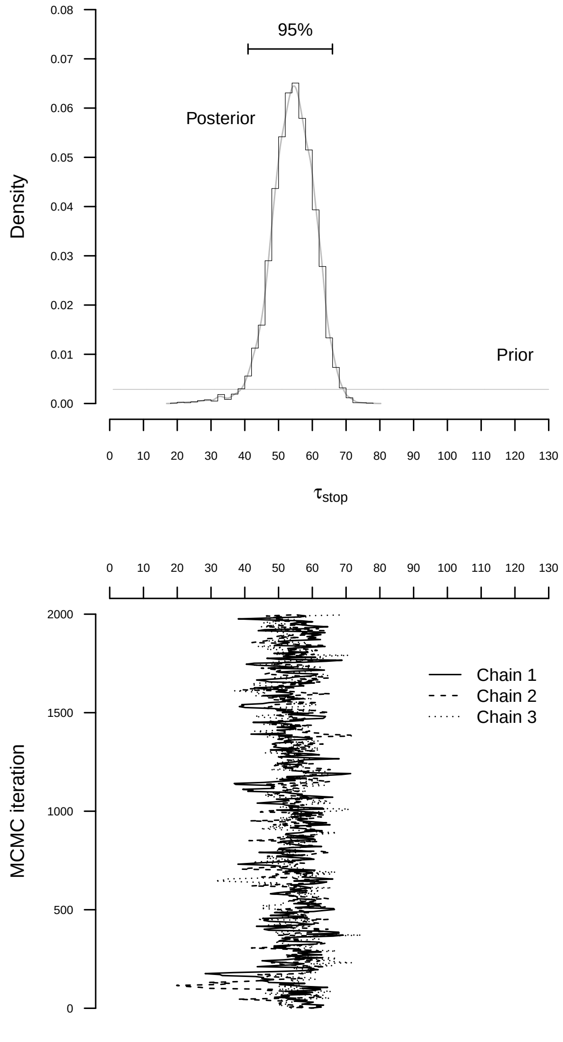


Figure 4.4 *Illustration of MCMC-based Bayesian estimation for the τ_{stop} parameter with the individual BPA.* The histogram in the top panel figure shows the posterior distribution of τ_{stop}. The corresponding gray line indicates the fit of a nonparametric density estimator. The horizontal black line at the top of the top panel shows the 95% Bayesian credible interval. The horizontal gray line at the bottom of the top panel shows the prior distribution of τ_{stop}. The solid, dashed and dotted lines in the bottom panel figure represent the different sequences of values (i.e., MCMC chains) sampled from the posterior distribution of τ_{stop}. To create the histogram in the top panel, the sampled values were first collected across the three chains and then projected onto the x–axis of the top panel figure (see also Matzke et al., 2013).

4.2 Releasing the BEESTS

BEESTS is a cross-platform open-source software for the estimation of SSRT distributions with the Bayesian parametric approach (Matzke et al., 2013). BEESTS relies on Python for parameter estimation and on R (R Core Team, 2012) for the post-processing of the posterior distribution of the model parameters. Specifically, BEESTS uses the Python-based toolboxes kabuki (Wiecki, Sofer, & Frank, 2013) and PyMC (Patil, Huard, & Fonnesbeck, 2010) to construct the model and to generate samples from the posterior distribution of the model parameters using Metropolis-within-Gibbs sampling (Tierney, 1994), respectively. For computational efficiency, the likelihood functions are coded in Cython (Behnel et al., 2011). Once the model parameters are estimated, BEESTS relies on R to compute summary statistics for the posterior distribution of the model parameters and to assess the quality of the parameter estimates. As shown in Figure 4.5, BEESTS is equipped with an easy-to-use graphical user interface (GUI).

4.3 Installation

BEESTS is a stand-alone and open source software released under the Affero General Public License. BEESTS runs on Windows (Windows XP and Windows 7) and OS X (Mountain Lion) operating systems. The software is freely available at `http://dora.erbe-matzke.com/software.html`. To install BEESTS on Windows, download `BEESTS-1.2.zip` and unpack the zip file at any desired location on your computer. Start the GUI by clicking on `BEESTS.exe`. To install BEESTS on OS X, download `BEESTS-1.2.dmg`, double-click the file, and install it on your computer.

4.4 Loading Data

The top panels of Figure 4.6 show the required data format for the analysis. Data files should be saved as csv (i.e., comma-separated values) files. For the individual analysis, the first row of the data file must contain the column names `"ss_presented","inhibited","ssd",and "rt".` The remaining rows contain the data for each go and stop-signal trial. For the hierarchical analysis, the first row of the data file must additionally contain the column name `"subj_idx".` See Table 4.1 for instructions on response coding and the `examples` folder in BEESTS for examples of the data format.

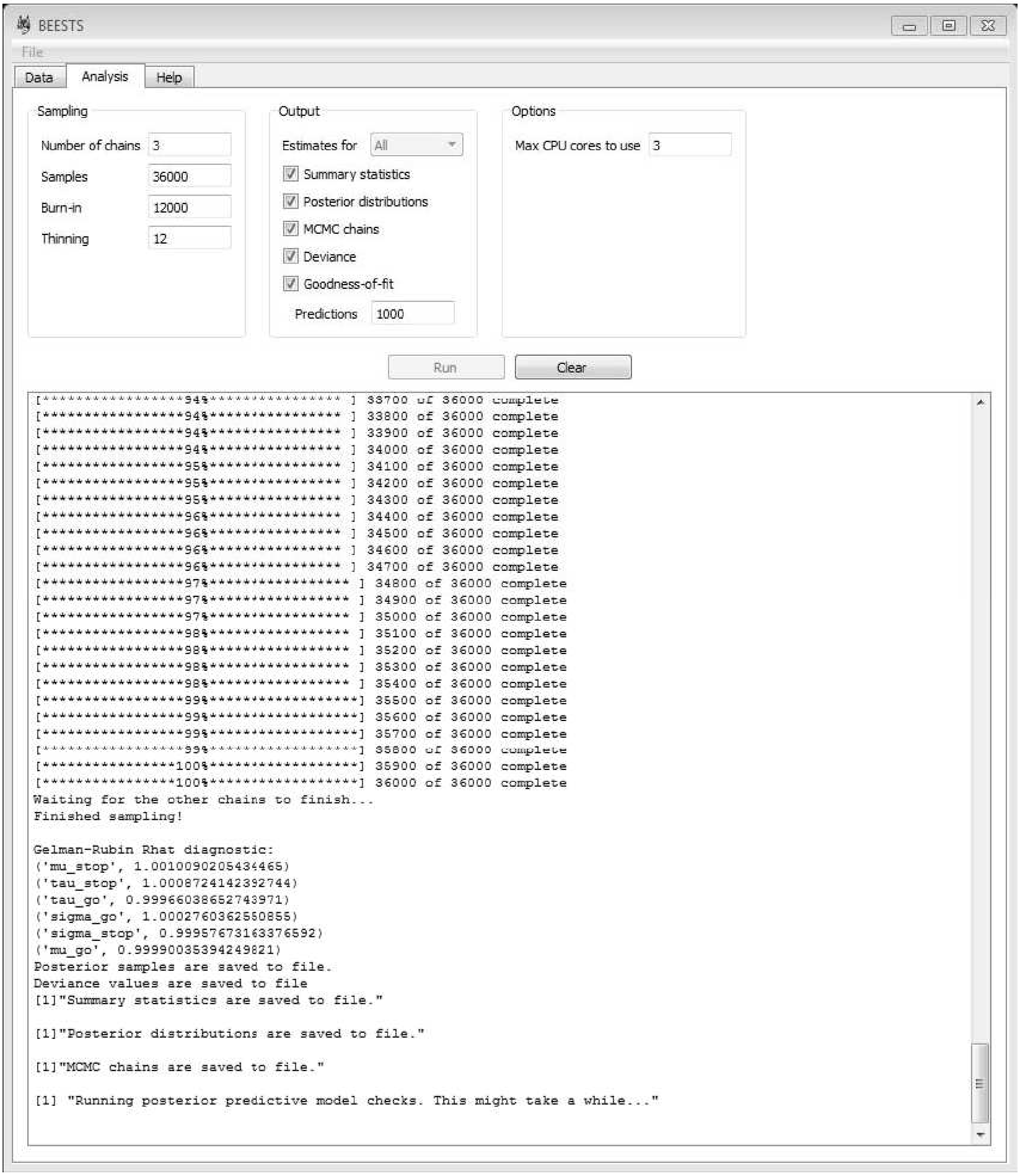
To load the data file, click on `Open` in the `File` menu and follow the instructions. Based on the data format, BEESTS automatically infers whether an individual or hierarchical analysis is appropriate: data files without the `"subj_idx"` column are analyzed with the individual BPA, whereas data files with the `"subj_idx"` column are analyzed with the hierarchical BPA.

4.5 Analysis

Once the data are loaded, users can specify the details of the MCMC sampling, the required output, and the preferred number of CPU cores used by BEESTS.

Sampling

BEESTS allows users to specify the following aspects of the sampling run. Typical values of the input arguments are shown in Figure 4.5.



Number of chains 3

Samples 36000

Burn-in 12000

Thinning 12

Estimates for All

☑ Summary statistics

☑ Posterior distributions

☑ MCMC chains

☑ Deviance

☑ Goodness-of-fit

Predictions 1000

Max CPU cores to use 3

```
[*****************94%******************  ] 33700 of 36000 complete
[*****************94%******************  ] 33800 of 36000 complete
[*****************94%******************  ] 33900 of 36000 complete
[*****************94%******************  ] 34000 of 36000 complete
[*****************95%******************  ] 34100 of 36000 complete
[*****************95%******************  ] 34200 of 36000 complete
[*****************95%******************  ] 34300 of 36000 complete
[*****************96%******************  ] 34400 of 36000 complete
[*****************96%******************  ] 34500 of 36000 complete
[*****************96%******************  ] 34600 of 36000 complete
[*****************96%******************  ] 34700 of 36000 complete
[*****************97%******************* ] 34800 of 36000 complete
[*****************97%******************* ] 34900 of 36000 complete
[*****************97%******************* ] 35000 of 36000 complete
[*****************98%******************* ] 35100 of 36000 complete
[*****************98%******************* ] 35200 of 36000 complete
[*****************98%******************* ] 35300 of 36000 complete
[*****************98%******************* ] 35400 of 36000 complete
[*****************99%********************] 35500 of 36000 complete
[*****************99%********************] 35600 of 36000 complete
[*****************99%********************] 35700 of 36000 complete
[*****************99%********************] 35800 of 36000 complete
[****************100%********************] 35900 of 36000 complete
[****************100%********************] 36000 of 36000 complete
Waiting for the other chains to finish...
Finished sampling!

Gelman-Rubin Rhat diagnostic:
('mu_stop', 1.0010090205434465)
('tau_stop', 1.0008724142392744)
('tau_go', 0.99966038652743971)
('sigma_go', 1.0002760362550855)
('sigma_stop', 0.99957673163376592)
('mu_go', 0.99990035394249821)
Posterior samples are saved to file.
Deviance values are saved to file
[1]"Summary statistics are saved to file."

[1]"Posterior distributions are saved to file."

[1]"MCMC chains are saved to file."

[1] "Running posterior predictive model checks. This might take a while..."
```

Figure 4.5 *Graphical user interface for BEESTS.* See text for details.

Table 4.1 Response Coding for the Hierarchical BEESTS Analysis

"subj_idx"	"ss_presented"	"inhibited"	"ssd"	"rt"
1	0	-999	-999	656
1	1	0	300	469
1	1	1	300	-999

Note. The `"subj_idx"` column contains the participant number. The `"ss_presented"` column contains the trial type, where go trials are coded with 0 and stop-signal trials are coded with 1. The `"inhibited"` column contains the inhibition data, where signal-respond trials are coded with 0 (i.e., unsuccessful inhibition), signal-inhibit trials are coded with 1 (i.e., successful inhibition), and go trials are coded with -999. The `"ssd"` column contains the stop-signal delay in ms., where go trials are coded with -999. The `"rt"` column contains the go RT for go trials and the signal-respond RT for signal-respond trials in ms., where signal-inhibit trials are coded with -999.

Number of Chains

Use the `Number of chains` option to specify the number of MCMC chains, i.e., sequences of values sampled from the posterior distribution of the parameters. The start values are automatically set to the maximum a posteriori probability (MAP) estimates of the parameters.

Samples

Use the `Samples` option to specify the total number of MCMC samples per chain.

Burn-In

Use the `Burn-in` option to specify the number of burn-in samples to discard at the beginning of each chain.

Thinning

Use the `Thinning` option to specify the degree of thinning within each chain. For instance, a thinning factor of 12 means that only every 12^{th} MCMC sample will be retained.

Output

All output will be saved in the directory where the data file is located. BEESTS automatically saves the posterior samples from each chain to a separate csv file (e.g., `name.datafile_parameters1.csv`, `name.datafile_parameters2.csv`,etc.). If multiple chains are run, BEESTS automatically displays the $\hat{R}$ statistic for each model parameter (see Figure 4.5).

As shown in Figure 4.5, BEESTS allows users to request the following additional output. If `Estimates for` is set to `All` in a hierarchical analysis, BEESTS will provide the selected output options (i.e., summary statistics, density plots of the posterior distributions, and MCMC trace plots) for the group-level parameters *and* for each participant separately. If `Estimates for` is set to `Only-group`, BEESTS will provide the selected output options only for the group-level parameters.

Summary Statistics

Use the `Summary statistics` option to obtain a csv file with the summary statistics (i.e., mean, standard deviation, and quantiles) of the posterior distribution of the model parameters and of the corresponding mean and standard deviation of the go and SSRT distribution (see Equation 4.3 and Equation 4.4).

Posterior Distributions

Use the `Posterior distributions` option to obtain a pdf file with the density plots of the posterior and the prior distribution of the model parameters.

MCMC Chains

Use the `MCMC chains` option to obtain a pdf file with trace plots for the MCMC chains of the model parameters.

Deviance

Use the `Deviance` option to obtain the deviance values from each chain in a separate csv file (e.g., `name.datafile_deviance1.csv,name.datafile_deviance2.csv,`etc.). The deviance values may be used to compute the Deviance Information Criterion (DIC, e.g., Spiegelhalter, Best, Carlin, & van der Linde, 2002) measure of model selection.

Goodness-of-Fit

Use the `Goodness-of-fit` option to assess the absolute goodness-of-fit of the model using posterior predictive model checks. As explained in Matzke et al. (2013), the adequacy of the model can be assessed by generating predicted data using the posterior distributions of the parameters. If the model adequately describes the data, the predictions based on the model parameters should closely approximate the observed data. The model checks can be formalized by computing posterior predictive p values (e.g., Gelman & Hill, 2007; Gelman et al., 1996, but see Bayarri & Berger, 1998). Extreme p values close to 0 or 1 indicate that the BPA does not describe the observed data adequately.

For each individual participant, BEESTS uses the median of the observed and predicted signal-respond RTs as test statistics. The `Predictions` option can be used to specify the number of predicted data sets. BEESTS then randomly samples the specified number of parameter vectors from the joint posterior of the individual go and stop parameters. Next, BEESTS generates the specified number of predicted stop-signal data sets for each SSD using the corresponding number of stop-signal trials and the chosen parameter vectors. For each predicted data set, BEESTS then computes the median signal-respond RT. Lastly, for each SSD, BEESTS computes the one-sided posterior predictive p value given by the fraction of times that the predicted median signal-respond RT is greater than the observed median signal-respond RT. Corresponding two-sided p values can be computed as $2 \times \min(p, 1 - p)$. Note however that two-sided p values are well defined only when the test statistic has a symmetric distribution. Note also that BEESTS assesses model fit on all SSDs that contain at least one observed signal–respond RT. In order to obtain stable median signal-respond RTs, however, we advise users to interpret the results only on SSDs with a reasonable number of observed signal-respond RTs.

The output of the posterior predictive model checks consists of (1) a csv file listing for each SSD the number of observed signal-respond RTs, the observed median signal-respond RT, the average of the predicted median signal-respond RTs, and the one-sided and two-sided posterior predictive p value and (2) a pdf file with a graphical summary of the model checks using violin plots. Violin plots (e.g., Hintze & Nelson, 1998) combine information available from density plots with information about summary statistics in the form of box plots. Note that irrespective of the type of analysis (individual or hierarchical), the goodness-of-fit of the model is assessed on a participant level using the parameter values of the individual participants (see Figure 4.8).

Options: Max CPU Cores to Use

Use the `Max CPU cores to use` option to specify the number of CPU cores to use during the sampling process. If multiple MCMC chains are requested, BEESTS can run the chains in parallel by allocating each chain to a different CPU core in order to increase speed. The default number of CPU cores used by BEESTS is the number of cores available on the computer minus one.

Running the Analysis

Once the details of the sampling process and the required output are specified, start the analysis by clicking on `Run`. As shown in Figure 4.5, BEESTS automatically displays the progress of the sampling. If multiple MCMC chains are run in parallel, BEESTS displays the progress of only one of the MCMC chains (i.e., the main process). The analysis can be stopped by "killing" the (parallel) processes in the Task Manager. Use the `Clear` command to clear the working space.

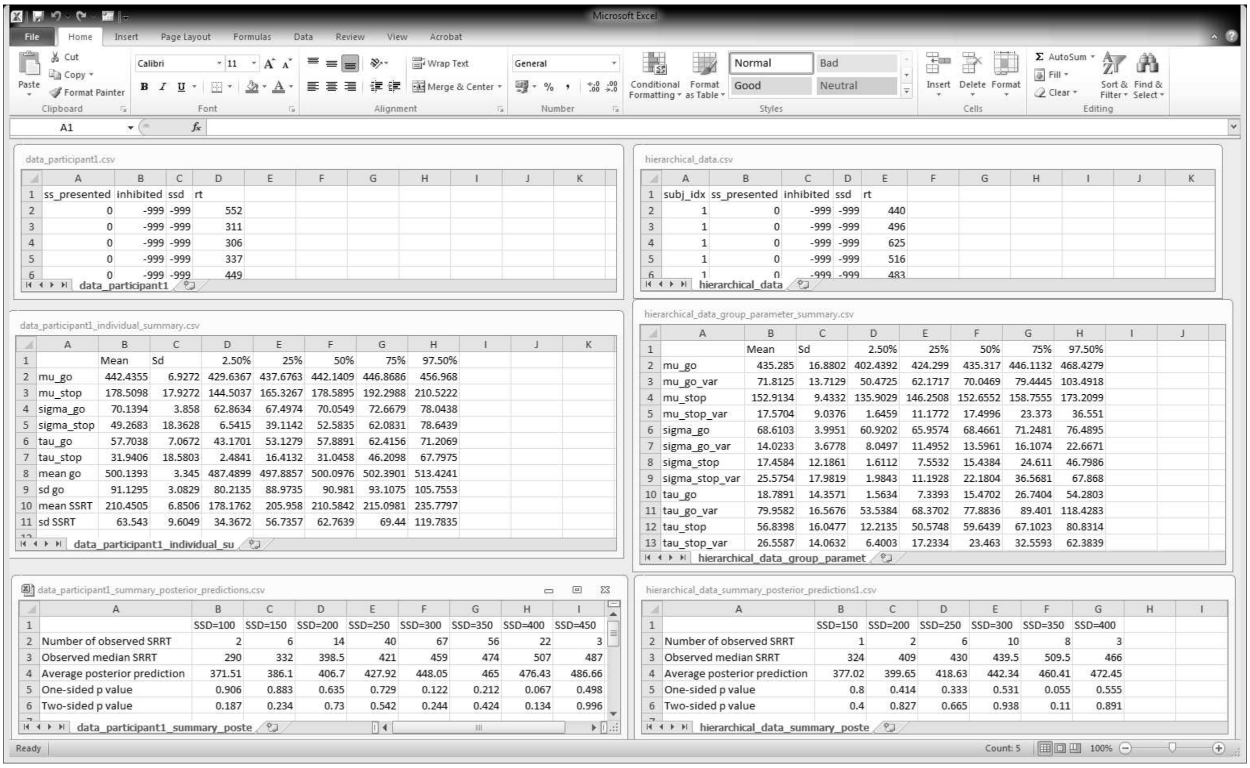
4.6 Empirical Data Examples: Individual and Hierarchical Analysis

In this section, we illustrate the use of BEESTS with the stop-signal data of 20 participants from the 40% stop-signal condition of the first experiment reported in Bissett and Logan (2011). The data set featured a relatively large number of 720 go trials and 480 stop-signal trials per participant. See Matzke et al. (2013) for the details on the data pre-processing and the model fitting. For all of the participants, the BEESTS implementation yielded parameter estimates that are highly similar to the ones obtained from the WinBUGS routine. For a comparison of the parameter estimates from the BEESTS and the WinBUGS implementation, the reader is referred to the supplemental materials and to the empirical data examples in Matzke et al. (2013).

Due to relatively high autocorrelations between the parameters, we ran long chains, discarded the beginning of the chains as burn-in and thinned each chain. The results reported below are based on 6,000 retained samples, using `Number of chains = 3`, `Samples = 36000`, `Burn-in = 12000`, and `Thinning = 12`.

Individual Analysis

In this section, we present the results of fitting the data of Participant 1 with the individual BPA. See the `examples` folder for the data set. Using three CPU cores, the sampling took approximately 23 minutes with BEESTS. The same analysis took about 15 hours with WinBUGS. The top left panel of Figure 4.6 shows the required data format for the individual analysis. Figure 4.7 shows the posterior and prior distributions (left panel; option `Posterior distributions`) and the MCMC chains (right panel; option `MCMC chains`) for the six model parameters. The prior distributions are adequately updated; the posteriors are substantially narrower than the priors. The posterior distributions and the three MCMC chains do not show signs of convergence problems. All $\hat{R}$ values were lower than 1.05. The middle left panel of Figure 4.6 shows the summary statistics of the posterior distribution of the model parameters (option `Summary statistic`). The posterior distributions are estimated well as evidenced by the relatively small posterior standard deviations.



data_participant1.csv

	A	B	C	D	E	F	G	H	I	J	K
1	ss_presented	inhibited	ssd	rt							
2	0	-999	-999	552							
3	0	-999	-999	311							
4	0	-999	-999	306							
5	0	-999	-999	337							
6	0	-999	-999	449							

data_participant1

hierarchical_data.csv

	A	B	C	D	E	F	G	H	I	J	K
1	subj_idx	ss_presented	inhibited	ssd	rt						
2	1	0	-999	-999	440						
3	1	0	-999	-999	496						
4	1	0	-999	-999	625						
5	1	0	-999	-999	516						
6	1	0	-999	-999	483						

hierarchical_data

data_participant1_individual_summary.csv

	A	B	C	D	E	F	G	H	I	J	K
1		Mean	Sd	2.50%	25%	50%	75%	97.50%			
2	mu_go	442.4355	6.9272	429.6367	437.6763	442.1409	446.8686	456.968			
3	mu_stop	178.5098	17.9272	144.5037	165.3267	178.5895	192.2988	210.5222			
4	sigma_go	70.1394	3.858	62.8634	67.4974	70.0549	72.6679	78.0438			
5	sigma_stop	49.2683	18.3628	6.5415	39.1142	52.5835	62.0831	78.6439			
6	tau_go	57.7038	7.0672	43.1701	53.1279	57.8891	62.4156	71.2069			
7	tau_stop	31.9406	18.5803	2.4841	16.4132	31.0458	46.2098	67.7975			
8	mean go	500.1393	3.345	487.4899	497.8857	500.0976	502.3901	513.4241			
9	sd go	91.1295	3.0829	80.2135	88.9735	90.981	93.1075	105.7553			
10	mean SSRT	210.4505	6.8506	178.1762	205.958	210.5842	215.0981	235.7797			
11	sd SSRT	63.543	9.6049	34.3672	56.7357	62.7639	69.44	119.7835			

data_participant1_individual_su

hierarchical_data_group_parameter_summary.csv

	A	B	C	D	E	F	G	H	I	J
1		Mean	Sd	2.50%	25%	50%	75%	97.50%		
2	mu_go	435.285	16.8802	402.4392	424.299	435.317	446.1132	468.4279		
3	mu_go_var	71.8125	13.7129	50.4725	62.1717	70.0469	79.4445	103.4918		
4	mu_stop	152.9134	9.4332	135.9029	146.2508	152.6552	158.7555	173.2099		
5	mu_stop_var	17.5704	9.0376	1.6459	11.1772	17.4996	23.373	36.551		
6	sigma_go	68.6103	3.9951	60.9202	65.9574	68.4661	71.2481	76.4895		
7	sigma_go_var	14.0233	3.6778	8.0497	11.4952	13.5961	16.1074	22.6671		
8	sigma_stop	17.4584	12.1861	1.6112	7.5532	15.4384	24.611	46.7986		
9	sigma_stop_var	25.5754	17.9819	1.9843	11.1928	22.1804	36.5681	67.868		
10	tau_go	18.7891	14.3571	1.5634	7.3393	15.4702	26.7404	54.2803		
11	tau_go_var	79.9582	16.5676	53.5384	68.3702	77.8836	89.401	118.4283		
12	tau_stop	56.8398	16.0477	12.2135	50.5748	59.6439	67.1023	80.8314		
13	tau_stop_var	26.5587	14.0632	6.4003	17.2334	23.463	32.5593	62.3839		

hierarchical_data_group_paramet

data_participant1_summary_posterior_predictions.csv

	A	B	C	D	E	F	G	H	I
1		SSD=100	SSD=150	SSD=200	SSD=250	SSD=300	SSD=350	SSD=400	SSD=450
2	Number of observed SRRT	2	6	14	40	67	56	22	3
3	Observed median SRRT	290	332	398.5	421	459	474	507	487
4	Average posterior prediction	371.51	386.1	406.7	427.92	448.05	465	476.43	486.66
5	One-sided p value	0.906	0.883	0.635	0.729	0.122	0.212	0.067	0.498
6	Two-sided p value	0.187	0.234	0.73	0.542	0.244	0.424	0.134	0.996

data_participant1_summary_poste

hierarchical_data_summary_posterior_predictions1.csv

	A	B	C	D	E	F	G	H	I
1		SSD=150	SSD=200	SSD=250	SSD=300	SSD=350	SSD=400		
2	Number of observed SRRT	1	2	6	10	8	3		
3	Observed median SRRT	324	409	430	439.5	509.5	466		
4	Average posterior prediction	377.02	399.65	418.63	442.34	460.41	472.45		
5	One-sided p value	0.8	0.414	0.333	0.531	0.055	0.555		
6	Two-sided p value	0.4	0.827	0.665	0.938	0.11	0.891		

hierarchical_data_summary_poste

Figure 4.6 *BEESTS input and output.* The left panels show input and output for the individual analysis. The right panels show input and output for the hierarchical analysis. The top panels show the required data format. The middle panels show the output of the `Summary statistic` option. For the hierarchical analysis, only the group-level mean and group-level variability (i.e., standard deviation) parameters are shown. The bottom panels show partial output for the `Goodness-of-fit` option for Participant 1 in the Bissett and Logan (2011) experiment. SRRT = signal-respond RT.

The go parameters are generally estimated more precisely than the stop parameters because the go parameters are estimated based on the go RTs as well as the signal-respond RTs and are therefore better constrained by the data.

The bottom left panel of Figure 4.6 shows the summary of the posterior predictive model checks (option `Goodness-of-fit`) using 1,000 samples from the joint posterior of the model parameters (`Samples = 1000`). As mentioned above, we advise users to assess model fit only on SSDs with a reasonable number of observed signal-respond RTs. For instance, we assessed goodness-of-fit only on SSDs with at least 10 observed signal-respond RTs. The one-sided p values on these five SSDs (i.e., 200, 250, 300, 350, and 400 ms) are far from 0 or 1 and the two-sided p values are all above 0.05. The left panel of Figure 4.8 shows the corresponding graphical summary for the model checks. For the selected SSDs, the observed median signal-respond RTs (i.e., black triangles) are well within the 2.5th and 97.5th percentile of the predicted median signal-respond RTs (see gray violin plots), and are adequately approximated by the median of the predicted median signal-respond RTs (i.e., white circles). The results of the posterior predictive model checks indicated thus that the BEESTS analysis appropriately accounted for the observed data.

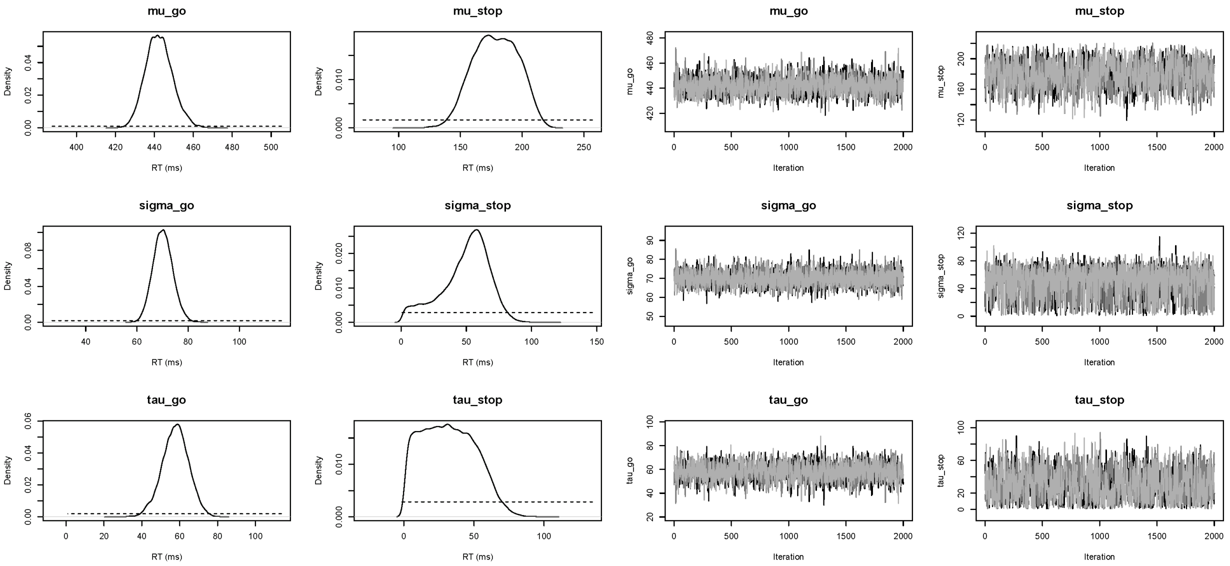


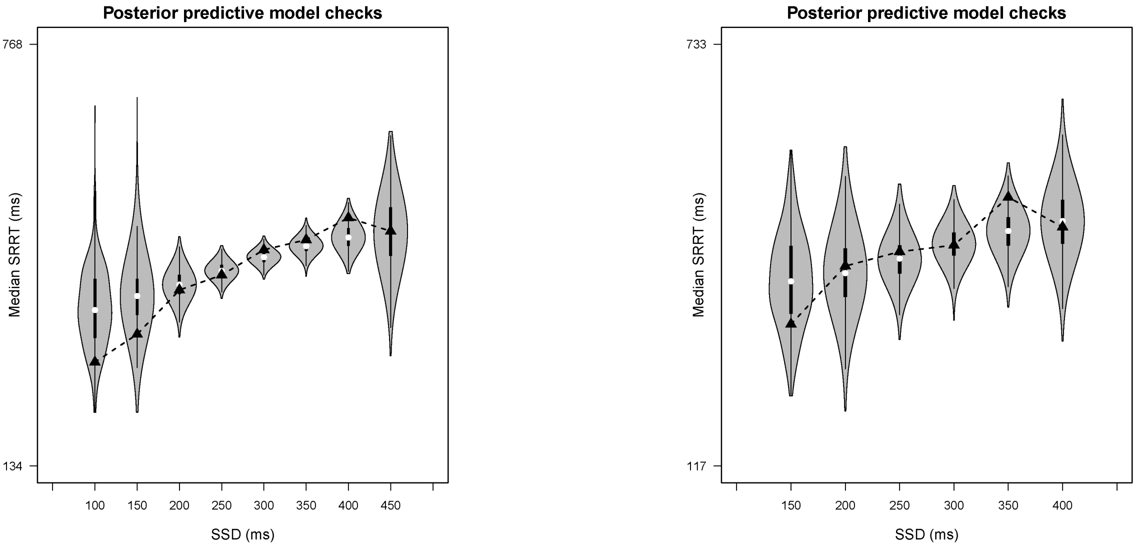
Figure 4.7 *Posterior (black solid lines) and prior distributions (black dotted lines; left panel) and MCMC chains (right panel) of the model parameters for Participant 1 in the Bissett and Logan (2011) data set obtained with the individual BPA.*

Hierarchical Analysis

As explained above, the hierarchical approach has the potential to provide accurate parameter estimates with relatively few observations per participant. To illustrate the benefits of the hierarchical approach over the individual BPA with scarce data, this section presents the results of fitting a subsample of the observations from the Bissett and Logan (2011) data set with the hierarchical as well as the individual BPA. For each of the 20 participants, we fit a randomly selected 90 go RTs, 30 signal-respond RTs, and 30 successful inhibitions with the hierarchical BPA. We then compared the results from the hierarchical analysis to the results from fitting the same subsample of data with the individual BPA. Using three CPU cores, the hierarchical analysis took approximately 3.5 hours with BEESTS. The same analysis took about 100 hours with WinBUGS.

The top right panel of Figure 4.6 shows the required data format for the hierarchical analysis. Figure 4.9 shows the posterior and prior distributions (top panel) and the MCMC chains (bottom panel) for the group-level mean and standard deviation parameters. The prior distribution of the group-level parameters are adequately updated; the posteriors are substantially narrower than the priors and the chains have mixed well. The $\hat{R}$ values for all group-level and individual parameters were lower than 1.05. The middle right panel of Figure 4.6 shows the summary statistics of the posterior distribution of the group-level mean and standard deviation parameters. The posterior distributions are estimated relatively precisely. Note that if the `Estimates for All` option is selected, BEESTS also produces output (i.e., density plots of the posteriors, MCMC trace plots, and summary statistics) for the individual go and stop parameters for each participant separately.

The bottom right panel of Figure 4.6 shows the summary of the posterior predictive model checks for Participant 1 using 1,000 samples from the joint posterior of the participant-level model parameters obtained with the hierarchical BPA. All posterior predictive p values are well within an acceptable range. Note, however, that the median signal-respond RTs —observed and predicted—



(a) Individual BPA

(b) Hierarchical BPA

Figure 4.8 *Results of the posterior predictive model checks for Participant 1 in the Bissett and Logan (2011) data set with the individual (panel a) and the hierarchical (panel b) BPA.* See text for a detailed description of the posterior predictive analyses. For each SSD, the figures show the observed median signal-respond RT (black triangle), a density plot of the predicted median signal-respond RTs (gray violin plot), a boxplot ranging from the 25^{th} to the 75^{th} percentile of the predicted median signal-respond RTs, and the median of the predicted median signal-respond RTs (white circle). SRRT = signal-respond RT.

are based on only a few observations. The right panel of Figure 4.8 shows the corresponding graphical summary of the posterior predictive model checks. All observed median signal-respond RTs are well within the range of the median signal-respond RTs predicted by the joint posterior of the model parameters. Due to the scarcity of the data, however, there is large uncertainty in the predicted median signal-respond RTs. Compare the results of the posterior predictive model checks in the bottom two panels of Figure 4.8. The violin plots in the left panel show the predicted median signal-respond RTs from the individual analysis of the data of Participant 1 based on the full 1,200 trials. The violin plots in the right panel show the predicted median signal-respond RTs from the hierarchical analysis of the data of Participant 1 based on a subsample of only 150 trials. Because the hierarchical analysis is based on substantially fewer observations than the individual analysis of the full data set presented in the previous section, the predicted median signal-respond RTs in the right panel are more spread out than the predicted median signal-respond RTs in the left panel. Posterior predictive p values resulting from such unstable observed and predicted median signal-respond RTs should be interpreted with caution.

To illustrate the benefits of the hierarchical approach over the individual BPA with scarce data, we compared the parameter estimates from the hierarchical analysis with estimates obtained from the individual analysis of the same subsample of 150 trials. As mentioned above, hierarchical modeling generally results in more accurate and less variable estimates than single-level estimation. Figure 4.10 shows the posterior distribution of the stop parameters of Participant 1 obtained with the hierarchical and the individual BPA using the same subsample of 150 observations. The gray density plots show the posterior distribution of the stop parameters from the hierarchical BPA. The

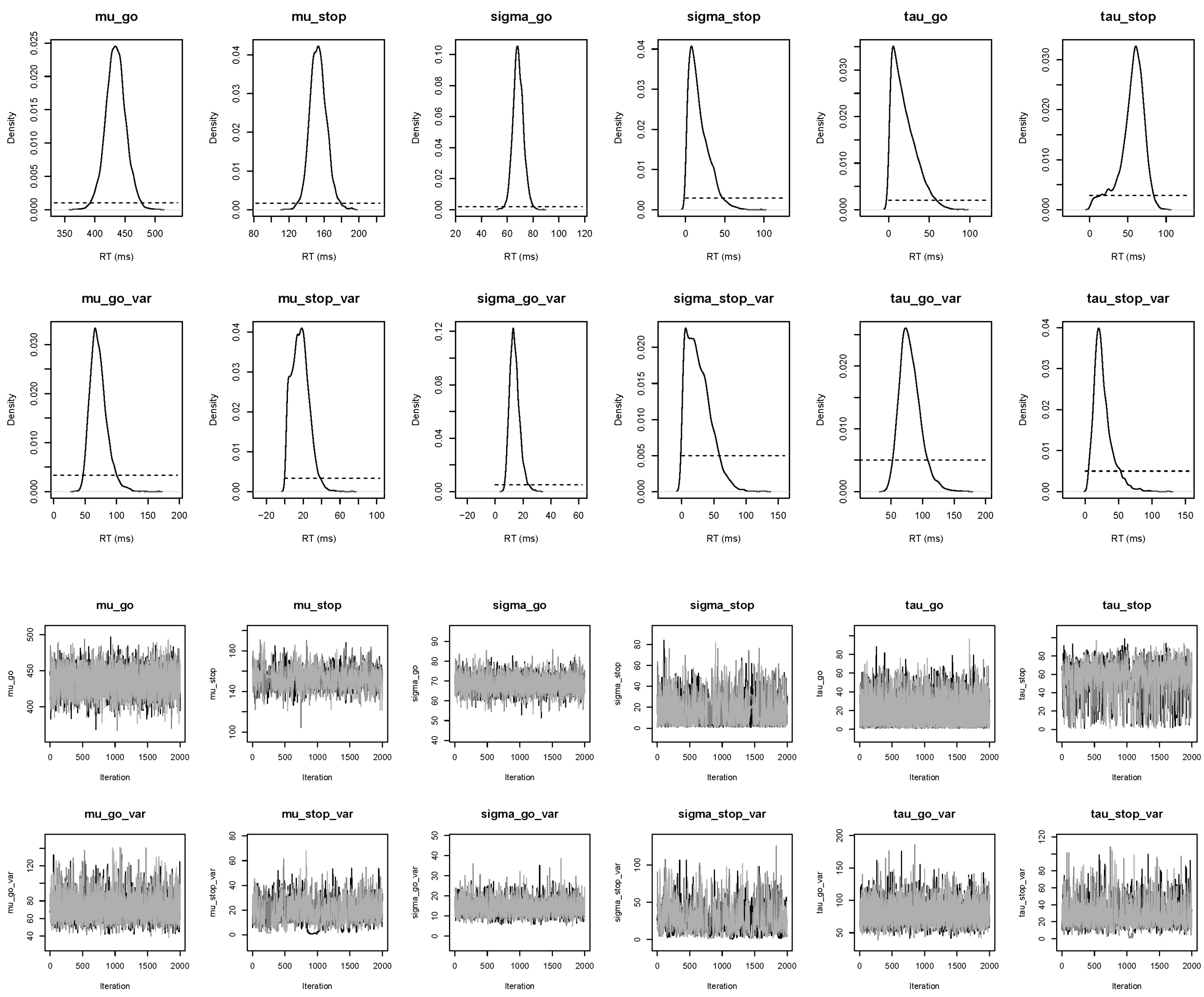


Figure 4.9 *Posterior distributions and MCMC chains of the group-level model parameters in the Bissett and Logan (2011) data set obtained with the hierarchical BPA.* The first and third rows show posterior (black solid line) and prior distributions (black dotted line) and MCMC trace plots for the group-level mean parameters, respectively. The second and fourth rows show posterior and prior distributions and trace plots for the group-level variability (i.e., group-level standard deviation) parameters, respectively.

black density plots show the posterior distribution of the stop parameters from the individual analysis. The posterior distributions of the stop parameters estimated with the hierarchical approach are less variable (i.e., smaller 95% Bayesian credible interval) than the posteriors estimated with the individual BPA. Also, the posterior medians from the hierarchical analysis are —as expected— shrunk towards their corresponding group mean (see also Matzke et al., 2013).

4.7 Discussion

The horse-race model presents various opportunities to estimate the latency of response inhibition in the stop-signal paradigm. Most methods, however, only focus on deriving a summary measure

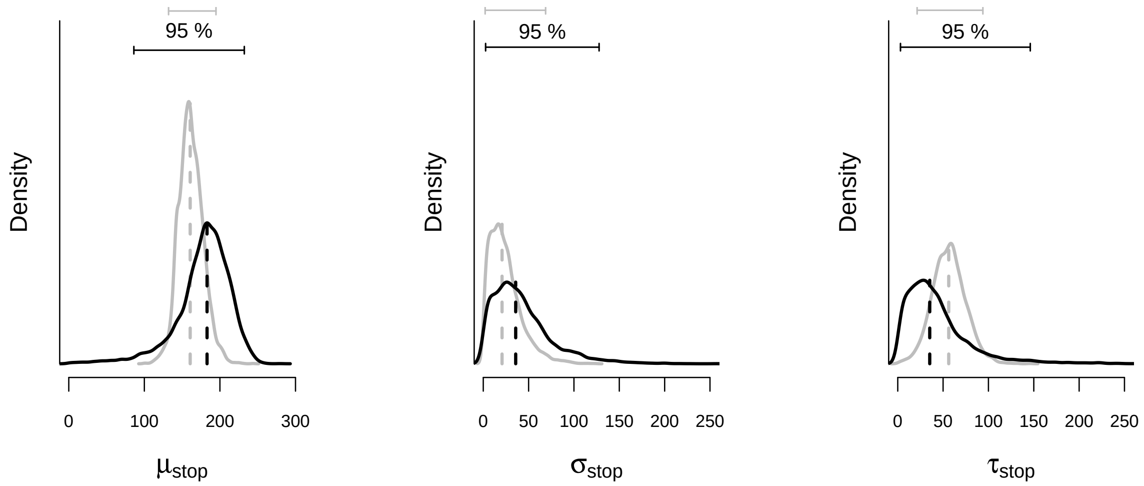


Figure 4.10 *Posterior distribution of the stop parameters estimated from a subsample of the data of Participant 1 with the individual and the hierarchical BPA.* The solid black and gray lines show the posterior distribution of the stop parameters and the corresponding 95% Bayesian credible intervals obtained with the individual and the hierarchical BPA, respectively. The dashed black and gray lines show the median of the posterior distributions obtained with the individual and the hierarchical BPA, respectively (see also Matzke et al., 2013).

of SSRT. Recently, Matzke et al. (2013) have developed a Bayesian parametric approach (BPA) that allows for the estimation of the entire distribution of stopping latencies. The goal of the present paper was to promote the widespread application of the Bayesian analysis of stop-signal data by introducing BEESTS, a relatively fast and user-friendly software implementation of the BPA. BEESTS provides users with a range of output options, such as summary statistics of the posterior distribution of the parameters and various diagnostic tools to assess the quality of the estimates. Importantly, BEESTS is equipped with an easy-to-use graphical user interface.

BEESTS can be applied to individual as well as hierarchical stop-signal data. The advantage of the individual approach lies in its simplicity. The advantage of the hierarchical approach lies in its potential to provide accurate parameter estimates with relatively few observations per participant. The choice between the individual and the hierarchical approach in practical applications depends on a delicate balance between the quality of the data, the number of participants, the number of trials per participant, and whether users are interested in obtaining accurate parameter estimates on the participant level in order to examine individual differences or focus on group comparisons and are satisfied with interpreting only the group-level parameters. Prior to data collection, users are encouraged to generate synthetic data with varying number of trials and participants, fit the data in BEESTS, and inspect the parameter estimates in order to assess the expected uncertainty of the model parameters under the different scenarios and modeling approaches.

BEESTS assumes that go RTs and SSRTs are ex-Gaussian distributed and relies on Bayesian parameter estimation to obtain estimates for the go and stop parameters. Note, however, that the BPA itself does not hinge on the particular parametric form used to summarize the distributions, nor is it heavily influenced by the exact choice of the prior distributions. In our experience, the ex-Gaussian assumption and the corresponding (group-level and hyper) prior distributions implemented in BEESTS provide a reasonable default setting. Nevertheless, interested users may adapt the source code (`https://github.com/twiecki/stopsignal`) to accommodate alternative parametric assumptions or different prior settings. Also, the posterior predictive model check implemented in BEESTS using the median signal-respond RT is only one of many possible approaches to assess the goodness-of-fit of the model. Users may adapt the source code to implement posterior predictive model checks using alternative test statistics (see Matzke et al., 2013).

Conclusion

Here we introduced a user-friendly software package —BEESTS— that allows for the efficient estimation of entire SSRT distributions using MCMC sampling. BEESTS allows researchers to rigorously address important questions about the variability of stopping latencies, such as the relationship between mean SSRT and SSRT variance. Similarly, BEESTS enables investigators to assess differences in the shape of go RT and SSRT distributions between clinical populations or experimental groups. BEESTS therefore facilitates the interpretation of stop-signal data and may open fruitful new avenues for response inhibition research.

Part II

Multinomial Processing Tree Models

Chapter 5

Bayesian Estimation of Multinomial Processing Tree Models with Heterogeneity in Participants and Items

This chapter is currently in press as:
Dora Matzke, Conor V. Dolan, William H. Batchelder, and Eric-Jan Wagenmakers (in press).
Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items.
Psychometrika.[1]

Abstract

Multinomial processing tree (MPT) models are theoretically motivated stochastic models for the analysis of categorical data. Here we focus on a crossed-random effects extension of the Bayesian latent-trait pair-clustering MPT model. Our approach assumes that participant and item effects combine additively on the probit scale and postulates (multivariate) normal distributions for the random effects. We provide a WinBUGS implementation of the crossed-random effects pair-clustering model and an application to novel experimental data. The present approach may be adapted to handle other MPT models.

5.1 Introduction

Multinomial processing tree (MPT) models are theoretically motivated stochastic models for the analysis of categorical data. MPT models can be used to measure the contribution of the different cognitive processes that determine performance in various experimental paradigms. Due to their simplicity, MPT models have become increasingly popular over the last decades and have been applied to a variety of areas in cognitive psychology (for reviews, see Batchelder & Riefer, 1999; Erdfelder et al., 2009).

[1] The final publication is available at `http://link.springer.com/article/10.1007/s11336-013-9374-9`.

MPT models assume that the observed category responses follow a multinomial distribution. MPT models reparametrize the category probabilities of the multinomial distribution in terms of a number of model parameters that are assumed to represent underlying cognitive processes. The category probabilities are generally expressed as nonlinear functions of the underlying model parameters. Specifically, MPT models assume that the observed response categories result from one or more hypothesized sequences of cognitive events, a structure that can be represented by a rooted tree architecture such as the one depicted in Figure 5.1. The formal properties of MPT models are described by Hu and Batchelder (1994), Purdy and Batchelder (2009), and Riefer and Batchelder (1988). For computer software for fitting and testing MPT models, see for instance Hu and Phillips (1999), Moshagen (2010), and Wickelmaier (2011).

Traditionally, statistical inference for MPT models is carried out on data that are aggregated across participants and items using the classical maximum likelihood approach (e.g., Hu & Batchelder, 1994). This approach relies on the assumption of homogeneity in participants and items, that is, the assumption that participants and items do not differ substantively in terms of the cognitive processes or characteristics represented by the model parameters. However, heterogeneity in participants and items is more likely to be the rule rather than the exception. For example, participant variables such as age and IQ are likely to influence performance on many cognitive tests, and the same holds for item variables such as word frequency and word length. The cognitive processes represented by the model parameters may not only be variable, but may also be highly correlated. For example, two cognitive abilities that both reflect, say, some aspect of memory retrieval are likely to be related, resulting in correlations between the model parameters representing these abilities. Most importantly, in the presence of parameter heterogeneity, the analysis of aggregated data may bias parameter estimation and statistical inference (e.g., Ashby, Maddox, & Lee, 1994; Clark, 1973; Curran & Hintzman, 1995; Estes, 1956; Hintzman, 1980, 1993; Klauer, 2006; Rouder & Lu, 2005; J. B. Smith & Batchelder, 2008).

In recent years, researcher have become increasingly interested in developing approaches to MPT modeling that incorporate parameter heterogeneity (e.g., Klauer, 2006, 2010; Rouder, Lu, Morey, Sun, & Speckman, 2008; J. B. Smith & Batchelder, 2010). These attempts typically involve Bayesian hierarchical or multilevel modeling that allows the model parameters to vary either over participants or over items in a statistically specified way (e.g., Farrell & Ludwig, 2008; Gelman, Carlin, Stern, & Rubin, 2003; Gelman & Hill, 2007; Gill, 2002; M. D. Lee, 2011; M. D. Lee & Newell, 2011; M. D. Lee & Wagenmakers, 2013; Nilsson et al., 2011; Rouder & Lu, 2005; Shiffrin et al., 2008).

A prominent approach to deal with parameter heterogeneity in MPT models is the recently developed latent-trait method (Klauer, 2010). The latent-trait approach relies on Bayesian hierarchical modeling and postulates a multivariate normal distribution for the probit transformed parameters. The latent-trait approach deals with parameter heterogeneity as a result of differences either between participants or between items, but not both. In many situations, however, it is reasonable to assume that the model parameters differ both between the participants and between the particular items used in an experiment. In this case, both sources of variability —participant and item— should be modeled as random effects.

The goal of the present paper is therefore threefold. First, we extend Klauer's (2010) latent-trait approach to accommodate heterogeneity in participants as well as items. Second, we illustrate the use of the resulting crossed-random effects approach with novel experimental data. Lastly, to facilitate the use of Bayesian hierarchical methods in MPT modeling, we provide software implementations of the latent-trait and the crossed-random effects approach using WinBUGS (Bayesian inference Using Gibbs Sampling for Windows; Lunn et al., 2012; Lunn, Spiegelhalter, Thomas, & Best, 2009; Lunn et al., 2000). WinBUGS is a general purpose statistical software for Bayesian analysis that implements the Markov chain Monte Carlo (MCMC; Gamerman & Lopes, 2006; Gilks et al., 1996) algorithm necessary for Bayesian parameter estimation (for an introduction for psychologists, see Kruschke, 2010b; M. D. Lee & Wagenmakers, 2013; Sheu & O'Curry, 1998). We will use the pair-clustering model —one of the most extensively studied MPT models— as an example. However, the crossed-random effects approach presented here may in principle be adapted to handle many other MPT models as well.

The paper is organized as follows. The first section introduces various methods to accommodate parameter heterogeneity in MPT models. The second section introduces the pair-clustering MPT model in more detail. The third section presents the WinBUGS implementation of the latent-trait pair-clustering model. The fourth section presents the crossed-random effects pair-clustering model with the corresponding WinBUGS implementation and describes the results of applying the model to novel experimental data. The fifth section concludes the paper.

5.2 Parameter Heterogeneity in MPT Models

The data for an MPT model consist of category responses from several participants to each of a set of items. MPT model parameters, θ_p, $p = 1, ..., P$, represent probabilities of latent cognitive capacities, such as attending to an item, storing an item in memory, retrieving an item from memory, detecting the source of an item, making an inference, or guessing a response. Such parameters are functionally independent and each has parameter space $[0, 1]$.

Parameter estimation and statistical inference for MPT models is traditionally carried out on response category frequencies aggregated over participants and items using maximum likelihood methods (e.g., Hu & Batchelder, 1994). This approach is based on the assumption of parameter homogeneity. If this assumption is violated, the analysis of aggregated data may lead to erroneous conclusions. The consequences of variability are especially troubling for nonlinear models, such as MPT models. In particular, reliance on aggregated data in the presence of parameter heterogeneity may lead to biased parameter estimates, the underestimation of confidence intervals, and the inflation of Type I error rates (e.g., Batchelder, 1975; Batchelder & Riefer, 1999; Heathcote, Brown, & Mewhort, 2000; Klauer, 2006; Riefer & Batchelder, 1991; Rouder & Lu, 2005). Moreover, the specific pattern of the parameter correlations can greatly influence the magnitude of the deleterious effects of unmodeled parameter heterogeneity (Klauer, 2006).

In recent years, a growing number of researchers has started to use cognitive models that accommodate heterogeneity in participants and/or items (e.g., DeCarlo, 2002; Karabatsos & Batchelder, 2003; M. D. Lee, 2011; M. D. Lee & Webb, 2005; Navarro, Griffiths, Steyvers, & Lee, 2006; Rouder & Lu, 2005; Rouder et al., 2007, 2003). In the context of MPT models, Klauer (2006) and J. B. Smith and Batchelder (2008) proposed statistical tests for detecting parameter heterogeneity. Moreover, a number of approaches that deal with parameter heterogeneity are now available for MPT models.

These approaches rely on hierarchical modeling and postulate population-level (hyper) distributions for the model parameters. The population-level distributions describe the variability in parameters either across participants or across items (e.g., Gelman et al., 2003; Gelman & Hill, 2007; Gill, 2002). For instance, Klauer (2006; see also Stahl & Klauer, 2007) proposed the use of latent-class MPT models with discrete population-level distributions to model the between-participant variability and the correlations between the model parameters. In contrast, J. B. Smith and Batchelder (2010) proposed to capture the between-participant variability of the model parameters using independent beta distributions (see also Batchelder & Riefer, 2007; Karabatsos & Batchelder, 2003; Riefer & Batchelder, 1991).

Here we will focus on yet another alternative —the latent-trait approach— that assumes a multivariate normal distribution for the participant differences in the probit transformed parameters and accounts for the correlations between the model parameters (Klauer, 2010). The latent-trait approach relies on Bayesian parameter estimation, but the MCMC algorithm for estimating the model parameters is currently not implemented in any off-the-shelf software package.

All the above described alternatives deal with parameter heterogeneity as a result of differences either between participants or between items, but not both, and rely on data that are aggregated either over items or over participants. It is, however, often reasonable to assume that the model parameters vary between participants as well as between items. In such situations, participant and item differences should be modeled as crossed-random effects (Clark, 1973) and inference should be based on participant-by-item data.

In psychometrics, there is a long tradition of simultaneously modeling variability in participants and items (e.g., De Boeck, 2008; Lord & Novick, 1986). In cognitive psychology, in contrast, such modeling constitutes a relatively recent trend (e.g., Baayen, 2008). For instance, Rouder and Lu (2005) and Rouder et al. (2007) have recently developed hierarchical signal detection models that incorporate random participant and item effects. In MPT modeling, attempts to simultaneously model heterogeneity in participants and items are scarce.

Augmenting MPT models with participant and item variability requires a separate parameter for each participant-item combination, θ_{ijp}, where $i = 1, ..., I$ indexes the participants, $j = 1, ..., J$ indexes the items, and $p = 1, ..., P$ indexes the model parameters in $\boldsymbol{\theta} = (\theta_{ijp})$. This requirement leads to $I \times J \times P$ parameters for only $I \times J$ data points, resulting in problems with model identification. We can reduce the number of parameters by using, for example, a reparametrization of the two-parameter Rasch model (e.g., Fischer & Molenaar, 1995). We can then model each participant-item combination using

$$\theta_{ijp} = \frac{\alpha_{ip}\beta_{jp}}{\alpha_{ip}\beta_{jp} + (1-\alpha_{ip})(1-\beta_{jp})}, \tag{5.1}$$

for $\alpha_{ip}, \beta_{jp} \in (0, 1)$ (Batchelder, 1998, 2009). Here α_{ip} and β_{jp} denote the i^{th} participant effect and the j^{th} item effect relating to parameter p, respectively. Karabatsos and Batchelder (2003) developed this Rasch model approach for the General Condorcet MPT Model. Batchelder and Crowther (1997) also used a Rasch model decomposition and modeled the logit transformed participant-item parameters as additive functions of the participant and item effects. See De Boeck and Partchev (2012) for an alternative approach to model heterogeneity in participants and items in MPT models using item response theory.

In the present paper we will explore an alternative that extends Klauer's (2010) latent-trait approach to simultaneously deal with heterogeneity in participants and items. Specifically, we will model the probit transformed θ_{ijp} parameters as additive combinations of participant and item effects. The participant and item effects are then assumed to come from (multivariate) normal distributions. Rouder et al. (2008) used a similar approach for a simple hierarchical process dissociation model, where they assumed the additivity of the probit transformed participant and item effects and modeled these using multivariate normal priors (see also Rouder & Lu, 2005; Rouder et al., 2007).

To summarize, a number of hierarchical approaches are now available for MPT models to deal with heterogeneity introduced either by the participants or by the items. The latest among these methods, Klauer's (2010) latent-trait approach, assumes a multivariate normal distribution for the probit transformed parameters and incorporates the possibility of parameter correlations. The latent-trait approach deals with parameter heterogeneity as a result of differences either between

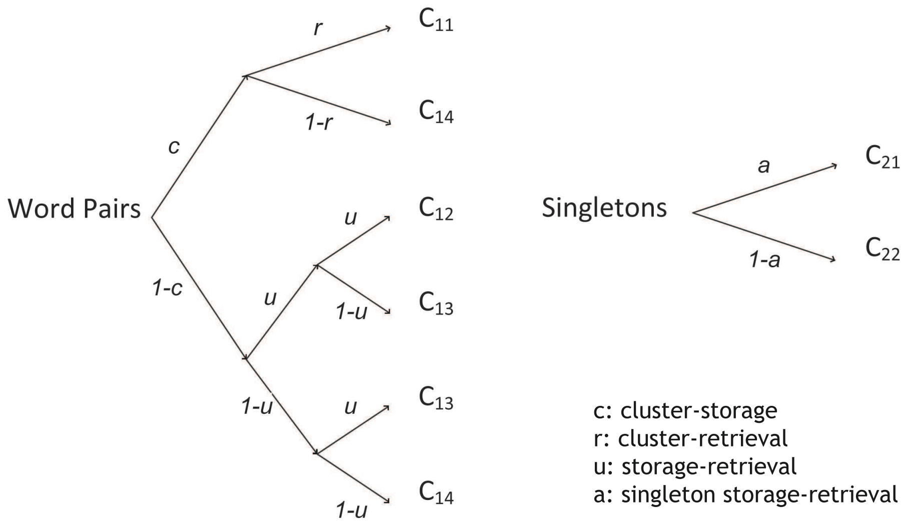


Figure 5.1 *Multinomial processing tree for the pair-clustering paradigm.*

participants or between items, but not for both sources. The latent-trait approach may readily be augmented to accommodate crossed-random effects by assuming additivity of participant and item effects on the probit scale.

5.3 The Pair-Clustering MPT Model

The pair-clustering model —one of the most extensively studied MPT models— was developed for the measurement of the storage and retrieval processes that underlie performance in the pair-clustering paradigm (e.g., Batchelder & Riefer, 1980, 1986). The pair-clustering paradigm involves a free recall memory experiment, where participants study a list of words that consists of two types of items: semantically related word pairs (e.g., dog-cat, father-son) and singletons (i.e., unpaired words, such as paper and train). Participants are presented with the study list in a word-by-word fashion, such as dog - paper - father - train - cat - son - etc. After the presentation of the study list, participants are required to recall, in any order, as many words as they can. The general finding is that semantically related word pairs are recalled consecutively, as a ‘pair-cluster’.

Since its development, the pair-clustering model has facilitated the interpretation of numerous free recall phenomena, such as retroactive inhibition and the effects of presentation rate and stimulus spacing (see Batchelder & Riefer, 1999). Moreover, the pair-clustering model has been used successfully to investigate memory deficits in various age groups and clinical populations (e.g., Bröder, Herwig, Teipel, & Fast, 2008; Golz & Erdfelder, 2004; Riefer & Batchelder, 1991; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002; see Batchelder & Riefer, 2007 for a review).

The architecture of the pair-clustering model can be represented by a rooted tree structure shown in Figure 5.1. The responses of each participant fall into two independent category systems, namely responses to word pairs and responses to singletons. Each category system $k = 1, 2$ is modeled by a separate subtree of the multinomial model, where each subtree consists of a finite number of branches terminating in one of the response categories C_{kl}, $l = 1, ..., L_k$. The recall of word pairs is scored into four response categories ($L_1 = 4$): C_{11}, both members of a word pair are recalled consecutively; C_{12}, both members of a word pair are recalled but not consecutively; C_{13}, only one member of a word pair is recalled; and C_{14}, neither member of a word pair is recalled. The recall of singletons is scored in two response categories ($L_2 = 2$): C_{21}, singleton is recalled; and C_{22}, singleton is not recalled.

The pair-clustering model explains the observed data by reparametrizing the category probabilities, $Pr(C_{kl})$, of the multinomial distribution in terms of $p = 1, ..., 4$ functionally independent model parameters $\boldsymbol{\theta} = (c, r, u, a)$, with $\theta_p \in (0, 1)$. Parameter c represents the probability that a word pair is clustered and stored in memory. Parameter r is the conditional probability that a word pair is retrieved from memory, given that it was clustered. Parameter u is the conditional probability that a member of a word pair is stored and retrieved from memory, given that the word pair was not stored as a cluster. As the u parameter taps both the storage and retrieval of unclustered words, it is typically regarded as a nuisance parameter. Parameter a is the probability that a singleton is stored and retrieved from memory. As illustrated later, it is frequently assumed that $a = u$, i.e., the probability that a singleton is stored and retrieved (a) equals the probability that a member of a word pair is stored and retrieved, given that it was not clustered (u). The pair-clustering model has four free response categories and it features at most four model parameters. The identification of the pair-clustering model has been established elsewhere (e.g., Batchelder & Riefer, 1986).

According to the model, if a word pair is successfully clustered and retrieved with joint probability cr, the two members of the word pair are retrieved consecutively, resulting in recall category C_{11}. If a word pair is successfully clustered (c) but is not retrieved (1-r), neither member of the word pair is retrieved, resulting in recall category C_{14}. The model thus assumes that clustered pairs are either retrieved as a pair or are not retrieved at all. In contrast, if word pairs are not clustered (1-c), either member of the word pair can be stored and retrieved independently with probability u, resulting in recall category C_{12} or C_{13}. Retrieved items from unclustered word pairs are thus not recalled consecutively.

The probabilities of the six response categories are expressed in terms of the model parameters as follows:

$$\begin{aligned}
Pr(C_{11}|\boldsymbol{\theta}) &= cr \\
Pr(C_{12}|\boldsymbol{\theta}) &= (1-c)u^2 \\
Pr(C_{13}|\boldsymbol{\theta}) &= (1-c)2u(1-u) \\
Pr(C_{14}|\boldsymbol{\theta}) &= c(1-r) + (1-c)(1-u)^2 \\
Pr(C_{21}|\boldsymbol{\theta}) &= a \\
Pr(C_{22}|\boldsymbol{\theta}) &= 1-a.
\end{aligned} \tag{5.2}$$

The raw data in category system k consist of the response of a given participant $i = 1, ..., I$ to a particular item $j = 1, ..., J_k$, represented by a vector of length L_k. For a given participant-word pair combination, the raw data $\mathbf{n}_{ij,1}$ thus consist of a vector of length $L_1 = 4$, where the entry n_{ijl} equals 1 if the response of participant i to word pair j falls into response category l, and zero otherwise. For example, if participant i recalls both members of word pair j consecutively (i.e., response category C_{11}), the raw data are given by the vector $(1, 0, 0, 0)$. Similarly, for a given participant-singleton combination, the raw data $\mathbf{n}_{ij,2}$ consist of a vector of length $L_2 = 2$, where n_{ijl} equals 1 if the response of participant i to singleton j falls into response category l, and zero otherwise. For example, if participant i does not recall singleton j (i.e., response category C_{22}), the raw data are given by the vector $(0, 1)$. Traditional analysis of pair-clustering data assumes that observations over participant and items are independent and identically distributed. Parameter estimation is generally carried out on category responses summed over participants and items (e.g., Batchelder & Riefer, 1986).

5.4 The Latent-Trait Pair-Clustering Model

The main goal of the present paper is to augment Klauer's (2010) Bayesian latent-trait approach to handle heterogeneity in both participants and items. To facilitate this, we first introduce the latent-trait approach in more detail and provide a WinBUGS implementation of the latent-trait pair-clustering model. We then report the results of a parameter recovery study. In what follows we assume that the items are homogeneous and use the latent-trait approach to model individual differences between participants. Note, however, that the latent-trait approach may just as well be used to capture the variability between items instead of participants. In this case, we would assume that participants are homogeneous and model the differences between the items.

Introduction to the Latent-Trait Approach

The symbols and notation used in the text, figures, and the WinBUGS scripts are summarized in Table 5.1. As we focus on parameter heterogeneity as a result of individual differences between participants, the raw data are aggregated over the J_1 word pairs and the J_2 singletons but not over the $i = 1, ..., I$ participants. The data of participant i consist thus of the frequency of responses, n_{ikl}, falling into recall category C_{kl}, $k = 1, 2$, $l = 1, ..., L_k$.

For each participant i in each category system k, the observed category frequencies are assumed to follow a multinomial distribution with category probabilities $Pr(C_{kl}|\boldsymbol{\theta}_i)$. Formally, let B_{klm} be the m^{th} branch terminating in C_{kl}, $m = 1, ..., M_{kl}$. The probability that participant i follows branch B_{klm} is given by

$$Pr(B_{klm}|\boldsymbol{\theta}_i) = \prod_{p=1}^{P} \theta_{ip}^{v_{klmp}} (1-\theta)_{ip}^{w_{klmp}}, \tag{5.3}$$

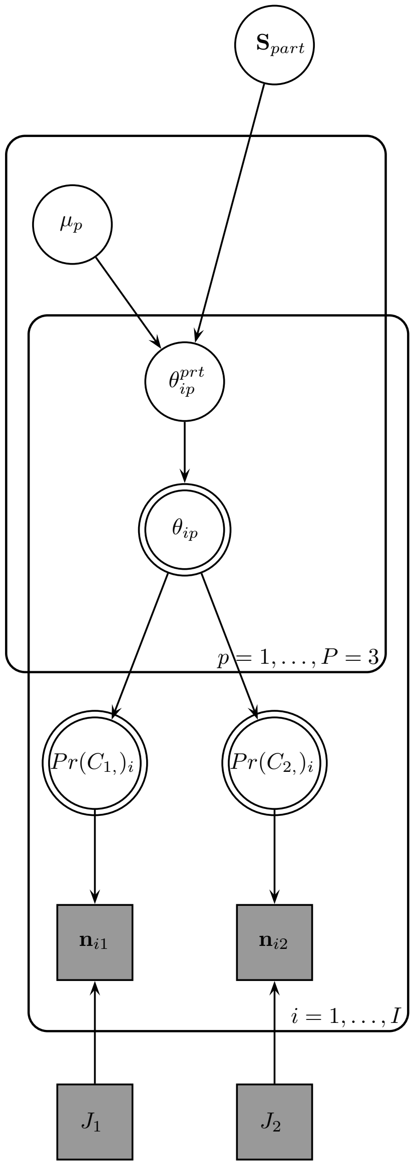
where $v_{klmp} \geq 0$ and $w_{klmp} \geq 0$ are the number of nodes on branch B_{klm} that is associated with parameter θ_p, $p = 1, ..., P$, and $1 - \theta_p$, respectively. The probability of each response category is given by adding the probabilities of all the branches that lead to that category:

$$Pr(C_{kl}|\boldsymbol{\theta}_i) = \sum_{m=1}^{M_{kl}} \prod_{p=1}^{P} \theta_{ip}^{v_{klmp}} (1-\theta_{ip})^{w_{klmp}}. \tag{5.4}$$

The data of participant i across the two category systems, $\mathbf{n}_i = (\mathbf{n}_{i1}, \mathbf{n}_{i2})$, are assumed to follow a multinomial distribution:

$$Pr(\mathbf{N}_i = \mathbf{n}_i|\boldsymbol{\theta}_i) = \prod_{k=1}^{K} \left\{ \frac{J_k!}{n_{ik1}! \times n_{ik2}! \times ... \times n_{ikL_k}!} \prod_{l=1}^{L_k} [Pr(C_{kl}|\boldsymbol{\theta}_i)]^{n_{ikl}} \right\}. \tag{5.5}$$

The latent-trait approach relies on Bayesian hierarchical modeling that allows the individual model parameters θ_{ip} to vary over participants in a statistically specified way. The method postulates a multivariate normal distribution to capture the between-participant variability and the correlations between the model parameters. The latent-trait approach relies on MCMC sampling to approximate the posterior distributions of the model parameters. In what follows, we present an easy-to-use WinBUGS implementation of the latent-trait approach that enables researchers to obtain samples from the posterior distribution of the model parameters.



$$\mathbf{S}_{part} \sim \text{Scaled} - \text{Inverse} - \text{Wishart}\big(\mathbf{W}, df = P+1, \boldsymbol{\xi}_{part}\big)$$
$$\xi_{part_p} \sim \text{Uniform}\big(0, 100\big)$$
$$\mu_p \sim \text{Normal}\big(0, 1\big)$$

$$\boldsymbol{\theta}_i^{prt} \sim \text{Multivariate} - \text{Normal}\Big(\left(\mu_1, \ldots, \mu_P\right), \mathbf{S}_{part}^{-1}\Big)$$
$$\theta_{ip} = \phi\left(\theta_{ip}^{prt}\right)$$

$$Pr(C_{11})_i = \theta_{i1} \times \theta_{i2}$$
$$Pr(C_{12})_i = (1 - \theta_{i1}) \times \theta_{i3}^2$$
$$Pr(C_{13})_i = (1 - \theta_{i1}) \times 2 \times \theta_{i3} \times (1 - \theta_{i3})$$
$$Pr(C_{14})_i = \theta_{i1} \times (1 - \theta_{i2}) + (1 - \theta_{i1}) \times (1 - \theta_{i3})^2$$

$$Pr(C_{21})_i = \theta_{i3}$$
$$Pr(C_{22})_i = (1 - \theta_{i3})$$

$$\mathbf{n}_{i1} \sim \text{Multinomial}\left(Pr(C_{1,})_i, J_1\right)$$
$$\mathbf{n}_{i2} \sim \text{Multinomial}\left(Pr(C_{2,})_i, J_2\right)$$

Figure 5.2 *Graphical model for the latent-trait pair-clustering model.* $\theta_{i1} = c_i$, $\theta_{i2} = r_i$, and $\theta_{i3} = u_i$. Note. To maintain consistency with the WinBUGS syntax, the multivariate normal and independent normal distributions are parametrized in terms of the precision (i.e., inverse variance).

WinBUGS Implementation of the Latent-Trait Pair-Clustering Model

The graphical model for the WinBUGS implementation of the latent-trait pair-clustering model is shown in Figure 5.2. Observed variables are represented by shaded nodes and unobserved variables are represented by unshaded nodes. Continuous variables are represented by circles and discrete variables are represented by squares. The graph structure indicates dependencies between the nodes and the plates represent independent replications (e.g., M. D. Lee, 2008). The graphical model depicts the basic pair-clustering model for I participants responding to J_1 word pairs and J_2 singletons, with the constraint that $a = u$. The corresponding WinBUGS script is available in the supplemental materials at `http://dora.erbe-matzke.com/publications.html`.

Table 5.1 Notation.

Notation	Explanation
K	Number of category systems
L_k	Number of response categories in category system k
I	Number of participants
J_k	Number of items in category system k
P_k	Number of parameters in category system k
C_{kl}	Response category l in category system k
M_{kl}	Number of branches terminating in C_{kl}
B_{klm}	m^{th} branch terminating in C_{kl}
$n_{ij,kl}$	Response (i.e., 0 or 1) of participant-item combination ij in C_{kl}
θ_{ijp_k}	Parameter p of participant-item combination ij in category system k (i.e., c, r, u for $k = 1$; a for $k = 2$)
$v_{kl,mp}$	Number of nodes on B_{klm} associated with θ_{p_k}
$w_{kl,mp}$	Number of nodes on B_{klm} associated with $1 - \theta_{p_k}$
$\theta^{prt}_{ijp_k}$	Probit transformed parameter p of participant-item combination ij in category system k
μ_{p_k}	Group mean for parameter $\theta^{prt}_{p_k}$
$\mu_{\mu_{p_k}}$	Mean of normal prior for μ_{p_k}
$\sigma_{\mu_{p_k}}$	Standard deviation of normal prior for μ_{p_k}
$\delta^{raw}_{part_{ip_k}}$	i^{th} unscaled participant effect relating to parameter p_k
$\xi_{part_{p_k}}$	Scaling factor for the participant effects relating to parameter p_k
$\delta_{part_{ip_k}}$	i^{th} scaled participant effect relating to parameter p_k
$\mathbf{T}_{part}$	Unscaled variance-covariance matrix of participant effects
$\mathbf{S}_{part}$	Scaled variance-covariance matrix of participant effects
$\sigma_{part_{p_k}}$	Scaled standard deviation of participant effects relating to parameter p_k
$\rho_{part_{p_k p'_k}}$	Correlation between participant effects relating to parameter p_k and p'_k
$\delta^{raw}_{item_{jp_k}}$	j^{th} unscaled item effect relating parameter p_k
$\xi_{item_{p_k}}$	Scaling factor for the item effects relating to parameter p_k
$\delta_{item_{jp_k}}$	j^{th} scaled item effect relating to parameter p_k
$\mathbf{T}_{item}$	Unscaled variance-covariance matrix of item effects*
$\mathbf{S}_{item}$	Scaled variance-covariance matrix of item effects*
$\lambda_{item_{p_k}}$	Unscaled standard deviation of item effects relating to parameter p^{**}_k
$\sigma_{item_{p_k}}$	Scaled standard deviation of item effects relating to parameter p_k
$\rho_{item_{p_k p'k}}$	Correlation between item effects relating to parameter p_k and $p\prime^{*}_k$

Note. For the latent-trait approach, the k subscript of the parameter index p is suppressed throughout the text because $u_i = a_i$. The * indicates item parameters that are used only for the real data example featuring correlated item effects. The ** indicates item parameters that are used only for the parameter recovery study featuring uncorrelated item effects.

Data

For each participant, the data for word pairs, $\mathbf{n}_{i1}$, follow a multinomial distribution, with category probabilities $Pr(C_{11}|\boldsymbol{\theta}_i)$, $Pr(C_{12}|\boldsymbol{\theta}_i)$, $Pr(C_{13}|\boldsymbol{\theta}_i)$, $Pr(C_{14}|\boldsymbol{\theta}_i)$, and J_1. For each participant, the data for singletons, $\mathbf{n}_{i2}$, follow a multinomial distribution with $Pr(C_{21}|\boldsymbol{\theta}_i)$, $Pr(C_{22}|\boldsymbol{\theta}_i)$, and J_2.

Prior Distributions

The basic model depicted in Figure 5.2 assumes three parameters per participant ($P = 3$): $\boldsymbol{\theta}_i = (c_i, r_i, u_i)$. Thus, we assume that $a_i = u_i$. The individual model parameters θ_{ip} are transformed from the probability scale to the real line using a probit link so that the transformed parameters θ^{prt}_{ip} are given by $\Phi^{-1}(\theta_{ip})$, where Φ is the standard normal cumulative distribution function. The use of probit transformed probabilities has a long history in psychometrics, and is also common practice in Bayesian cognitive modeling (e.g., Rouder & Lu, 2005; Rouder et al., 2008, 2007). To model participant heterogeneity and parameter correlations, we assume that the probit transformed parameters $\boldsymbol{\theta}_i^{prt}$ follow a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\mathbf{S}_{part}$. The θ_{ip}^{prt} parameters are reparametrized as follows:

$$\theta_{ip}^{prt} = \mu_p + \delta_{part_{ip}}, \tag{5.6}$$

where μ_p is the group mean for parameter p and $\delta_{part_{ip}}$ is the i^{th} participant's deviation from it. The $\boldsymbol{\delta}_{part_i}$ parameters are then drawn from a zero-centered multivariate distribution with variance-covariance matrix $\mathbf{S}_{part}$.

Hyper-Prior Distributions

The population-level $\boldsymbol{\mu}$ and $\mathbf{S}_{part}$ parameters are estimated from the data and therefore require prior distributions. The priors for the μ_p parameters are independent normal distributions with $\mu_{\mu_p} = 0$ and $\sigma^2_{\mu_p} = 1$. Note that the original formulation of the latent-trait approach (Klauer, 2010) assumes independent normal distributions with $\mu_{\mu_p} = 0$ and $\sigma^2_{\mu_p} = 100$. However, we prefer to use $\sigma^2_{\mu_p} = 1$ because it corresponds to a uniform distribution on the probability scale (Rouder & Lu, 2005).

The prior for the variance-covariance matrix $\mathbf{S}_{part}$ is a scaled Inverse-Wishart distribution. The Inverse-Wishart is a frequently used prior for variance-covariance matrices (Gelman & Hill, 2007). The Inverse-Wishart prior has two parameters: the degrees of freedom that is set to one plus the number of free participant parameters $(1 + P)$ and the scale matrix that is set to the $P \times P$ identity matrix $(\mathbf{W})$. The advantage of the Inverse-Wishart is that it results in an uninformative uniform prior distribution between -1 and 1 for the $\rho_{pp'}$ correlation parameters. The disadvantage is that the Inverse-Wishart with $1 + P$ degrees of freedom imposes a very restrictive prior on the standard deviations. To be able to estimate the standard deviations more freely, we augment the Inverse-Wishart with a set of scale parameters, $\boldsymbol{\xi}_{part} = [\xi_{part_1}, ..., \xi_{part_P}]$ (Gelman & Hill, 2007). The resulting scaled Inverse-Wishart distribution still implies a uniform prior distribution for the correlation parameters, but it allows the standard deviations to be estimated more freely than does the Inverse-Wishart. The variance-covariance matrix $\mathbf{S}_{part}$ is then modeled as

$$\mathbf{S}_{part} = \text{Diag}(\boldsymbol{\xi}_{part})\ \mathbf{T}_{part}\ \text{Diag}(\boldsymbol{\xi}_{part}), \tag{5.7}$$

where Diag($\boldsymbol{\xi}$) is a diagonal matrix containing the scale parameters. $\mathbf{T}_{part}$ follows an Inverse-Wishart distribution with $1 + 3$ degrees of freedom, with a scale matrix that is set to the 3×3 identity matrix. The standard deviations can be obtained by

$$\sigma_{part_p} = |\xi_{part_p}| \times \sqrt{T_{part_{pp}}}. \tag{5.8}$$

The correlation parameters are given by

$$\rho_{part_{pp'}} = \frac{\xi_{part_p}\xi_{part_{p'}}T_{part_{pp'}}}{|\xi_{part_p}|\sqrt{T_{part_{pp}}} \times |\xi_{part_{p'}}|\sqrt{T_{part_{p'p'}}}}. \tag{5.9}$$

The ξ_{part_p} parameters are given uniform distributions ranging from 0 to 100 (e.g., Gelman & Hill, 2007). Klauer (2010) used normal distributions with a mean of one and a variance of 100 as prior for the scaling parameters. In our WinBUGS implementation, these priors occasionally resulted in convergence problems for the variance and the correlation parameters. Note that the use of redundant multiplicative parameters, such as ξ_{part_p}, has been reported to increase the rate of convergence in hierarchical models (Gelman & Hill, 2007). As a result of the new parametrization, Equation 5.6 can be reformulated as follows:

$$\theta_{ip}^{prt} = \mu_p + \xi_{part_p} \times \delta_{part_{ip}}^{raw}, \tag{5.10}$$

where μ_p is the group mean for parameter p, ξ_{part_p} is the scaling factor of the scaled Inverse-Wishart distribution, and $\delta_{part_{ip}}^{raw}$ is the i^{th} participant's unscaled deviation from the group mean.

Parameter Recovery Study

We conducted a series of parameter recovery studies to assess whether the WinBUGS implementation of the latent-trait pair-clustering model adequately recovers true parameter values. Here we report the results of a study where we generated free recall data for synthetic participants responding to a set of word pairs and singletons in two sessions of the pair-clustering task. The resulting datasets were fit with the latent-trait pair-clustering model using WinBUGS.

Methods

Each synthetic participant performed the pair-clustering task two consecutive times. For each participant, the data from the two sessions were scored into four category systems: word pairs and singletons for the first session and word pairs and singletons for the second session. We ran three sets of simulations, each comprising 100 datasets. First, each data set contained observations from 63 ($I = 63$) synthetic participants, responding to 10 word pairs ($J_1 = 10$) and 5 singletons ($J_2 = 5$) in each of the two sessions. Second, each data set contained observations from 63 participants, responding to 20 word pairs and 10 singletons in each of the two sessions. Third, each data set contained observations from 126 participants, responding to 10 word pairs and 5 singletons in each of the two sessions.

Similar to Klauer's (2010) recovery study, we used five parameters ($P = 5$) per participant across the two sessions: $\boldsymbol{\theta}_i = (c_{1_i}, r_i, u_{1_i}, c_{2_i}, u_{2_i})$. The following parameter constraints were imposed: $r_{1_i} = r_{2_i}$, $a_{1_i} = u_{1_i}$, and $a_{2_i} = u_{2_i}$. The generating population-level parameter values are shown in Figure 5.3. We conducted several recovery studies using alternative true parameter values. The results were essentially the same as the ones reported here. Note that the details of the recovery study, including the true parameter values and the number of participants and items, are identical to those used in Klauer's paper.

For each analysis reported in this article, we ran three MCMC chains and used randomly generated overdispersed starting values to confirm that the chains have converged to the stationary distribution. Convergence is confirmed if the individual chains are indistinguishable from each other. Convergence was formally assessed with the $\hat{R}$ statistic (Brooks & Gelman, 1998; Gelman & Rubin, 1992), a quantitative measure of convergence that compares the within-chain variance to the between-chain variance. The results reported in this article are based on analyses where $\hat{R}$ for all parameters of interest (i.e., group means, random effects, and the standard deviation and the correlation of the random effects) is lower than 1.05. In light of the possibility of high autocorrelations between successive MCMC samples, we ran relatively long MCMC chains and thinned each chain by retaining samples from only every 3^{rd} iteration.

The latent-trait pair-clustering model was fit to the synthetic datasets using WinBUGS. For each data set, we discarded the first 2,000 samples of each chain as burn-in and based inference on a total of 54,000 recorded samples.

Results

The results of the recovery study for the group-level parameters are shown in Figure 5.3. We follow Klauer's (2010) practice and use the median and the standard deviation to summarize the posterior distribution of the parameters. Also, the posterior median is often preferable over the posterior mode or the posterior mean for non-symmetric or heavy tailed posterior distributions. Note that the group c_1, r, u_1, c_2, and u_2 parameters are reported on the probability scale, while their standard deviations and correlations are reported on the probit scale. The group parameters and their standard deviations are recovered relatively well using the posterior median even for the first set of simulations with relatively few participants and very few items. Naturally, as the number of items or the number of participants increases, the bias, the posterior standard deviation, and the standard error of the recovered parameters decrease. The storage-retrieval u_1 and u_2 parameters and their standard deviations are estimated most precisely, as indicated by the small posterior standard deviation of the estimates. The cluster-retrieval r parameter and its standard deviation are estimated the least precisely as evidenced by the greater posterior uncertainty of the estimates, especially for the first set of simulations.

With respect to the correlation parameters, the results are less clear-cut. Similar to Klauer's (2010) findings, the posterior median underestimates the parameter correlations especially in datasets with few participants and items. The posterior standard deviations are rather large, indicating large uncertainty in the estimates. Nevertheless, as the number of participants or the number of items increases, the bias, the posterior standard deviations and the standard error of the recovered correlations decrease. As for the standard deviations, correlations involving the cluster-retrieval r parameter are the least well estimated, especially for the first set of simulations.

To sum up, the results of the simulation study indicated that the WinBUGS version of the latent-trait pair-clustering model adequately recovered the true parameter values. In the next section, we extend the latent-trait pair-clustering model and the corresponding WinBUGS script to handle heterogeneity in both participants and items.

5.5 The Crossed-Random Effects Pair-Clustering Model

In many applications of MPT models, it is reasonable to assume that the model parameters do not only differ between participants but also between the items used in a particular experiment. We should then treat both participant and items effects as random, define parameters for each participant-item combination and base statistical inference on the unaggregated data. This section introduces a crossed-random effects pair-clustering model that is based on an extension of Klauer's (2010) latent-trait approach. Our crossed-random effects model assumes that the participant and item effects combine additively on the probit scale. The participant and item effects are modeled with multivariate normal and independent normal distributions, respectively, with means and (co)variances estimated from the data.

Introduction to the Crossed-Random Effects Approach

In the crossed-random effects pair-clustering model, statistical inference is based on unaggregated participant-by-item data. In a given category system k, $k = 1, 2$, the raw category responses of each participant-item combination, $i = 1, ..., I$, $j = 1, ..., J_k$, are assumed to follow a multinomial distribution with category probabilities $Pr(C_{kl}|\boldsymbol{\theta}_{ij_k})$, $l = 1, ..., L_k$, where $\boldsymbol{\theta}_{ij_k}$ contains the $p = 1, ..., P_k$ model parameters of participant-item combination ij in category system k.

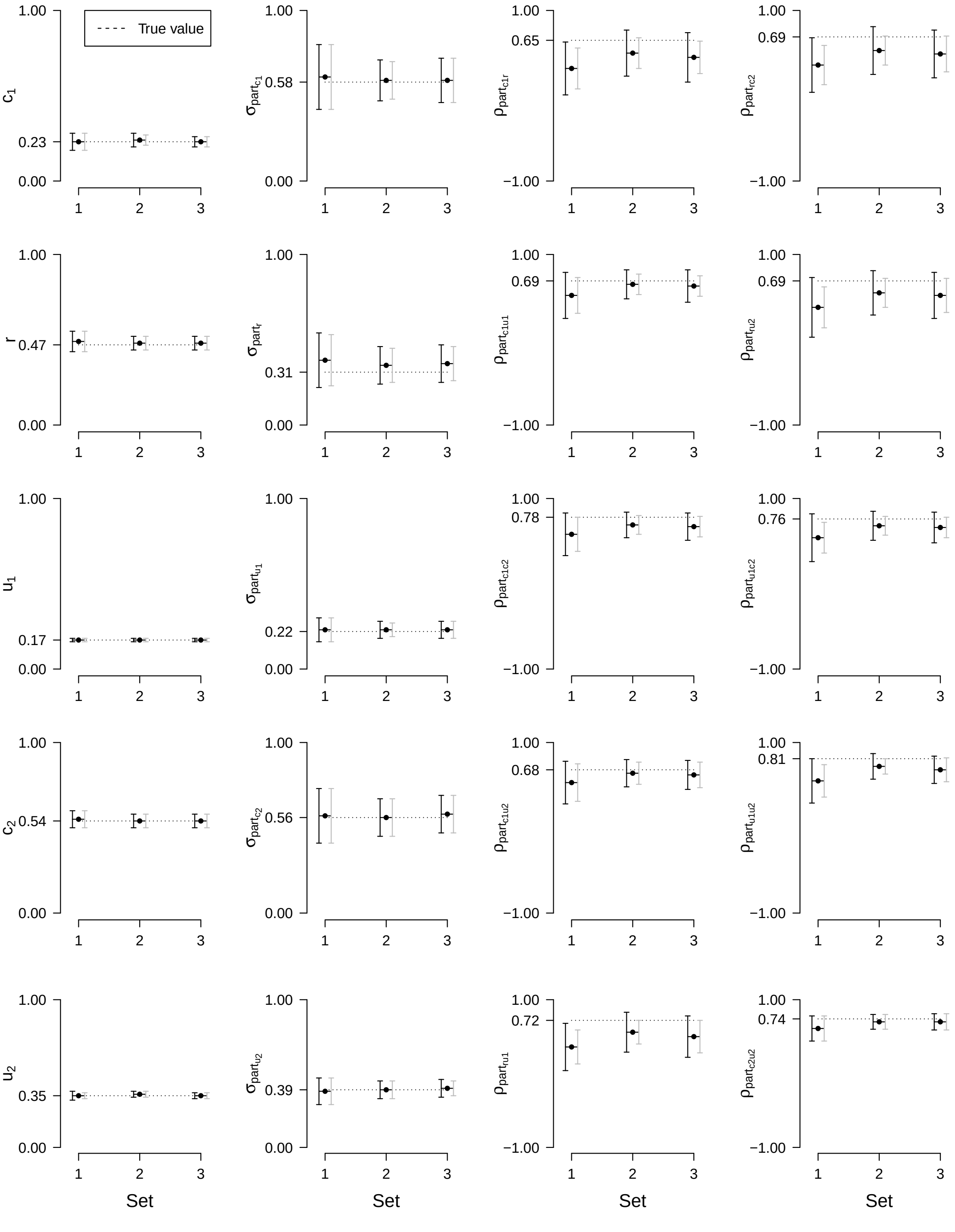


Figure 5.3 *Posterior medians from the parameter recovery study for the latent-trait pair-clustering model using WinBUGS.* Each set of simulations consisted of 100 datasets. The black bullets indicate the mean of the posterior median of the parameters across the 100 replications. The black vertical lines are based on the mean of the posterior standard deviation across the 100 replications. The gray vertical lines indicate the standard error of the posterior median across the 100 replications.

The requirement of a separate parameter for each participant-item combination leads to a very large number of parameters, resulting in problems of model identification. To reduce the number of parameters, we assume that the probit transformed parameters are given by the additive combination of participant and item effects (e.g., Rouder & Lu, 2005; Rouder et al., 2008, 2007). More formally, the crossed-random effects pair-clustering model assumes that the probit transformed participant-item parameters in category system k are given by

$$\theta^{prt}_{ijp_k} = \mu_{p_k} + \delta_{part_{ip_k}} + \delta_{item_{jp_k}}, \tag{5.11}$$

where μ_{p_k} is the group mean for parameter p in category system k, and $\delta_{part_{ip_k}}$ and $\delta_{item_{jp_k}}$ are the i^{th} participant effect and the j^{th} item effect, respectively. We postulate a multivariate normal distribution to describe variability between participants and independent normal distributions to capture the variability between items. The participant effects are thus allowed to be correlated a priori, whereas the item effect are not. Naturally, we may model the correlations between the item effects —similar to the participant effects— using a multivariate normal distribution. The possibility to incorporate correlated participant *and* correlated item effects will be demonstrated shortly using experimental data. The next section presents an easy-to-use WinBUGS implementation of the crossed-random effects pair-clustering model.

WinBUGS Implementation of the Crossed-Random Effects Pair-Clustering Model

The graphical model for the WinBUGS implementation of the crossed-random effect pair-clustering model is shown in Figure 5.4. The graphical model depicts the basic pair-clustering model for I participants responding to J_1 word pairs and J_2 singletons. The corresponding WinBUGS script is available in the supplemental materials.

Data

The raw data of each participant-word pair combination, $\mathbf{n}_{ij,1}$, follow a multinomial distribution, with category probabilities $Pr(C_{11}|\boldsymbol{\theta}_{ij_1})$, $Pr(C_{12}|\boldsymbol{\theta}_{ij_1})$, $Pr(C_{13}|\boldsymbol{\theta}_{ij_1})$, $Pr(C_{14}|\boldsymbol{\theta}_{ij_1})$. Similarly, the raw data for each participant-singleton combination, $\mathbf{n}_{ij,2}$, follow a multinomial distribution, with category probabilities $Pr(C_{21}|\boldsymbol{\theta}_{ij_2})$, $Pr(C_{22}|\boldsymbol{\theta}_{ij_2})$.

Prior Distributions

The crossed-random effects pair-clustering model posits a separate parameter for each participant-item combination in each category system k. These θ_{ijp_k} parameters are transformed from the probability scale to the real line using the probit link. As given in Equation 5.11, the probit transformed parameters $\theta^{prt}_{ijp_k}$ are given by the additive combination of participant and item effects.

In the category system for word pairs, the model assumes three participant effects for each participant (i.e., $\delta_{part_{ic}}$, $\delta_{part_{ir}}$, and $\delta_{part_{iu}}$) and three item effects for each word pair (i.e., $\delta_{item_{jc}}$, $\delta_{item_{jr}}$, and $\delta_{item_{ju}}$). The model postulates thus three parameters for each participant-word pair combination ($P_1 = 3$): $\boldsymbol{\theta}_{ij_1} = (c_{ij}, r_{ij}, u_{ij})$. For singletons, the model assumes one participant effect per participant ($\delta_{part_{ia}}$) and one item effect per singleton ($\delta_{item_{ja}}$). The model postulates thus one parameter for each participant-singleton combination ($P_2 = 1$) : $\theta_{ij_2} = a_{ij}$.

In the basic pair-clustering model depicted in Figure 5.4, the constraint that $a = u$ may be implemented as follows. First, the group mean of the singleton storage-retrieval a parameter is constrained to be equal to the group mean of the storage-retrieval u parameter: $\mu_a = \mu_u$. Second,

$$\mathbf{S}_{part} \sim \text{Scaled} - \text{Inverse} - \text{Wishart}\big(\mathbf{W}, df = P+1, \boldsymbol{\xi}_{part}\big)$$
$$\xi_{part_p} \sim \text{Uniform}\big(0, 100\big)$$
$$\sigma^2_{item_p} \sim \text{Scaled} - \text{Inverse} - \text{Gamma}\big(1, 1, \xi_{item_p}\big)$$
$$\xi_{item_p} \sim \text{Uniform}\big(0, 100\big)$$
$$\mu_p \sim \text{Normal}\big(0, 1\big)$$
$$\boldsymbol{\delta}_{part_i} \sim \text{Multivariate} - \text{Normal}\Big(\left(0,0,0\right), \mathbf{S}^{-1}_{part}\Big)$$
$$\delta_{item_{jp}} \sim \text{Normal}\big(0, \sigma^2_{item_p}{}^{-1}\big)$$
$$\delta_{item_{ja}} \sim \text{Normal}\big(0, \sigma^2_{item_3}{}^{-1}\big)$$

$$\theta^{prt}_{ijp} = \mu_p + \delta_{part_{ip}} + \delta_{item_{jp}}$$
$$\theta^{prt}_{ija} = \mu_3 + \delta_{part_{i3}} + \delta_{item_{ja}}$$
$$\theta_{ijp} = \phi\left(\theta^{prt}_{ijp}\right)$$
$$\theta_{ija} = \phi\left(\theta^{prt}_{ija}\right)$$

$$Pr(C_{11})_{ij} = \theta_{ij1} \times \theta_{ij2}$$
$$Pr(C_{12})_{ij} = (1 - \theta_{ij1}) \times \theta^2_{ij3}$$
$$Pr(C_{13})_{ij} = (1 - \theta_{ij1}) \times 2 \times \theta_{ij3} \times (1 - \theta_{ij3})$$
$$Pr(C_{14})_{ij} = \theta_{ij1} \times (1 - \theta_{ij2}) + (1 - \theta_{ij1}) \times (1 - \theta_{ij3})^2$$

$$Pr(C_{21})_{ij} = \theta_{ija}$$
$$Pr(C_{22})_{ij} = (1 - \theta_{ija})$$

$$\mathbf{n}_{ij1} \sim \text{Multinomial}\left(Pr(C_{1,})_{ij}, 1\right)$$
$$\mathbf{n}_{ij2} \sim \text{Multinomial}\left(Pr(C_{2,})_{ij}, 1\right)$$

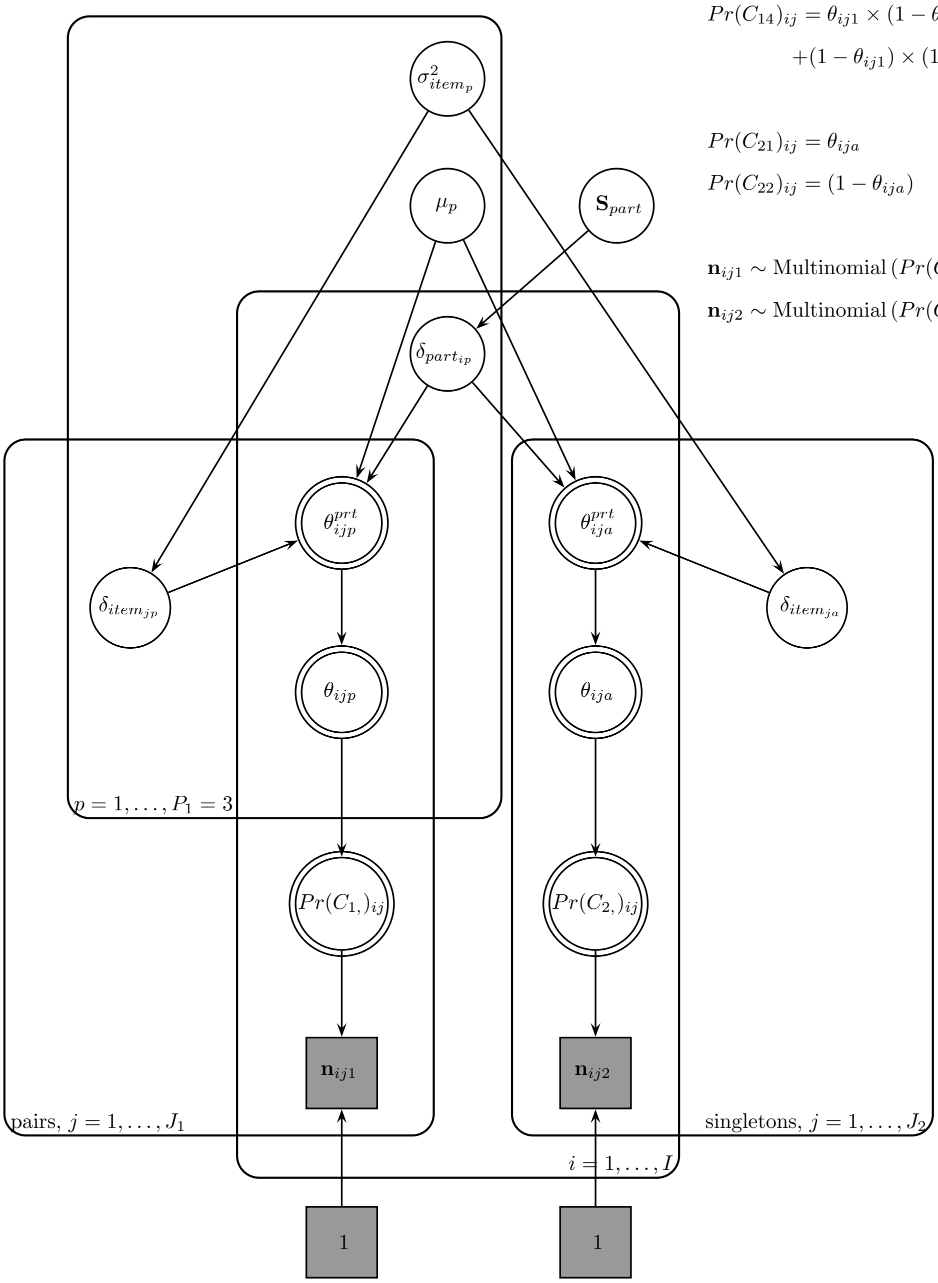


Figure 5.4 *Graphical model for the crossed-random effects pair-clustering model.* $\theta_{ij1} = c_{ij}$, $\theta_{ij2} = r_{ij}$, $\theta_{ij3} = u_{ij}$. Note. To maintain consistency with the WinBUGS syntax, the multivariate normal and independent normal distributions are parametrized in terms of the precision (i.e., inverse variance).

note that the basic model assumes that each participant is presented with J_1 word pairs and J_2 singletons. We are thus able to place across-category system constraints on the participant effects, because responses from a given participant are available in both category systems: $\delta_{part_{ia}} = \delta_{part_{iu}}$. Third, we are unable to place across-category system constraints on the items effects because a given item appears in only one of the category systems: responses to each of the J_1 word pairs are only available in the first category system, while responses to each of the J_2 singletons are only available in the second category system. Nevertheless, we may assume that the standard deviation of the item effects relating to a and u are equal: $\sigma_{item_a} = \sigma_{item_u}$. A possibility for across-category system constraints on the item effects will be illustrated shortly using experimental data.

The $\boldsymbol{\delta}_{part_i}$ parameters are assumed to come from a zero-centered multivariate normal distribution, with variance-covariance matrix $\mathbf{S}_{part}$ estimated from the data. The $\delta_{item_{jp_k}}$ parameters are drawn from zero-centered independent normal distributions, with the standards deviations $\sigma_{item_{p_k}}$ estimated from the data.

Hyper-Prior Distributions

The priors for the grand mean μ_{p_k} parameters are weakly informative independent normal distributions with $\mu_{\mu_{p_k}} = 0$ and $\sigma^2_{\mu_{p_k}} = 1$. The prior for $\mathbf{S}_{part}$ is a scaled Inverse-Wishart distribution. The degrees of freedom of the scaled Inverse-Wishart equals one plus the number of free participant effects. In the model shown in Figure 5.4, we postulate three participant effects across the two category systems, resulting in four degrees of freedom. The scale matrix is set to the 3×3 identity matrix ($\mathbf{W}$). The scaling factor ξ_{part} parameters of the Inverse-Wishart are given uniform distributions ranging from 0 to 100. The standard deviations and the correlations of the participant effects can be obtained using Equation 5.8 and 5.9, respectively.

The priors for the $\sigma^2_{item_{p_k}}$ variance parameters are independent scaled inverse gamma distributions with $\alpha = 1$ and $\beta = 1$. The inverse gamma distribution with α and β set to low values, such as 1, 0.01, or 0.001 is a frequently used prior for variance parameters (e.g., Spiegelhalter, Thomas, Best, Gilks, & Lunn, 2003). In order to increase the rate of convergence, we augment each variance parameter with a redundant multiplicative scaling parameter ξ_{item}, a technique called parameter expansion (Gelman & Hill, 2007). In the expanded model, the item standard deviations are given by

$$\sigma_{item_{p_k}} = |\xi_{item_{p_k}}| \times \lambda_{item_{p_k}}, \tag{5.12}$$

where $\xi_{item_{p_k}}$ is the scaling factor and $\lambda_{item_{p_k}}$ is the unscaled item standard deviation for parameter p in category system k. The ξ_{item} parameters are given uniform distributions ranging from 0 to 100. As a result of expanding the model with the ξ_{part} and ξ_{item} parameters, Equation 5.11, can be reformulated as follows:

$$\theta^{prt}_{ijp_k} = \mu_{p_k} + \xi_{part_{p_k}} \times \delta^{raw}_{part_{ip_k}} + \xi_{item_{p_k}} \times \delta^{raw}_{item_{jp_k}}, \tag{5.13}$$

where $\delta^{raw}_{part_{ip_k}}$ and $\delta^{raw}_{item_{jp_k}}$ are the unscaled effects for participant i and item j relating to parameter p in category system k, respectively.

Parameter Recovery Study

We conducted a series of parameter recovery studies to examine whether the crossed-random effects pair-clustering model adequately recovers true parameter values. Here we report the results of a study where we generated free recall data for synthetic participants responding to the same set of word pairs and the same set of singletons in two sessions of the pair-clustering task. We analyzed the resulting datasets with the crossed-random effects pair-clustering model using WinBUGS.

Methods

Each synthetic participant performed the pair-clustering task two consecutive times using the same set of word pairs and the same set of singletons. For each participant-word pair combination, the data from the two sessions were scored into two separate category systems. Similarly, for each participant-singleton combination, the data from the two sessions were scored into two separate category systems. We conducted three sets of simulations, each comprising 100 synthetic datasets. First, each data set contained observations from 63 ($I = 63$) synthetic participants, responding to the same set of 10 word pairs ($J_1 = 10$) and the same set of 5 singletons ($J_2 = 5$) in each of the two sessions. Second, each data set contained observations from 63 participants, responding to 20 word pairs and 10 singletons in each of the two sessions. Third, each data set contained observations from 126 participants, responding to 10 word pairs and 5 singletons in each of the two sessions. We used five ($P_1 = 5$) parameters for each participant-word pair combination: $\boldsymbol{\theta}_{ij_1} = (c_{1,ij}, r_{ij}, u_{1,ij}, c_{2,ij}, u_{2,ij})$. The cluster-retrieval r parameter was thus constrained to be equal across the two sessions, $r_{1,ij} = r_{2,ij} = r_{ij}$. We used two ($P_2 = 2$) parameters for each participant-singleton combination: $\boldsymbol{\theta}_{ij_2} = (a_{1,ij}, a_{2,ij})$.

As the same set of word pairs and singletons were used across the two sessions, the J_1 items effects relating to c, r, and u, and the J_2 item effects relating to a were assumed to be equal across the two sessions. We followed the approach described earlier to implement the constraint that $a = u$. The generating parameter values for the population-level parameters are shown in Figure 5.5.

The crossed-random effects pair-clustering model was fit to the synthetic datasets using WinBUGS. As before, we monitored samples from every 3^{rd} iteration, we discarded the first 2,000 samples of each chain as burn-in, and based inference on a total of 54,000 recorded samples.

Results

The results of the recovery study for the group-level model parameters are shown in Figure 5.5. As before, the group c_1, r, u_1, c_2, and u_2 parameters are reported on the probability scale, while the standard deviations and the correlations are reported on the probit scale. The group parameters and the participant and item effect standard deviations are approximated well using the posterior median even for the first set of simulations with relatively few participants and very few items. Again, the storage-retrieval u_1 and u_2 parameters and the corresponding standard deviations are estimated most precisely and the cluster-retrieval r parameter and the corresponding standard deviations are estimated least precisely. As the number of items or the number of participants increases, the bias, the posterior standard deviation, and the standard error of the recovered parameters decrease.

With respect to the participant effect correlations, the results are again less straightforward. The posterior median underestimates the parameter correlations, especially in the first set of simulations with relatively few participants and very few items. The posterior standard deviations are quite large, suggesting large uncertainty in the estimates. Naturally, increasing the number of participants or the number of items decreases the bias, the posterior standard deviation, and the standard error of the recovered correlations. Again, correlations involving the cluster-retrieval r parameter are the least well estimated.

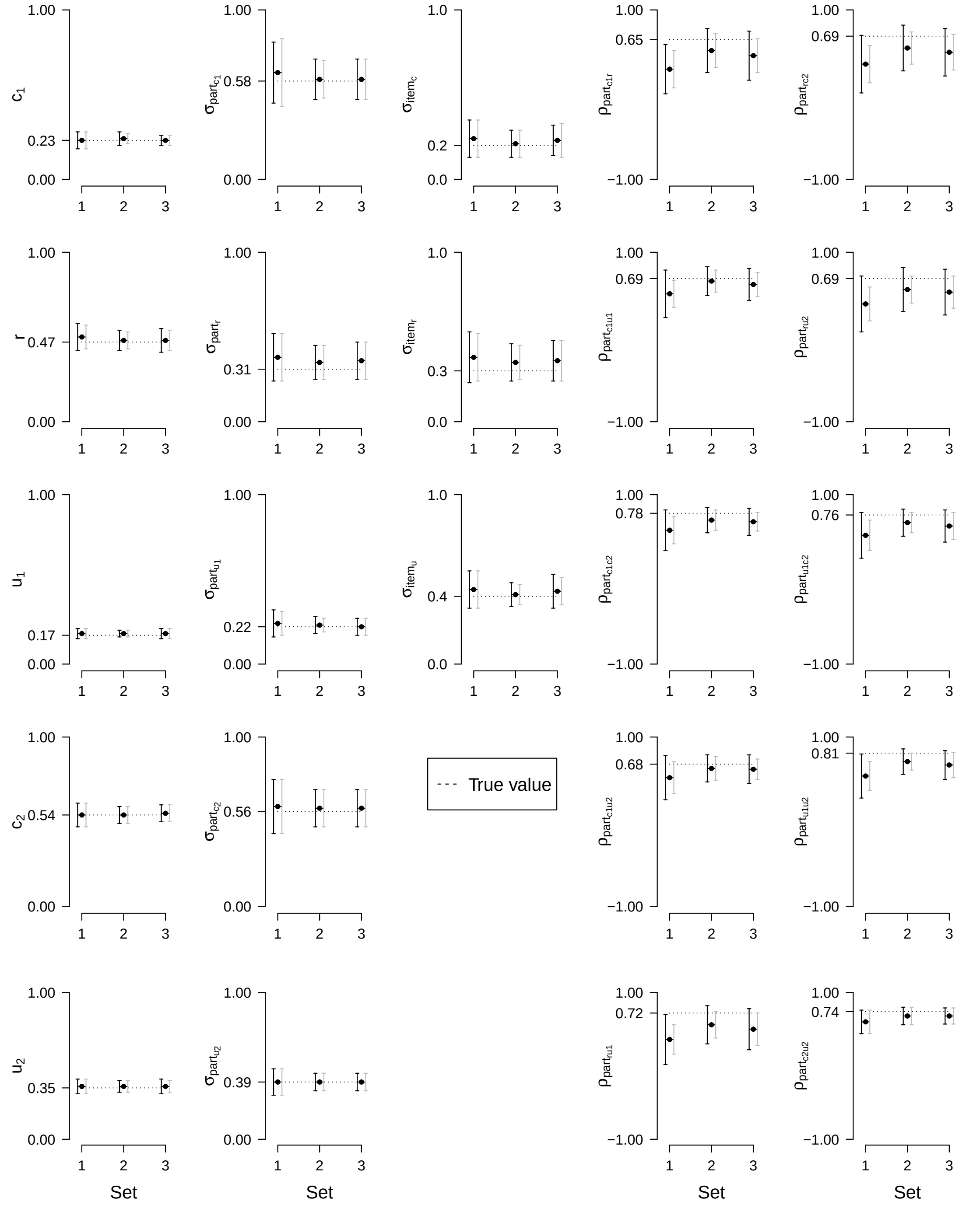


Figure 5.5 *Posterior medians from the parameter recovery study for the crossed-random effects pair-clustering model using WinBUGS.* Each set of simulations consisted of 100 datasets. The black bullets indicate the mean of the posterior median of the parameters across the 100 replications. The black vertical lines are based on the mean of the posterior standard deviation across the 100 replications. The gray vertical lines indicate standard error of the posterior median across the 100 replications.

To sum up, the results of the simulation study indicated that the WinBUGS implementation of the crossed-random effects pair-clustering model adequately recovered the true parameter values. In the next section, we apply the model to novel experimental data and illustrate the possibility to incorporate correlated participant as well as correlated item effects.

Fitting Real Data: A Pair-Clustering Experiment on Word Frequency

In order to illustrate the use of the crossed-random effects pair-clustering model and the possibility to incorporate correlated participant as well as correlated item effects, we applied the model to novel experimental data that featured orthographically related word pairs and the manipulation of word frequency. A common finding in memory research is that free recall performance is better for pure lists of high frequency (HF) words than for pure lists of low frequency (LF) words (e.g., Deese, 1960; Hall, 1954; Postman, 1970; Sumby, 1963). For mixed lists of both HF and LF words, however, the HF advantage is often eliminated (e.g., DeLosh & McDaniel, 1996; C. P. Duncan, 1974; Gregg, 1976). Models of free recall performance typically explain this pure list-mixed list word frequency paradox in terms of differences in the relative contribution of order/relational processing and item specific processing (e.g., DeLosh & McDaniel, 1996; Merritt, DeLosh, & McDaniel, 2006). The word frequency effect has never been investigated using the pair-clustering paradigm. The goal of the present experiment was therefore to demonstrate the word frequency effect in pair-clustering and to use the cross-random effects approach to explore the changes in cognitive processes that underlie the pure list-mixed list paradox. Moreover, contrary to previous applications of the pair-clustering paradigm, we employed orthographically related word pairs in order to examine orthographic clustering effects in free recall.

Methods

All 70 participants were undergraduate psychology students from the University of Amsterdam. Five participants did not comply with the instructions and the requirements of the experiment (e.g., making notes of the presented words, not being native speaker of Dutch, answering a mobile phone during the experimental session) and were excluded from all subsequent analyses. The remaining 65 participants (44 females) were native Dutch speakers, with a mean age of 22 years. Participation was rewarded either with course credits or with 7 euro.

The experimental stimulus pool consisted of 45 HF and 45 LF word pairs. The stimuli are available in the supplemental materials. The HF words had a mean occurrence of 185.03 per million and the LF words had a mean occurrence of 2.51 per million. Word length varied between 3 and 7 letters, with a mean length of 4.27 and 4.36 for HF and LF words, respectively. The word pairs were orthographically related Dutch nouns, where the two members of each word pair differed only in terms of one consonant (e.g., book-cook and house-mouse). Each word was orthographically similar only to its pair and orthographically dissimilar to all other words in the stimulus pool.

Each participant was presented with six experimental lists: two lists consisting of 10 HF word pairs and 5 HF singletons (i.e., pure HF lists), two lists consisting of 10 LF word pairs and 5 LF singletons (i.e., pure LF lists), one list consisting of 5 HF and 5 LF word pairs and 3 HF and 2 LF singletons, and one list consisting of 5 HF and 5 LF word pairs and 2 HF and 3 LF singletons (i.e., mixed lists). The study words were randomized across participants. For each participant, 30 HF and 30 LF word pairs were randomly assigned to the different experimental lists. The remaining 15 HF and 15 LF pairs were used to create singletons by randomly selecting one of the two members of each word pair. The 15 HF and 15 LF singletons were then randomly assigned to the different experimental lists. Word pairs and singletons were randomly intermixed within each list, with the constraint that the lag between the presentation of the two members of a given word pair was at least two and at most five words. The order of list presentation was randomized across participants.

Apart from the experimental stimulus items, each list contained 6 primacy buffer items at the beginning and 6 recency buffer items at the end of the list. The buffer items were orthographically dissimilar to each other and to the experimental stimuli. The pure HF lists contained only HF buffers, the pure LF lists contained only LF buffers, and the mixed lists contained six HF and six LF buffers that were randomly assigned to the 12 buffer positions. In total, each experimental list consisted of 37 words: 12 buffer items, 10 word pairs and 5 singletons.

The presentation of the six experimental lists was preceded by a practice test session. The mixed frequency practice list consisted of 10 orthographically related word pairs, 5 singletons, and 12 buffer items. Words in the practice list were orthographically dissimilar to words in the experimental lists.

Testing took place in small groups of maximum eight participants using personal computers. At the beginning of the testing session, participants read the instructions and signed the informed consent. The instructions emphasized the orthographic similarity of the words to encourage participants to cluster related word pairs. After the practice session, participants were presented with the six experimental lists. Words were presented one at a time on the computer screen at a rate of 4 sec/word. After the presentation of each list, participants were instructed to recall and type in the words without paying attention the their presentation order. After each 3 minute recall period, participants were given a 1 minute break during which they played the popular computer game Tetris.

Behavioral Results

Buffer items were excluded from all subsequent analyses. Data were collapsed per *list type* (pure vs. mixed) and *word frequency* (HF vs. LF), resulting in the following four conditions: (1) one pure HF condition consisting of 20 HF word pairs and 10 HF singletons originally presented in the two pure HF lists, (2) one pure LF condition consisting of 20 LF word pairs and 10 LF singletons originally presented in the two pure LF lists, (3) one mixed HF condition consisting of 10 HF word pairs and 5 HF singletons originally presented in the two mixed lists, and (4) one mixed LF condition consisting of 10 LF word pairs and 5 LF singletons originally presented in the two mixed lists. The data are available in the supplemental materials.

As shown in the upper left panel of Figure 5.6, the free recall data demonstrated the typical pure list-mixed list word frequency paradox. Recall performance was better for the pure HF condition than for the pure LF condition; however, in the mixed condition the HF advantage was largely eliminated. We formally assessed the *word frequency* × *list type* interaction using Bayesian null hypothesis testing (see, e.g., Masson, 2011; Raftery, 1995; Wagenmakers, 2007). Specifically, we used the Bayesian information criterion (BIC) approximation to the Bayes factor (e.g., Raftery, 1999) to compute the posterior probabilities of the null and the alternative hypotheses. We assumed that the H_0 and the H_A are equally likely a priori, i.e., $P(H_0)/P(H_A) = 1$. The resulting posterior probability of 0.89 for the alternative hypothesis, $P_{BIC}(H_A|\text{Data})$, provides positive evidence for the presence of the *word frequency* × *list type* interaction (e.g., Raftery, 1995).

Model Fitting

Each participant $i = 1, ..., 65$ was presented with each HF stimulus pair $j = 1, ..., J_{HF} = 45$ either in the HF pure or in the HF mixed condition. A given participant therefore observed a specific HF stimulus pair either as a word pair or as a singleton, and either in the pure or in the mixed condition.

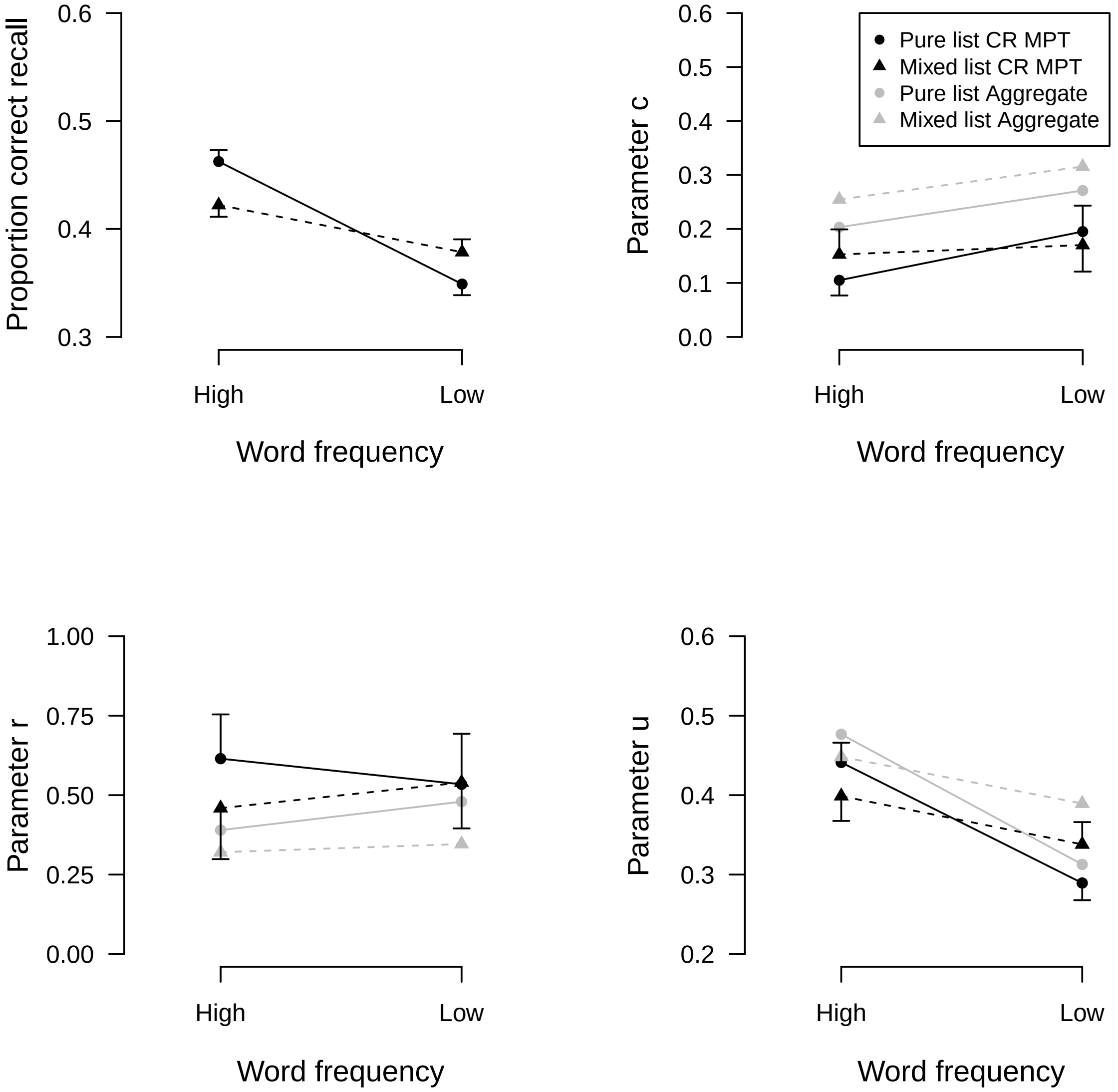


Figure 5.6 *Mean proportion of correct recall across participants and posterior medians for the group-level c, r, and u parameters for each condition of the word frequency experiment.* CR = crossed-random effects. For the recall proportions, the vertical lines show the standard errors. For the model parameters, the black circles and triangles show the posterior median of the group-level parameters from the crossed-random effects analysis of the pure and the mixed list, respectively. The black vertical lines indicate the size of the posterior standard deviation of the group-level parameters. The gray circles and triangles show parameter estimates from the aggregate analysis of the pure and the mixed list, respectively.

Similarly, each participant was presented with each LF stimulus pair $j = 1, ..., J_{LF} = 45$ either in the LF pure or the LF mixed condition. A given participant therefore observed a specific LF stimulus pair either as a word pair or as a singleton, and either in the pure or in the mixed condition. However, the additive structure of the model parameters enables us to estimate parameters for each participant-stimulus pair combination c_{ij}, r_{ij}, u_{ij}, a_{ij} for each of the four conditions.

The key group-level c, r and u parameters were free to vary across the four conditions. We imposed the following parameter constraints. Note that the constraints were chosen purely on the basis of inspection of the unconstrained parameter estimates. Formal model selection for MPT models using Bayes factors (e.g., Kass & Raftery, 1995) is beyond the scope of this article. The present analysis merely serves as an illustration of parameter estimation in the crossed-random effects pair-clustering model. First, as information on each participant and each stimulus pair was available in both category systems, we were able to place across-category system constraints on the participant as well as the item effects, resulting in $a_{ij} = u_{ij}$ for each participant-stimulus pair combination in each condition. Second, we constrained the participant effects relating to the cluster-retrieval r parameter $\delta_{part_{i_r}}$ to be equal across the four conditions. Lastly, we assumed that the item effects δ_{itemj} for c, r, and u are the same regardless whether the stimulus pair is shown in the pure condition or in the mixed condition. To illustrate the possibility to incorporate correlated participant as well as correlated item effects, we modeled both types of random effects —$\boldsymbol{\delta}_{part_i}$, and $\boldsymbol{\delta}_{itemHF_j}$ and $\boldsymbol{\delta}_{itemLF_j}$— using multivariate normal distributions, with variance-covariance matrices estimated from the data.

The crossed-random effects model was fit to the data set using WinBUGS. We monitored samples from every 3^{rd} iteration, we discarded the first 8,000 samples of each chain as burn-in, and based inference on a total of 72,000 recorded samples. Examples of thinned and un-thinned MCMC chains are available in the supplemental materials.

The posterior medians and the posterior standard deviations of the estimated group parameters c, r, and u for each condition are shown in Figure 5.6. Both the cluster-storage c and the cluster-retrieval r parameters indicate that participants indeed stored and retrieved orthographically similar words in clusters. The value of the cluster-retrieval r parameter is within the range of values typically encountered in the pair-clustering paradigm. The cluster-storage c parameter is somewhat lower than in typical applications using semantically related word pairs (e.g., Riefer et al., 2002). Nevertheless, these results indicate that, in the present experiment, orthographic relatedness fostered clustered storage and clustered retrieval.

Figure 5.6 also shows that the group parameters are estimated relatively well as indicated by the reasonable posterior standard deviations. Because the pure conditions featured twice as many items as each of the two mixed conditions, the group parameters are estimated slightly better in the HF and LF pure conditions than in the HF and LF mixed conditions. Note also that the cluster-retrieval r parameter is estimated less precisely than the cluster-storage c and storage-retrieval u parameters. This result is not surprising because the response categories involving the cluster-retrieval r parameter (i.e., C_{11}) are reached infrequently due to the relatively low value of the cluster-storage c parameter. The cluster-retrieval r parameter is therefore less well constrained by the data than the other group parameters.

To explore the effects of the experimental manipulations on the model parameters, we computed Bayesian p values for the c, r, and u parameters in the HF pure vs. LF pure and the HF mixed vs. LF mixed comparisons. Specifically, for each parameter, we computed the proportion of posterior samples where μ_{HF} is smaller (or larger) than μ_{LF} (see also Klauer, 2010). The storage-retrieval u parameter mirrors the behavioral results and demonstrates the typical word frequency paradox ($p < 0.01$ for $\mu_{u_{HFP}} < \mu_{u_{LFP}}$ and $p = 0.04$ for $\mu_{u_{HFM}} < \mu_{u_{LFM}}$). This result is to be expected because the u parameter quantifies the joint probability of the storage and retrieval of unclustered words. In contrast, the posterior medians of the c and r parameters show an entirely different pattern for the *word frequency* × *list type* interaction. With respect to the cluster-storage parameter, c is lower in the pure HF condition than in the pure LF conditions and does not differ between the mixed HF and mixed LF conditions ($p = 0.04$ for $\mu_{c_{LFP}} < \mu_{c_{HFP}}$ and $p = 0.39$ for $\mu_{c_{LFM}} < \mu_{c_{HFM}}$). Lastly, with respect to the cluster-retrieval parameter, r does not seem to differ between the pure LF and pure HF conditions, but it appears to be lower in the mixed HF condition than in the mixed LF condition ($p = 0.68$ for $\mu_{r_{LFP}} < \mu_{r_{HFP}}$ and $p = 0.36$ for $\mu_{r_{LFM}} < \mu_{r_{HFM}}$). Note, however, that the Bayesian p value for the HF mixed vs. LF mixed comparison is not convincing; the posterior distribution of the $\mu_{r_{HFM}}$ and $\mu_{r_{LFM}}$ parameters overlap considerably as a result of the larger posterior uncertainty in estimating the r parameter (see bottom left panel in Figure 5.6).

We also assessed the effects of the experimental manipulations on the model parameters without taking into account the uncertainty of the parameter estimates. For each parameter, we computed the $P_{BIC}(H_A|\text{Data})$ for the *word frequency* × *list type* interactions shown in Figure 5.6 using the posterior median of the participant parameters (i.e., $\mu + \delta_{part_i}$). For all three parameters c, r, and u, we obtained $P_{BIC}(H_A|\text{Data}) > 0.99$, a result that provides very strong evidence for the presence of the *word frequency* × *list type* interaction.

The model-based analysis uncovered a number of interesting phenomena that were not apparent in the behavioral results. First, in the pure condition, participants are slightly more likely to cluster LF than HF word pairs, suggesting that orthographic similarity is more readily apparent for LF words than for HF words. Alternatively, participants may strategically compensate for the difficulty of encoding LF words in the pure condition by paying more attention to their orthographic similarity. Second, in the mixed condition, participants are more likely to recall clustered LF word pairs than clustered HF word pairs. This result suggests that once intra-word associations are created, LF word pairs in the mixed condition are easier to recall, possibly as a result of their distinctiveness in a mixed list environment.

For comparison, we aggregated the word frequency data across participants and items and computed maximum likelihood parameter estimates using the closed form expressions presented in Batchelder and Riefer (1986). The aggregate results are presented in Figure 5.6 using the solid and dashed gray lines. Similar to the crossed-random effects analysis, the u parameter from the aggregate analysis mirrored the word frequency paradox apparent in the behavioral data. In contrast, the c and r parameters from the aggregate analysis did not reproduce the pattern of the *word frequency* × *list type* interaction from the crossed-random effects analysis.

The posterior distributions of the participant and item standard deviations are shown in Figure 5.7. The standard deviations are estimated most precisely for the participant and item effects involving the storage-retrieval u parameter. Standard deviations involving the cluster-retrieval r parameters are estimated with the largest posterior uncertainty due to the relatively low value of the cluster-storage c parameter across all conditions. Evidence for heterogeneity in participants is convincing for all participant standard deviations, with the exception of $\sigma_{part_{cLFmixed}}$, a parameter for which the lower bound of the 95% Bayesian credible interval approaches zero (i.e., 0.02). Heterogeneity in items is evident for all item standard deviations, with the exception of $\sigma_{item_{cHF}}$ and $\sigma_{item_{rHF}}$, with a lower bound of 0.04 and 0.01, respectively.

The posterior medians and standard deviations for the participant and item effect correlations are shown in Table 5.2. Correlations between the participant effects relating to the storage-retrieval u parameter (i.e., $u_{part_{HFP}}$, $u_{part_{HFM}}$, $u_{part_{LFP}}$, $u_{part_{LFM}}$) are estimated most precisely as indicated by the small posterior standard deviations. In contrast, correlations involving the participant effect $c_{part_{LFM}}$ are generally the least well constrained by the data. Participant effects relating to the cluster-storage c parameter are relatively strongly correlated across the different conditions,

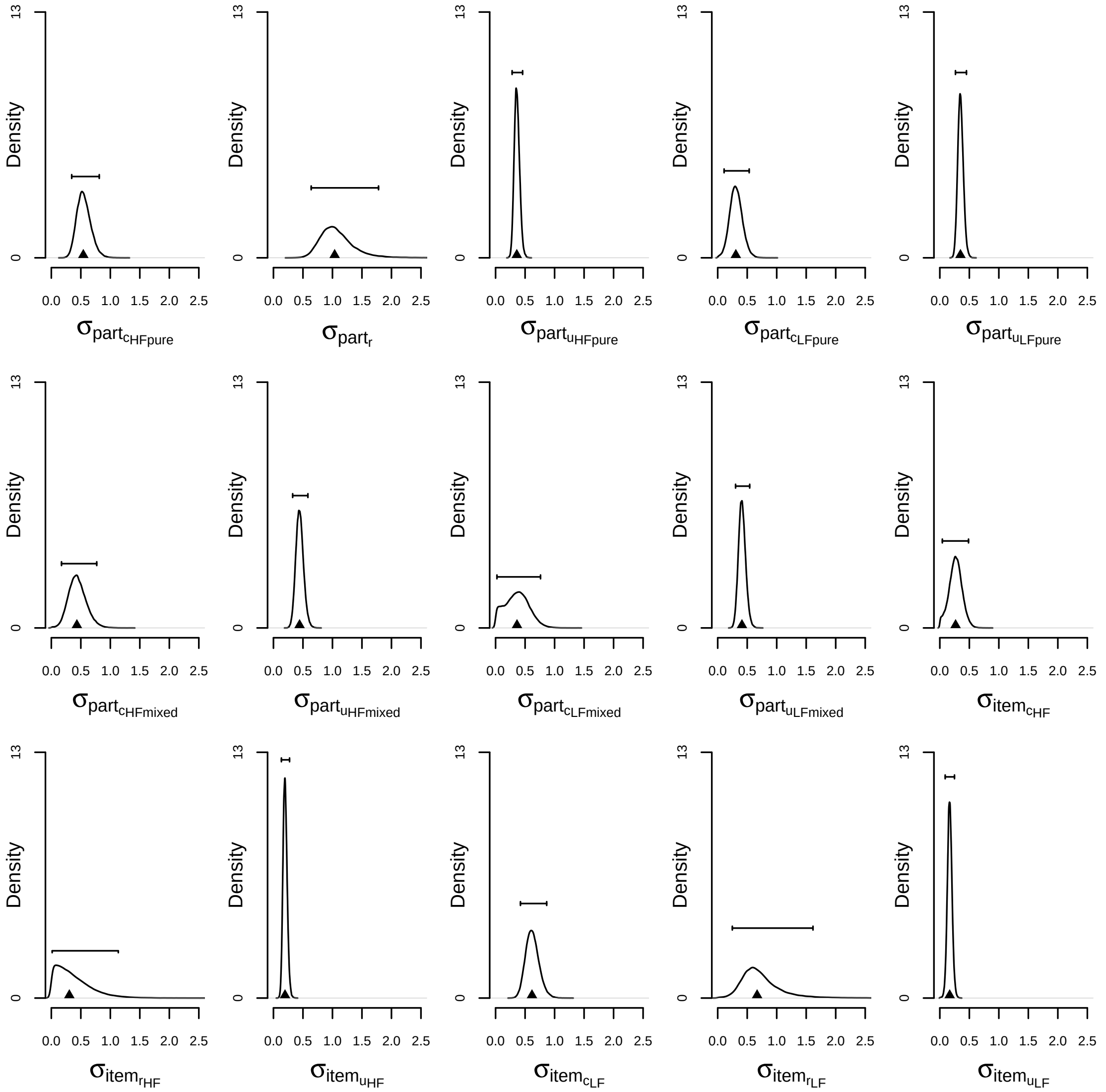


Figure 5.7 *Posterior distributions for the participant and item effect standard deviations for the word frequency experiment.* The black triangles show the median of the posterior distributions. The horizontal lines indicate the size of the 95% Bayesian credible intervals.

Table 5.2 Posterior Medians of the Correlation Parameters in the Word Frequency Experiment.

	$c_{part_{HFP}}$	$c_{part_{HFM}}$	$c_{part_{LFP}}$	$c_{part_{LFM}}$	r_{part}	$u_{part_{HFP}}$	$u_{part_{HFM}}$	$u_{part_{LFP}}$	$u_{part_{LFM}}$	$c_{item_{HF}}$	$r_{item_{HF}}$	u_{itemHF}	$c_{item_{LF}}$	$r_{item_{LF}}$	$u_{item_{LF}}$
$c_{part_{HFP}}$	1.00														
$c_{part_{HFM}}$	0.63	1.00													
	(0.19)														
$c_{part_{LFP}}$	0.55	0.58	1.00												
	(0.22)	(0.24)													
$c_{part_{LFM}}$	0.22	0.24	0.23	1.00											
	(0.33)	(0.34)	(0.32)												
r_{part}	-0.03	-0.03	0.12	0.00	1.00										
	(0.23)	(0.28)	(0.30)	(0.31)											
$u_{part_{HFP}}$	-0.56	-0.52	-0.41	0.02	0.24	1.00									
	(0.17)	(0.24)	(0.26)	(0.34)	(0.19)										
$u_{part_{HFM}}$	-0.47	-0.47	-0.30	-0.07	0.42	0.74	1.00								
	(0.20)	(0.27)	(0.30)	(0.36)	(0.18)	(0.10)									
$u_{part_{LFP}}$	-0.51	-0.47	-0.28	-0.03	0.40	0.74	0.79	1.00							
	(0.19)	(0.26)	(0.31)	(0.36)	(0.18)	(0.10)	(0.09)								
$u_{part_{LFM}}$	-0.56	-0.51	-0.32	-0.11	0.39	0.73	0.81	0.78	1.00						
	(0.18)	(0.26)	(0.29)	(0.36)	(0.19)	(0.11)	(0.09)	(0.10)							
$c_{item_{HF}}$										1.00					
$r_{item_{HF}}$										-0.22	1.00				
										(0.42)					
u_{itemHF}										0.34	-0.10	1.00			
										(0.30)	(0.41)				
$c_{item_{LF}}$													1.00		
$r_{item_{LF}}$													0.29	1.00	
													(0.32)		
$u_{item_{LF}}$													0.32	0.27	1.00
													(0.23)	(0.34)	

Note. HFP = high frequency pure condition; HFM = high frequency mixed condition; LFP = low frequency pure condition; LFM = low frequency mixed condition; $part$ = participant effect; $item$ = item effect. The standard deviation of the posterior distributions is shown in brackets.

Table 5.3 Results of the Posterior Predictive Model Checks: Aggregate and Covariance Structure Analysis.

Analysis	HF pure	HF mixed	LF pure	LF mixed
Aggregate	0.56	0.19	0.45	0.59
Participant covariances	0.61	0.40	0.27	0.51
Item covariances	0.65	0.49	0.67	0.86

Note. For the aggregate analysis, the data that are summed over items and averaged over participants. For the analysis of participant covariances, the data are summed only across the items. For the analysis of item covariances, the data are summed only across the participants. HF = high frequency; LF = low frequency.

suggesting that participants who tend to cluster orthographically related word pairs in one condition are likely to cluster also in the other conditions. Similarly, participant effects relating to the storage-retrieval u parameter are highly correlated across the different conditions, indicating that participants who are good at recalling unclustered words in one condition are also expected to perform well in the other conditions. The participant effects $c_{part_{HFP}}$, $c_{part_{HFM}}$ and $c_{part_{LFP}}$ show relatively strong negative correlations with the storage-retrieval u parameter across all conditions. The $c_{part_{LFM}}$ effect, however, seems to be uncorrelated with u. Participant effects relating to the cluster-storage c parameter are uncorrelated with participant effects for cluster-retrieval r. In contrast, participant effects relating to the storage-retrieval u parameter seem to correlate positively with r.

For HF items, the $c_{item_{HF}}$ effect is negatively correlated with the cluster-retrieval r parameter and is positively correlated with the storage-retrieval u parameter. The item effects $r_{item_{HF}}$ and $u_{item_{HF}}$ seem to be uncorrelated. For LF items, the items effects relating to the three parameters (i.e., $c_{item_{LF}}$, $r_{item_{LF}}$, and $u_{item_{LF}}$) are positively correlated. Note, however, that the correlations between the item effects —especially for HF items— are estimated rather imprecisely, as evidenced by the large posterior standard deviation of the estimates.

Assessing model fit

We used posterior predictive model checks (e.g., Gelman & Hill, 2007; Gelman et al., 1996) to examine whether the WinBUGS implementation of the crossed-random effects pair-clustering model with the chosen parameter constraints adequately describes the observed data. In posterior predictive model checks, we assess the adequacy of the model by generating new data (i.e., predictions) using samples from the joint posterior distribution of the estimated parameters. If our implementation of the crossed-random effects pair-clustering model adequately describes the modeled data, the predictions based on the model parameters should closely approximate the observed data.

We formalized the model checks with posterior predictive p values (e.g., Gelman & Hill, 2007; Gelman et al., 1996; Klauer, 2010). We first defined a test statistic T and for each of $d = 1, ..., 1, 200$ draws from the posterior distribution of the parameters, we computed its value for the observed data using the participant-item parameters, $T(data, \theta_{ij}^{d})$. We then generated new pair-clustering data for each draw d from the joint posterior and computed the value of T for each predicted data set, $T(data^{*,d}, \theta_{ij}^{d})$. The posterior predictive p value is given by the fraction of times that $T(data^{*,d}, \theta_{ij}^{d})$ is larger than $T(data, \theta_{ij}^{d})$. Extreme p values close to 0 (e.g., lower than 0.05) indicate that the model does not describe the observed data adequately.

Table 5.4 Results of the Posterior Predictive Model Checks: Participant and Item-Wise Analysis.

Analysis	HF pure	HF mixed	LF pure	LF mixed
Participant-wise	3%	2%	3%	0%
Item-wise	7%	2%	1%	4%

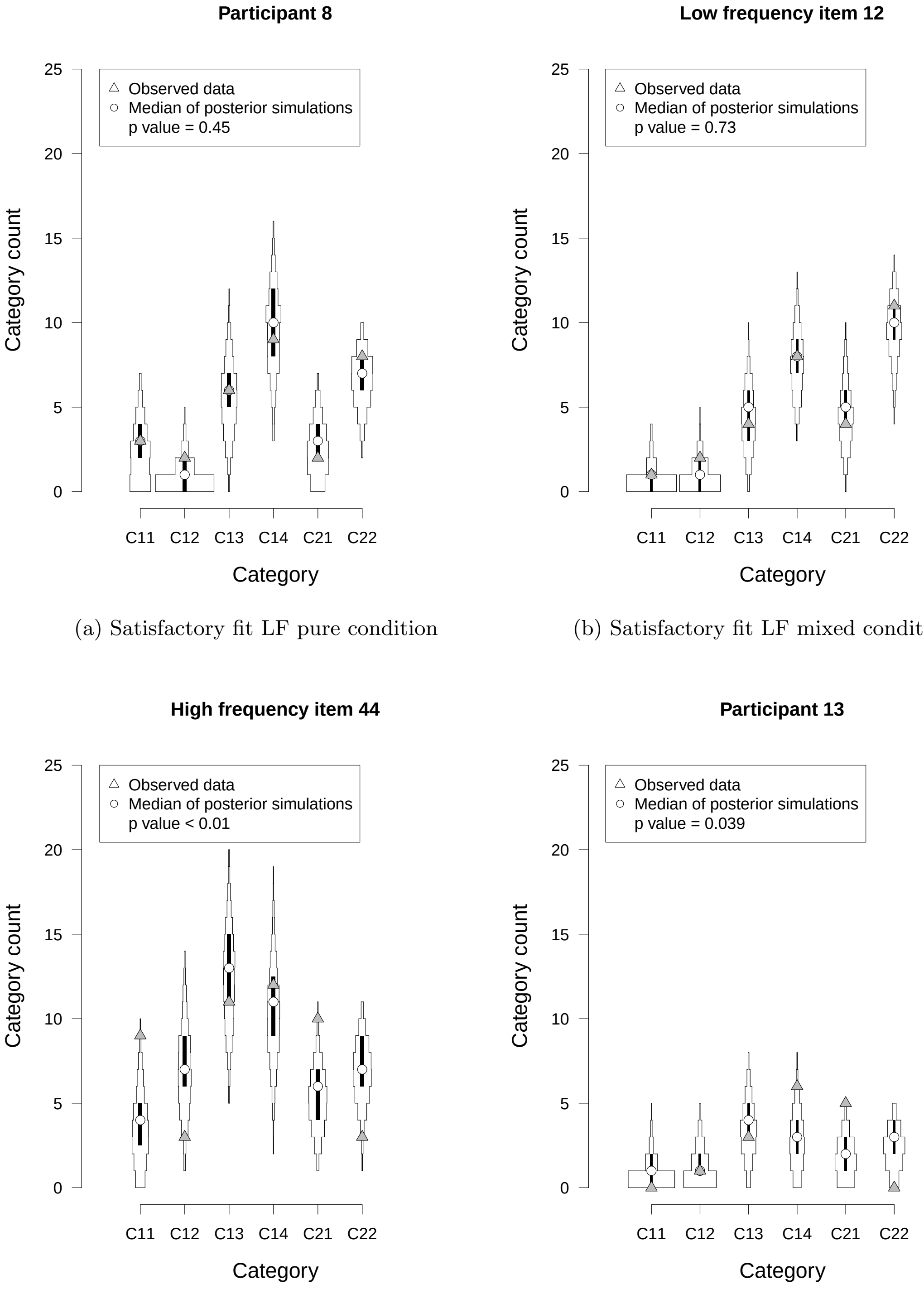
Note. HF = high frequency; LF = low frequency.

For each condition of the experiment, we conducted three sets of posterior predictive checks using Klauer's (2010) test statistics $T_1(data, \theta)$ and $T_2(data, \theta)$, which Klauer proposed to assess the recovery of the mean and the covariance of the observed category frequencies, respectively. First, we examined the recovery of the observed data that are summed over items and averaged over participants using T_1. Second, we examined the recovery of the covariance structure of the observed data that (1) are summed only across the items and (2) are summed only across the participants using T_2. Lastly, we examined the recovery of the participant-wise and item-wise frequency counts using T_1.

Table 5.3 shows the posterior predictive p values for the recovery of the aggregated category frequencies and the participant and item covariances. Table 5.4 shows the percentage of participants and items with posterior predictive p values lower than 0.05 for the participant and item-wise analysis. The results indicate that the crossed-random effects pair-clustering model adequately describes the aggregated data and the covariance structure of the observed category frequencies. Although the model fares somewhat better in predicting the observed participant-wise category frequencies, it also provides adequate predictions for the majority of the items.

Figure 5.8 shows examples of model fit for the participant and item-wise posterior predictive model checks. Each panel depicts a discrete violin plot (e.g., Hintze & Nelson, 1998) for each response category in each category system. Discrete violin plots conveniently combine information available from histograms with information about summary statistics in the form of box plots. The top panels of Figure 5.8 show examples of satisfactory model fit; the observed category frequencies (i.e., gray triangles) all fall well within the 2.5^{th} and 97.5^{th} percentiles of the posterior predictions. The bottom panels show examples of poor model fit; for most response categories, the observed category frequencies are severely over or underestimated by the posterior predictions.

In summary, our crossed-random effects pair-clustering model provided reasonable population-level parameter estimates in the word frequency experiment. Posterior predictive model checks indicated that the model resulted in participant-stimulus pair parameter estimates that adequately described the observed data. The storage-retrieval u parameter mimicked the pattern of the behavioral results and demonstrated the typical pure list-mixed list word frequency paradox. The cluster-storage c parameter showed a small clustering advantage for LF word pairs over HF word pairs in the pure condition, possibly as a result of strategy use or the enhanced accessibility of orthographic information for LF words. The cluster-retrieval r parameter showed a recall advantage for clustered LF word pairs over clustered HF word pairs in the mixed condition, possibly as a result of the distinctiveness of LF words in a mixed list environment.



(a) Satisfactory fit LF pure condition

(b) Satisfactory fit LF mixed condition

(c) Poor fit HF pure condition

(d) Poor fit HF mixed condition

Figure 5.8 *Examples of satisfactory (panel a and b) and poor (panel c and d) model fit in the word frequency orthographic clustering experiment* The gray triangles indicate the observed data that are summed over the items (panel a and d) or over the participants (panel b and c). The circles indicate the median of the predicted category frequencies over the 1,200 posterior simulations. The black area in each violin plot is a box plot, with the box ranging from the 25^{th} to the 75^{th} percentile of the posterior predictive samples.

5.6 Discussion

MPT models are theoretically motivated stochastic models for the analysis of categorical data. Traditionally, statistical analysis for MPT models is carried out on aggregated data, assuming homogeneity in participants and items. If this assumption is violated, the analysis of aggregated data may lead to erroneous conclusions. Fortunately, various methods are now available to incorporate heterogeneity either in participants or in items within MPT models.

Here we focused on Klauer's (2010) latent-trait approach that postulates a multivariate normal distribution to model individual differences between the probit transformed model parameters. We provided a WinBUGS implementation of the latent-trait pair-clustering model and demonstrated that it provides well calibrated parameter estimates in synthetic data. We then expanded the latent-trait pair-clustering model to incorporate item variability. The resulting crossed-random effects approach assumes that participant and item effects combine additively on the probit scale. The random effects are modeled using (multivariate) normal distributions. First, we used simulations to show that the WinBUGS implementation of the crossed-random effects approach adequately recovers the true parameter values. The group parameters and their standard deviations were recovered with little bias even in datasets with very few items per participant. Precise estimation of the corresponding correlation parameters required a larger sample size and/or a greater number of items. Second, we applied the crossed-random effects model to novel experimental data and examined the changes in cognitive processes that underlie the pure list-mixed list word frequency paradox.

Approaches that are based on the additivity of probit transformed participant and item effects have been recently proposed in other research contexts as well (e.g., Pratte & Rouder, 2011; Rouder & Lu, 2005; Rouder et al., 2008, 2007). Here we demonstrated that this type of crossed-random effects modeling can be extended to the pair-clustering MPT model. We chose the pair-clustering model as our running example because is it one of the most extensively studied MPT models and it has been widely used to investigate memory deficits in various age groups and clinical populations (e.g., Batchelder & Riefer, 2007). It is well-known that using items with varying difficulties aids the estimation of individual differences. The crossed-random effects extension therefore makes the pair-clustering paradigm better equipped for assessing individuals with memory deficits.

Although we focused exclusively on pair-clustering, the crossed-random effects approach may be extended to many other MPT models. The issue of model identification must, however, be carefully considered. Specifically, problems may arise in models, such as the source monitoring model (Batchelder & Riefer, 1990; Schmittmann, Dolan, Raijmakers, & Batchelder, 2010), where one or more subtrees are unidentified so that a given subtree has more parameters than free response categories. In such situations, parameter constraints are required between the category systems to reduce the number of parameters and identify the model. In many applications, however, each item features in only one of the category systems of the model. As a result, we cannot use across-subtree constraints for the item effects, resulting in parameters that are not identified at the level of the individual items. In these models, we can obtain information on each item in each category system by —as in the present experiment— randomizing the items across the participants and the experimental conditions or trial types. In this way, we can place across-subtree constraints on the item effects and, due to the additive structure of the model, we can estimate parameters for each participant-item combination. Note, however, that the present approach deals only with models that are identified for each participant after collapsing across items and for each item after collapsing across the participants. In paradigms where items are restricted to certain category systems, model identification remains an issue that requires further development.

A related issue concerns the storage-retrieval u parameter. We indexed the u parameter by word pairs rather than by individual items, assuming that the two members of a word pair are homogeneous. To the best of our knowledge, all previous applications of the pair-clustering model have used this homogeneity assumption. Nevertheless, in certain situations —as with asymmetric stimuli, such as category-exemplar word pairs— the homogeneity assumption will most likely be violated. In such situations, we may want to index the u parameter by individual items rather than by word pairs. To be able to estimate a separate u parameter for each item and, at the same time, maintain model identifiability, we may split up C_{13} in two response categories and record whether the first or the second member of the word pair has been recalled. In our experience, however, the extra degree of freedom resulting from recording the order of the recall of word pairs does not offer enough benefits to offset the sparseness resulting from splitting the response categories.

Throughout the article, we used WinBUGS to sample from the posterior distribution of the model parameters. WinBUGS is a user-friendly standard MCMC software that does not require substantial knowledge of the underlying sampling algorithm. The basic WinBUGS scripts can be easily extended to multiple testing conditions with various parameter constraints or can be adopted to accommodate other MPT models. Due to its generality, however, WinBUGS is not tailored to the particular model at hand. For models with zero-centered random effects, WinBUGS might be slow to converge as a result of the high autocorrelation between successive MCMC draws. WinBUGS then requires more samples from the posterior distribution of the parameters than a tailor-made Gibbs sampler that uses block-wise sampling for groups of correlated parameters (e.g., Klauer, 2010; Rouder et al., 2007). Nevertheless, WinBUGS is a helpful tool for fitting Bayesian hierarchical MPT models in general and the pair-clustering model in particular, as long as the convergence of the MCMC chains is carefully monitored. Of course, several alternatives to WinBUGS are now available. The OpenBUGS (Lunn et al., 2009) and JAGS (Plummer, 2003) projects, for instance, provide more options for block-wise sampling than does WinBUGS, but to the best of our knowledge, the development of blocked updating is still work in progress. For yet another —recently developed— alternative, see the Stan project (Stan Development Team, 2012).

Prior Distributions

The latent-trait approach and its crossed-random effects extension rely on Bayesian parameter estimation and as such require the specification of prior distributions. As uninformative priors might lead to unrealistic and poorly calibrated estimates, we followed Klauer's (2010) work and used weakly informative hyper-priors. Our priors for the group means are more informative and the priors for the standard deviations are more diffuse than the priors used in Klauer's original formulation of the latent-trait approach. Bayesian parameter estimation is, however, not sensitive to the choice of the prior distributions as long as sufficiently informative data are available. Consider, for example, uniform prior distributions with different ranges (i.e., 0-5, 0-10, and 0-100) for the scaling factor ξ parameters of the participant and item standard deviations. Although the priors for ξ influence the shape of the priors for σ_{part} and σ_{item}, the results of additional simulations suggest that the recovered parameter estimates are not sensitive to these choices.

In our crossed-random effects approach, we modeled the synthetic data using uncorrelated item effects, whereas we modeled the experimental data using correlated item effects. The two approaches thus differed in terms of prior assumptions; the first model assumed that the item effects are independent a priori, whereas the latter model allowed them to be correlated. With sufficiently informative data, however, the data quickly overwhelm the prior. The correlations between the a priori independent random effects may therefore also be examined using the posterior of the individual item parameters. Nevertheless, in small datasets, the assumption of a priori uncorrelated item effects may induce bias in the estimated correlations between the item parameters (e.g., Rouder et al., 2007).

If the item effects are likely to be correlated, one may capture these —similar to the participant effects— using a multivariate normal distribution. Modeling the item correlations, however, increases the amount of data that is necessary to obtain stable parameter estimates. For our experimental data, we were unable to derive sufficiently precise estimates for the item correlations despite the relatively large item pool. Similarly, additional simulations indicated that precise estimates of item correlations in the pair-clustering paradigm require a rather large number of items, a requirement that is often difficult to satisfy in clinical applications. Nevertheless, explicitly modeling the item correlations, even if they cannot be estimated precisely, has the potential to correct for bias that might result from fitting a simpler, but unrealistic model.

Conclusion

Here we introduced WinBUGS implementations of the latent-trait pair-clustering model and its crossed-random effects extension. The models allow researchers to analyze pair-clustering data without relying on aggregation and the underlying unrealistic assumption of parameter homogeneity. The WinBUGS implementation can in principle be adopted to accommodate other multinomial models and therefore provides a useful contribution to the growing arsenal of analysis techniques that address the issue of parameter heterogeneity in MPT models.

Chapter 6

Model Comparison and the Principle of Parsimony

This chapter is a modified version of:
Joachim Vandekerckhove, Dora Matzke, and Eric-Jan Wagenmakers (in press).
Model comparison and the principle of parsimony.
In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology.* Oxford, UK: Oxford University Press.

6.1 Introduction

At its core, the study of psychology is concerned with the discovery of plausible explanations for human behavior. For instance, one may observe that "practice makes perfect": as people become more familiar with a task, they tend to execute it more quickly and with fewer errors. More interesting is the observation that practice tends to improve performance such that most of the benefit is accrued early on, a pattern of diminishing returns that is well described by a power law (Logan, 1988; but see Heathcote et al., 2000). This pattern occurs across so many different tasks (e.g., cigar rolling, maze solving, fact retrieval, and a variety of standard psychological tasks) that it is known as the "power law of practice". Consider, for instance, the lexical decision task, a task in which participants have to decide quickly whether a letter string is an existing word (e.g., *sunscreen*) or not (e.g., *tolphin*). When repeatedly presented with the same stimuli, participants show a power law decrease in their mean response latencies; in fact, they show a power law decrease in the entire response time distribution, that is, both the fast responses and the slow responses speed up with practice according to a power law (Logan, 1992).

The observation that practice makes perfect is trivial, but the finding that practice-induced improvement follows a general law is not. Nevertheless, the power law of practice only provides a descriptive summary of the data and does not explain the reasons why practice should result in a power law improvement in performance. In order to go beyond direct observation and statistical summary, it is necessary to bridge the divide between observed performance on the one hand and the pertinent psychological processes on the other. Such bridges are built from a coherent set of assumptions about the underlying cognitive processes—a theory. Ideally, substantive psychological theories are formalized as quantitative models (Busemeyer & Diederich, 2010; Lewandowsky & Farrell, 2010). For example, the power law of practice has been explained by instance theory (Logan, 1992, 2002). Instance theory stipulates that earlier experiences are stored in memory as individual traces or instances; upon presentation of a stimulus, these instances race to be retrieved, and the winner of the race initiates a response. Mathematical analysis shows that, as instances are added to memory, the finishing time of the winning instance decreases as a power function. Hence, instance theory provides a simple and general explanation of the power law of practice.

For all its elegance and generality, instance theory has not been the last word on the power law of practice. The main reason is that single phenomena often afford different competing explanations. For example, the effects of practice can also be accounted for by Rickard's component power laws model (Rickard, 1997), Anderson's ACT-R model (Anderson et al., 2004), Cohen et al.'s PDP model (J. D. Cohen, Dunbar, & McClelland, 1990), Ratcliff's diffusion model (Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009; Ratcliff, 1978), or Brown and Heathcote's linear ballistic accumulator model (Brown & Heathcote, 2005, 2008; Heathcote & Hayes, 2012). When various models provide competing accounts of the same data set, it can be difficult to choose between them. The process of choosing between models is called model comparison, model selection, or hypothesis testing, and it is the focus of this chapter.

A careful model comparison procedure includes both qualitative and quantitative elements. Important qualitative elements include the plausibility, parsimony, and coherence of the underlying assumptions, the consistency with known behavioral phenomena, the ability to explain rather than describe data, and the extent to which model predictions can be falsified through experiments. Here we ignore these important aspects and focus solely on the quantitative elements. The single most important quantitative element of model comparison relates to the ubiquitous tradeoff between parsimony and goodness-of-fit (Pitt & Myung, 2002). The motivating insight is that the appeal of an excellent fit to the data (i.e., high descriptive adequacy) needs to be tempered to the extent that the fit was achieved with a highly complex and powerful model (i.e., low parsimony).

The topic of quantitative model comparison is as important as it is challenging; fortunately, the topic has received—and continues to receive—considerable attention in the field of statistics, and the results of those efforts have been made accessible to psychologists through a series of recent special issues, books, and articles (e.g., Grünwald, 2007; Myung, Forster, & Browne, 2000; Pitt & Myung, 2002; Wagenmakers & Waldorp, 2006). Here we discuss several procedures for model comparison, with an emphasis on minimum description length and the Bayes factor. Both procedures entail principled and general solutions to the tradeoff between parsimony and goodness-of-fit.

The outline of this chapter is as follows. The first section describes the principle of parsimony and the unavoidable tradeoff with goodness-of-fit. The second section summarizes the research of Wagenaar and Boer (1987) who carried out an experiment to compare three competing multinomial processing tree models (MPTs; Batchelder & Riefer, 1980); this model comparison exercise is used as a running example throughout the chapter. The third section outlines different methods for model comparison and applies them to Wagenaar and Boer's MPT models. We focus on two popular information criteria, the AIC and the BIC, on the Fisher information approximation of the minimum description length principle, and on Bayes factors as obtained from importance sampling. The fourth section contains conclusions and take-home messages.

6.2 The Principle of Parsimony

Throughout history, prominent philosophers and scientists have stressed the importance of parsimony. For instance, in the Almagest—a famous 2nd-century book on astronomy—Ptolemy writes: "We consider it a good principle to explain the phenomena by the simplest hypotheses that can be established, provided this does not contradict the data in an important way." Ptolemy's principle

Occam's razor (sometimes *Ockham's*) is named after the English philosopher and Franciscan friar Father William of Occam (c.1288-c.1348), who wrote "Numquam ponenda est pluralitas sine necessitate" (plurality must never be posited without necessity), and "Frustra fit per plura quod potest fieri per pauciora" (it is futile to do with more what can be done with less). Occam's metaphorical razor symbolizes the principle of parsimony: by cutting away needless complexity, the razor leaves only theories, models, and hypotheses that are as simple as possible without being false. Throughout the centuries, many other scholars have espoused the principle of parsimony; the list predating Occam includes Aristotle, Ptolemy, and Thomas Aquinas ("it is superfluous to suppose that what can be accounted for by a few principles has been produced by many"), and the list following Occam includes Isaac Newton ("We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. Therefore, to the same natural effects we must, so far as possible, assign the same causes."), Bertrand Russell, Albert Einstein ("Everything should be made as simple as possible, but no simpler"), and many others.

In the field of statistical reasoning and inference, Occam's razor forms the foundation for the principle of minimum description length (Grünwald, 2000, 2007). In addition, Occam's razor is automatically accommodated through Bayes factor model comparisons (e.g., Jeffreys, 1961; Jefferys & Berger, 1992; MacKay, 2003). Both minimum description length and Bayes factors feature prominently in this chapter as principled methods to quantify the tradeoff between parsimony and goodness-of-fit.

Box 6.1 Occam's razor.

of parsimony is widely known as Occam's razor (see Box 6.1); the principle is intuitive as it puts a premium on elegance. In addition, most people feel naturally attracted to models and explanations that are easy to understand and communicate. Moreover, the principle also gives ground to reject propositions that are without empirical support, including extrasensory perception, alien abductions, or mysticism. In an apocryphal interaction, Napoleon Bonaparte asked Pierre-Simon Laplace why the latter's book on the universe did not mention its creator, only to receive the curt reply "I had no need of that hypothesis".

However, the principle of parsimony finds its main motivation in the benefits that it bestows those who use models for prediction. To see this, note that empirical data are composed of a structural, replicable part and an idiosyncratic, non-replicable part. The former is known as the signal, and the latter is known as the noise (Silver, 2012). Models that capture all of the signal and none of the noise provide the best possible predictions to unseen data from the same source. Overly simplistic models, however, fail to capture part of the signal; these models underfit the data and provide poor predictions. Overly complex models, on the other hand, mistake some of the noise for actual signal; these models overfit the data and again provide poor predictions. Thus, parsimony is essential because it helps discriminate the signal from the noise, allowing better prediction and generalization to new data.

Goodness-of-Fit

"From the earliest days of statistics, statisticians have begun their analysis by proposing a distribution for their observations and then, perhaps with somewhat less enthusiasm, have checked on whether this distribution is true. Thus over the years a vast number of test procedures have appeared, and the study of these procedures has come to be known as goodness-of-fit" (D'Agostino & Stephens, 1986, p. v).

The *goodness-of-fit* of a model is a quantity that expresses how well the model is able to account for a given set of observations. It addresses the following question: Under the assumption that a certain model is a true characterization of the population from which we have obtained a sample, and given the best fitting parameter estimates for that model, how well does our sample of data agree with that model?

Various ways of quantifying goodness-of-fit exist. One common expression involves a Euclidean distance metric between the data and the model's best prediction (the least squared error or LSE metric is the most well-known of these). Another measure involves the likelihood function, which expresses the likelihood of observing the data under the model, and is maximized by the best fitting parameter estimates (Myung, 2000).

Parsimony

Goodness-of-fit must be balanced against model complexity in order to avoid overfitting—that is, to avoid building models that well explain the data at hand, but fail in out-of-sample predictions. The principle of parsimony forces researchers to abandon complex models that are tweaked to the observed data in favor of simpler models that can generalize to new data sets.

A common example is that of polynomial regression. Figure 6.1 gives a typical example. The observed data are the circles in both the left and right panels. Crosses indicate unobserved, out-of-sample data points to which the model should generalize. In the left panel, a quadratic function is fit to the 8 observed data points, whereas the right panel shows a 7$^{\text{th}}$ order polynomial function fitted to the same data. Since a polynomial of degree 7 can be made to contain any 8 points in the plane, the observed data are perfectly captured by the best fitting polynomial. However, it is clear that this function generalizes poorly to the unobserved samples, and it shows undesirable behavior for larger values of x.

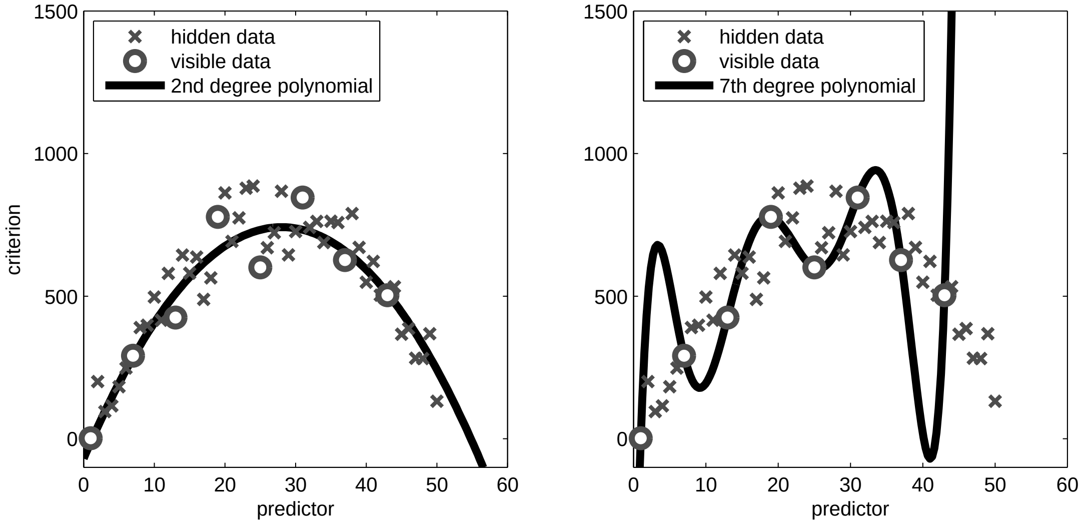


Figure 6.1 A polynomial regression of degree d is characterized by $\hat{y} = \sum_{i=0}^{d} a_i x^i$. This model has $d+1$ free parameters a_i; hence, in the right panel, a polynomial of degree 7 perfectly accounts for the 8 visible data points. This 7^{th} order polynomial, however, accounts poorly for the out-of-sample data points.

In sum, an adequate model comparison method needs to discount goodness-of-fit with model complexity. But how exactly can this be accomplished? As we will describe shortly, several model comparison methods are currently in vogue; all resulting from principled ideas on how to obtain *measures of generalizability*[1], meaning that these methods attempt to quantify the extent to which a model predicts unseen data from the same source (cf. Figure 6.1). Before outlining the details of various model comparison methods, we now introduce a data set that serves as a working example throughout the remainder of the chapter.

6.3 Example: Competing Models of Interference in Memory

For an example model comparison scenario, we revisit a study by Wagenaar and Boer (1987) on the effect of misleading information on the recollection of an earlier event. The effect of misleading postevent information was first studied systematically by E. F. Loftus, Miller, and Burns (1978); for a review of relevant literature, see Wagenaar and Boer (1987) and references therein.

Wagenaar and Boer (1987) proposed three competing theoretical accounts of the effect of misleading postevent information. To evaluate the three accounts, Wagenaar and Boer set up an experiment and introduced three quantitative models that translate each of the theoretical accounts into a set of parametric assumptions that together give rise to a probability density over the data, given the parameters.

Abstract Accounts

Wagenaar and Boer (1987) outlined three competing theoretical accounts of the effect of misleading postevent information on memory. Loftus' *destructive updating* model (DUM) posits that the conflicting information replaces and destroys the original memory. A *coexistence* model (CXM) asserts that an inhibition mechanism suppresses the original memory, which nonetheless remains viable though temporarily inaccessible. Finally, a *no-conflict* model (NCM) simply states that misleading postevent information is ignored, except when the original information was not encoded or already forgotten.

Experimental Design

The experiment by Wagenaar and Boer (1987) proceeded as follows. In Phase I, a total of 562 participants were shown a sequence of events in the form of a pictorial story involving a pedestrian-car collision. One picture in the story would show a car at an intersection, and a traffic light that was either red, yellow, or green. In Phase II, participants were asked a set of test questions with (potentially) conflicting information: Participants might be asked whether they remembered a pedestrian crossing the road when the car approached the "traffic light" (in the consistent group), the "stop sign" (in the inconsistent group) or the "intersection" (the neutral group). Then, in Phase III, participants were given a recognition test about elements of the story using picture pairs. Each pair would contain one picture from Phase I and one slightly altered version of the original picture. Participants were then asked to identify which of the pair had featured in the original story. A picture pair is shown in Figure 6.2, where the intersection is depicted with either a traffic light or a stop sign. Finally, in Phase IV, participants were informed that the correct choice in Phase III was the picture with the traffic light, and were then asked to recall the color of the traffic light.

[1] This terminology is due to Pitt and Myung (2002), who point out that measures often referred to as "model fit indices" are in fact more than mere measures of fit to the data—they combine fit to the data with parsimony and hence measure generalizability. We adopt their more accurate terminology here.

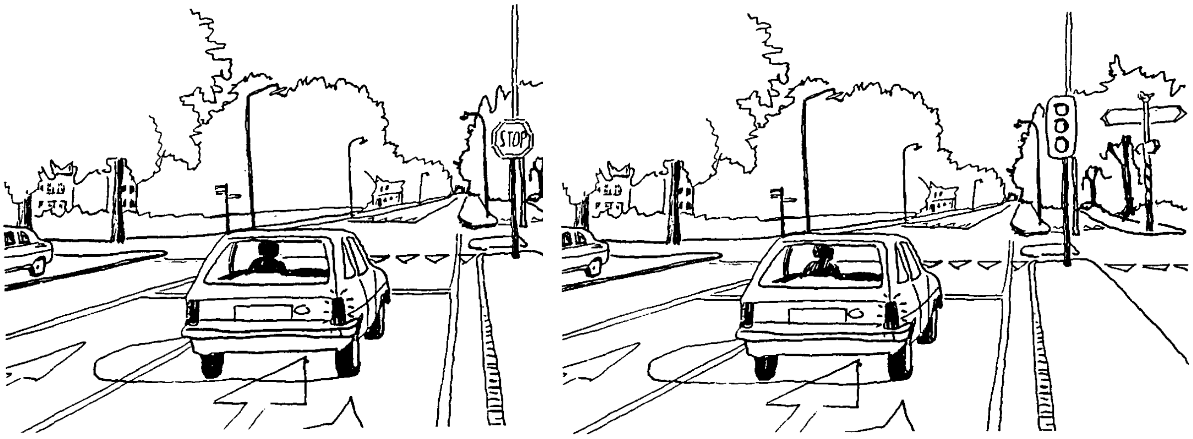


Figure 6.2 A pair of pictures from the third phase (i.e., the recognition test) of (Wagenaar & Boer, 1987, reprinted with permission), containing the critical episode at the intersection.

By design, this experiment should yield different response patterns depending on whether the conflicting postevent information destroys the original information (destructive updating model), only suppresses it temporarily (coexistence model), or does not affect the original information unless it is unavailable (no-conflict model).

Concrete Models

Wagenaar and Boer (1987) developed a series of MPT models (see Box 6.2) to quantify the predictions of the three competing theoretical accounts. Figure 6.3 depicts the no-conflict MPT model in the inconsistent condition. The figure is essentially a decision tree that is navigated from left to right. In Phase I of the collision narrative, the traffic light is encoded with probability p, and if so, the color is encoded with probability c. In Phase II, the stop sign is encoded with probability q. In Phase III, the answer may be known, or may be guessed correctly with probability 1/2, and in Phase IV the answer may be known or may be guessed correctly with probability 1/3. The probability of each path is given by the product of all the encountered probabilities, and the total probability of a response pattern is the summed probability of all branches that lead to it. For example, the total probability of getting both questions wrong is $(1-p) \times q \times 2/3 + (1-p) \times (1-q) \times 1/2 \times 2/3$. We would then, under the no-conflict model, expect that proportion of participants to fall in the response pattern with two errors.

The destructive updating model (Figure 2 in Wagenaar & Boer, 1987) extends the three-parameter no-conflict model by adding a fourth parameter d: the probability of destroying the traffic light information, which may occur whenever the stop sign was encoded. The coexistence model (Figure 3 in Wagenaar & Boer, 1987), on the other hand, posits an extra probability s that the traffic light is suppressed (but not destroyed) when the stop sign is encoded. A critical difference between the latter two is that a destruction step will lead to chance accuracy in Phase IV if every piece of information was encoded, whereas a suppression step will not affect the underlying memory and lead to accurate responding. Note here that if $s = 0$, the coexistence model reduces to the no-conflict model, as does the destructive updating model with $d = 0$. The models only make different predictions in the inconsistent condition, so that for the consistent and neutral conditions the trees are identical.

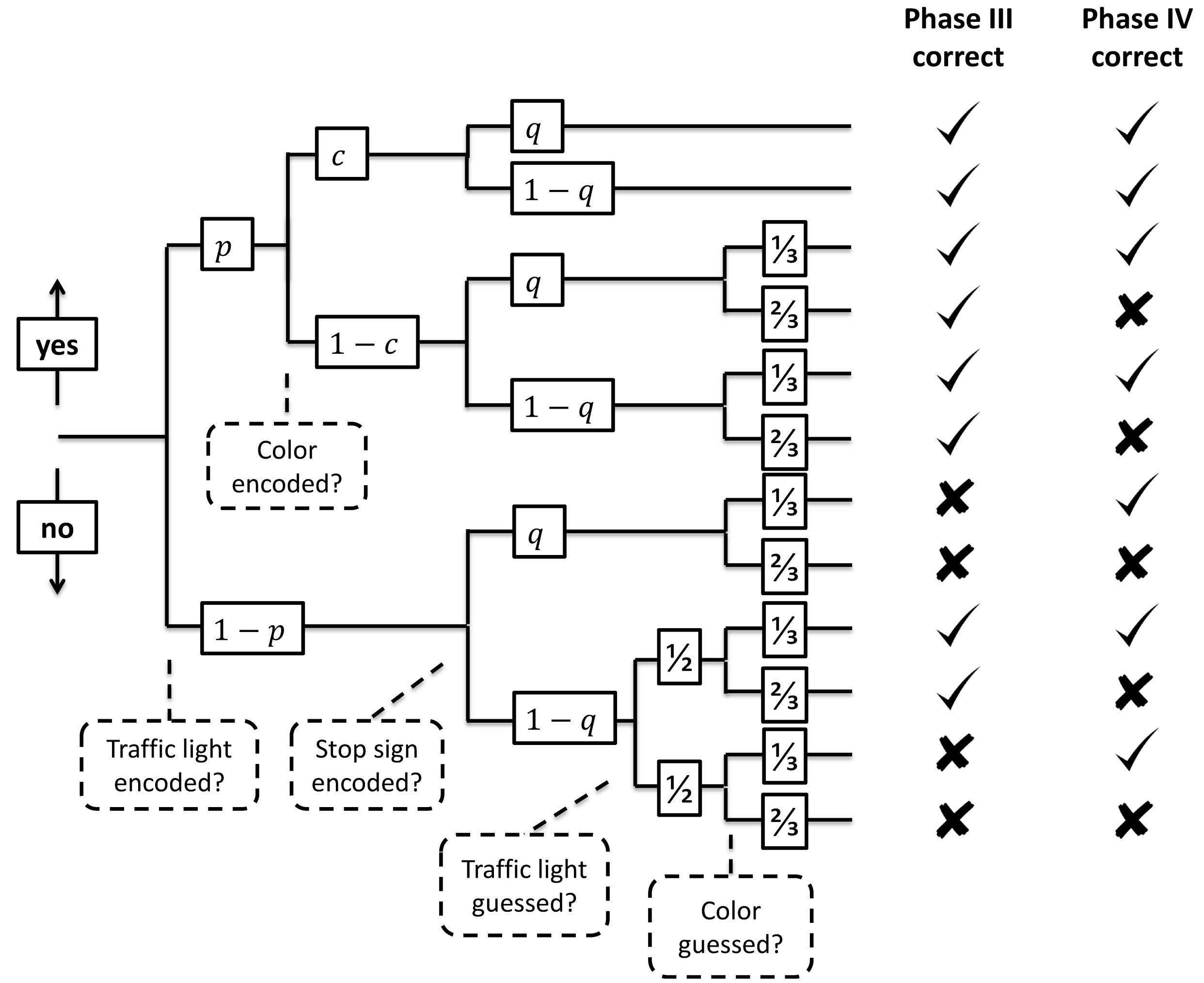


Figure 6.3 Multinomial processing tree representation of the inconsistent condition according to the no-conflict model (adapted from Wagenaar & Boer, 1987).

Previous Conclusions

After fitting the three competing MPT models, Wagenaar and Boer (1987) obtained the parameter point estimates in Table 6.1. Using a χ^2 model fit index, they concluded that “a distinction among the three model families appeared to be impossible in actual practice” (p. 304), after noting that the no-conflict model provides “an almost perfect fit” to the data. They propose, then, “to accept the most parsimonious model, which is the no-conflict model.” In the remainder of this chapter, we re-examine this conclusion using various model comparison methods.

6.4 Three Methods for Model Comparison

Many model comparison methods have been developed, all of them attempts to address the ubiquitous tradeoff between parsimony and goodness-of-fit. Here we focus on three main classes of interrelated methods: (1) AIC and BIC, the most popular information criteria; (2) minimum de-

Table 6.1 Parameter Point Estimates From Wagenaar and Boer (1987).

	p	c	q	d	s
No-conflict model (NCM)	0.50	0.57	0.50	n/a	n/a
Destructive updating model (DUM)	0.50	0.57	0.50	0.00	n/a
Coexistence model (CXM)	0.55	0.55	0.43	n/a	0.20

Multinomial processing tree models (Batchelder & Riefer, 1980; Chechile, 1973; Chechile & Meyer, 1976; Riefer & Batchelder, 1988) are psychological process models for categorical data. MPT models are used in two ways: as a psychometric tool to measure unobserved cognitive processes, and as a convenient formalization of competing psychological theories. Over time, MPTs have been applied to a wide range of psychological tasks and processes. For instance, MPT models are available for recognition, recall, source monitoring, perception, priming, reasoning, consensus analysis, the process dissociation procedure, implicit attitude measurement, and many other phenomena. For more information about MPTs, we recommend the review articles by Batchelder and Riefer (1999), Batchelder and Riefer (2007, pp. 24–32), and Erdfelder et al. (2009). The latter review article also discusses different software packages that can be used to fit MPT models. Necessarily missing from that list is the recently developed `R` package `MPTinR` (Singmann & Kellen, 2013) with which we have good experiences. As will become apparent throughout this chapter, however, our preferred method for fitting MPT models is Bayesian (Chechile & Meyer, 1976; Klauer, 2010; M. D. Lee & Wagenmakers, 2013; Matzke, Dolan, Batchelder, & Wagenmakers, in press; Rouder et al., 2008; J. B. Smith & Batchelder, 2010).

Box 6.2 Popularity of multinomial processing tree models.

scription length; (3) Bayes factors. Below we provide a brief description of each method and then apply it to the model comparison problem that confronted Wagenaar and Boer (1987).

Information Criteria

Information criteria are among the most popular methods for model comparison. Their popularity is explained by the simple and transparent manner in which they quantify the tradeoff between parsimony and goodness-of-fit. Consider for instance the oldest information criterion, AIC ("an information criterion"), proposed by Akaike (1973, 1974a):

$$\text{AIC} = -2 \ln p\left(y \mid \hat{\theta}\right) + 2k. \tag{6.1}$$

The first term $\ln p\left(y \mid \hat{\theta}\right)$ is the log maximum likelihood that quantifies goodness-of-fit, where y is the data set and $\hat{\theta}$ the maximum-likelihood parameter estimate; the second term $2k$ is a penalty for model complexity, measured by the number of adjustable model parameters k. The AIC estimates the expected information loss incurred when a probability distribution f (associated with the true data-generating process) is approximated by a probability distribution g (associated with the model under evaluation). Hence, the model with the lowest AIC is the model with the smallest expected information loss between reality f and model g, where the discrepancy is quantified by the Kullback-Leibler divergence $I(f, g)$ (for full details, see Burnham & Anderson, 2002). The AIC is unfortunately not *consistent*: as the number of observations grows infinitely large, AIC is not guaranteed to choose the true data generating model. Many researchers believe that the AIC tends to select complex models that overfit the data (O'Hagan & Forster, 2004; for a discussion see Vrieze, 2012).

Another information criterion, the BIC ("Bayesian information criterion") was proposed by G. Schwarz (1978):

$$\text{BIC} = -2 \ln p \left(y \mid \hat{\theta} \right) + k \ln n. \tag{6.2}$$

Here, the penalty term is $k \ln n$, where n is the number of observations. Hence, the BIC penalty for complexity increases with sample size, outweighing that of AIC as soon as $n \geq 8$. The BIC was derived as an approximation of a Bayesian hypothesis test using default parameter priors (the "unit information prior"; see below for more information on Bayesian hypothesis testing, and see Raftery, 1995, for more information on the BIC). The BIC is consistent: as the number of observations grows infinitely large, BIC is guaranteed to choose the true data generating model. Nevertheless, some researchers believe that in practical applications the BIC tends to select simple models that underfit the data (Burnham & Anderson, 2002).

Now consider a set of candidate models, $\mathcal{M}_i, i = 1, ..., m$, each with a specific IC (AIC or BIC) value. The model with the smallest IC value should be preferred, but the extent of this preference is not immediately apparent. For better interpretation we can calculate IC model weights (Akaike, 1974b; Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004); First, we compute, for each model i, the difference in IC with respect to the IC of the best candidate model:

$$\Delta_i = \text{IC}_i - \min \text{IC}. \tag{6.3}$$

This step is taken to increase numerical stability, but it also serves to emphasize the point that only differences in IC values are relevant. Next, we obtain the model weights by transforming back to the likelihood scale and normalizing:

$$w_i = \frac{\exp\left(-\Delta_i/2\right)}{\sum_{m=1}^{M} \exp\left(-\Delta_m/2\right)}. \tag{6.4}$$

The resulting AIC and BIC weights are called Akaike weights and Schwarz weights, respectively. These weights not only convey the relative preference among a set of candidate models, but also provide a method to combine predictions across multiple models using model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999). Both AIC and BIC rely on an assessment of model complexity that is relatively crude, as it is determined entirely by the number of free parameters but not by their functional form.

Application to Multinomial Processing Tree Models

In order to apply AIC and BIC to the three competing MPTs proposed by Wagenaar and Boer (1987), we first need to compute the maximum log likelihood. Note that the MPT model parameters determine the predicted probabilities for the different response outcome categories (cf. Figure 6.3 and Box 6.2); these predicted probabilities are deterministic parameters from a multinomial probability density function. Hence, the maximum log likelihood parameter estimates for an MPT model produce multinomial parameters that maximize the probability of the observed data (i.e., the occurrence of the various outcome categories).

Several software packages exist that can help find the maximum log likelihood parameter estimates for MPTs (e.g. Singmann & Kellen, 2013). With these estimates in hand, we can compute the information criteria described in the previous section. Table 6.2 shows the maximum log likelihood as well as AIC, BIC, and their associated weights (wAIC and wBIC; from Equation 6.4).

Table 6.2 AIC and BIC for the Wagenaar and Boer (1987) MPT Models.

	Log likelihood	k	AIC	wAIC	BIC	wBIC
No-conflict model (NCM)	-24.41	3	54.82	0.41	67.82	0.86
Destructive updating model (DUM)	-24.41	4	56.82	0.15	74.15	0.04
Coexistence model (CXM)	-23.35	4	54.70	0.44	72.03	0.10

Note. k is the number of free parameters.

Interpreting wAIC and wBIC as measures of relative preference, we see that the results in Table 6.2 are mostly inconclusive. According to wAIC, the no-conflict model and coexistence model are virtually indistinguishable, though both are preferable to the destructive updating model. According to wBIC, however, the no-conflict model should be preferred over both the destructive updating model and the coexistence model. The extent of this preference is noticeable but not decisive.

Minimum Description Length

The minimum description length principle is based on the idea that statistical inference centers around capturing regularity in data; regularity, in turn, can be exploited to compress the data. Hence, the goal is to find the model that compresses the data the most (Grünwald, 2007). This is related to the concept of Kolmogorov complexity—for a sequence of numbers, Kolmogorov complexity is the length of the shortest program that prints that sequence and then halts (Grünwald, 2007). Although Kolmogorov complexity cannot be calculated, a suite of concrete methods are available based on the idea of model selection through data compression. These methods, most of them developed by Jorma Rissanen, fall under the general heading of minimum description length (MDL; Rissanen, 1978, 1987, 1996, 2001). In psychology, the MDL principle has been applied and promoted primarily by Grünwald (2000), Grünwald (2007), Grünwald, Myung, and Pitt (2005), as well as Myung, Navarro, and Pitt (2006), Pitt and Myung (2002), and Pitt, Myung, and Zhang (2002).

Here we mention three versions of the MDL principle. First, there is the so-called *crude two-part code* (Grünwald, 2007); here, one sums the description of the model (in bits) and the description of the data encoded with the help of that model (in bits). The penalty for complex models is that they take many bits to describe, increasing the summed code length. Unfortunately, it can be difficult to define the number of bits required to describe a model.

Second, there is the Fisher information approximation (FIA; Pitt et al., 2002; Rissanen, 1996):

$$\text{FIA} = -\ln p\left(y \mid \hat{\theta}\right) + \frac{k}{2}\ln\left(\frac{n}{2\pi}\right) + \ln\int_{\Theta}\sqrt{\det\left[I(\theta)\right]}\,\mathrm{d}\theta, \tag{6.5}$$

where $I(\theta)$ is the Fisher information matrix of sample size 1. Note that FIA is similar to AIC and BIC in that it includes a first term that represents goodness-of-fit, and additional terms that represent a penalty for complexity. The second term resembles that of BIC, and the third term reflects a more sophisticated penalty that represents the number of distinguishable probability distributions that a model can generate (Pitt et al., 2002). Hence, FIA differs from AIC and BIC in that it also accounts for functional form complexity, not just complexity due to the number of free parameters. Note that FIA weights (or Rissanen weights) can be obtained by multiplying FIA by 2 and then applying Equations 6.3 and 6.4.

Table 6.3 Minimum Description Length Values for the Wagenaar and Boer (1987) MPT Models.

	Complexity	FIA	wFIA
No-conflict model (NCM)	6.44	30.86	0.44
Destructive updating model (DUM)	7.39	31.80	0.17
Coexistence model (CXM)	7.61	30.96	0.39

The third version of the MDL principle discussed here is normalized maximum likelihood (NML; Myung et al., 2006; Rissanen, 2001):

$$\text{NML} = \frac{p\left(y \mid \hat{\theta}_y\right)}{\int_x p\left(x \mid \hat{\theta}_x\right)}. \tag{6.6}$$

This equation shows that NML tempers the enthusiasm about a good fit to the observed data y (i.e., the numerator) to the extent that the model could also have provided a good fit to random data x (i.e., the denominator). NML is simple to state but can be difficult to compute; for instance, the denominator may be infinite and this requires additional measures to be taken (for details, see Grünwald, 2007).

Application to Multinomial Processing Tree Models

Using the parameter estimates from Table 6.1 and the code provided by Wu, Myung, and Batchelder (2010), we can compute the FIA for the three competing MPT models considered by Wagenaar and Boer (1987).[2] Table 6.3 displays, for each model, the FIA along with its associated complexity measure (the other one of its two constituent components, the maximum log likelihood, can be found in Table 6.2). The conclusions from the MDL analysis mirror those from the AIC measure, expressing a slight disfavor for the destructive updating model, and approximately equal preference for the no-conflict model versus the coexistence model.

Bayes Factors

In Bayesian model comparison, the posterior odds for models $\mathcal{M}_1$ and $\mathcal{M}_2$ are obtained by updating the prior odds with the diagnostic information from the data:

$$\frac{p(\mathcal{M}_1 \mid y)}{p(\mathcal{M}_2 \mid y)} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \times \frac{m(y \mid \mathcal{M}_1)}{m(y \mid \mathcal{M}_2)}. \tag{6.7}$$

Equation 6.7 shows that the change from prior odds $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ to posterior odds $p(\mathcal{M}_1 \mid y)/p(\mathcal{M}_2 \mid y)$ is given by the ratio of marginal likelihoods $m(y \mid \mathcal{M}_1)/m(y \mid \mathcal{M}_2)$, a quantity known as the *Bayes factor* (Jeffreys, 1961; Kass & Raftery, 1995). The log of the Bayes factor is often interpreted as the weight of evidence provided by the data (Good, 1985; for details, see Berger & Pericchi, 1996; Bernardo & Smith, 1994; Gill, 2002; O'Hagan, 1995).

Thus, when the Bayes factor $BF_{12} = m(y \mid \mathcal{M}_1)/m(y \mid \mathcal{M}_2)$ equals 5, the observed data y are 5 times more likely to occur under $\mathcal{M}_1$ than under $\mathcal{M}_2$; when BF_{12} equals 0.1, the observed data are 10 times more likely under $\mathcal{M}_2$ than under $\mathcal{M}_1$. Even though the Bayes factor has an

[2] Analysis using the `MPTinR` package from Singmann and Kellen (2013) gave virtually identical results.

unambiguous and continuous scale, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 6.4. We replaced the labels "not worth more than a bare mention" with "anecdotal", "decisive" with "extreme", and "substantial" with "moderate". These labels facilitate scientific communication but should be considered only as an approximate descriptive articulation of different standards of evidence.

Table 6.4 Evidence Categories for the Bayes Factor BF_{12} (Based on Jeffreys, 1961).

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for $\mathcal{M}_1$
30	—	100	Very strong evidence for $\mathcal{M}_1$
10	—	30	Strong evidence for $\mathcal{M}_1$
3	—	10	Moderate evidence for $\mathcal{M}_1$
1	—	3	Anecdotal evidence for $\mathcal{M}_1$
	1		No evidence
1/3	—	1	Anecdotal evidence for $\mathcal{M}_2$
1/10	—	1/3	Moderate evidence for $\mathcal{M}_2$
1/30	—	1/10	Strong evidence for $\mathcal{M}_2$
1/100	—	1/30	Very strong evidence for $\mathcal{M}_2$
	<	1/100	Extreme evidence for $\mathcal{M}_2$

Bayes factors negotiate the tradeoff between parsimony and goodness-of-fit and implement an automatic Occam's razor (Jefferys & Berger, 1992; MacKay, 2003; Myung & Pitt, 1997). To see this, consider that the marginal likelihood $m(y)$ can be expressed as $\int_\Theta p(y \mid \theta)p(\theta)\,\mathrm{d}\theta$: an average across the entire parameter space, with the prior providing the averaging weights. It follows that complex models with high-dimensional parameter spaces are not necessarily desirable—large regions of the parameter space may yield a very poor fit to the data, dragging down the average. The marginal likelihood will be highest for parsimonious models that use only those parts of the parameter space that are required to provide an adequate account of the data (M. D. Lee & Wagenmakers, 2013). By using marginal likelihood, the Bayes factor punishes models that hedge their bets and make vague predictions. Models can hedge their bets in different ways: by including extra parameters, by assigning very wide prior distributions to the model parameters, or by using parameters that have a complicated functional form. By computing a weighted average likelihood across the entire parameter space, the marginal likelihood (and, consequently, the Bayes factor) automatically takes all these aspects into account.

Bayes factors represent "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648) and "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378), but their application is not without challenges (Box 6.3). Below we show how these challenges can be overcome for the general class of MPT models. Next we compare the results of our Bayes factor analysis with those of the other model comparison methods using Jeffreys weights (i.e., normalized marginal likelihoods).

Application to Multinomial Processing Tree Models

In order to compute the Bayes factor, we seek to determine each model's marginal likelihood $m(y \mid \mathcal{M}_{(\cdot)})$. In the following, we omit the conditioning on a particular model. As indicated above, Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995) come with two main challenges, one practical and one conceptual. The practical challenge arises because Bayes factors are defined as the ratio of two marginal likelihoods, each of which requires integration across the entire parameter space. This integration process can be cumbersome and hence the Bayes factor can be difficult to obtain. Fortunately, there are many approximate and exact methods to facilitate the computation of the Bayes factor (e.g., Ardia, Baştürk, Hoogerheide, & van Dijk, 2012; Chen, Shao, & Ibrahim, 2002; Gamerman & Lopes, 2006); in this chapter we focus on BIC (a crude approximation), the Savage-Dickey density ratio (applies only to nested models) and importance sampling. The conceptual challenge that Bayes factors bring is that the prior on the model parameters has a pronounced and lasting influence on the result. This should not come as a surprise: the Bayes factor punishes models for needless complexity, and the complexity of a model is determined in part by the prior distributions that are assigned to the parameters. The difficulty arises because researchers are often not very confident about the prior distributions that they specify. To overcome this challenge, one can either spend more time and effort on the specification of realistic priors, or else one can choose default priors that fulfill general desiderata (e.g., Jeffreys, 1961; Liang et al., 2008). Finally, the robustness of the conclusions can be verified by conducting a sensitivity analysis in which one examines the effect of changing the prior specification (e.g., Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

Box 6.3 Two challenges for Bayes factors.

the marginal likelihood $m(y)$ is given by integrating the likelihood over the prior:

$$m(y) = \int p\left(y \mid \theta\right) p\left(\theta\right) \, \mathrm{d}\theta. \tag{6.8}$$

The most straightforward manner to obtain $m(y)$ is to draw samples from the prior $p(\theta)$ and average the corresponding values for $p(y \mid \theta)$:

$$m(y) \approx \frac{1}{N} \sum_{i=1}^{N} p\left(y \mid \theta_i\right), \qquad \theta_i \sim p(\theta). \tag{6.9}$$

For MPT models, this brute force integration approach may often be adequate. An MPT model usually has few parameters, and each is conveniently bounded from 0 to 1. However, brute force integration is inefficient, particularly when the posterior is highly peaked relative to the prior: in this case, draws from $p(\theta)$ tend to result in low likelihoods and only few chance draws may have high likelihood. This problem can be overcome by a numerical technique known as *importance sampling* (Hammersley & Handscomb, 1964).

In importance sampling, efficiency is increased by drawing samples from an importance density $g(\theta)$ instead of from the prior $p(\theta)$. Consider an importance density $g(\theta)$. Then,

$$\begin{aligned} m(y) &= \int p\left(y \mid \theta\right) p\left(\theta\right) \frac{g(\theta)}{g(\theta)} \, \mathrm{d}\theta \\ &= \int \frac{p\left(y \mid \theta\right) p\left(\theta\right)}{g(\theta)} g(\theta) \, \mathrm{d}\theta \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \frac{p\left(y \mid \theta_i\right) p\left(\theta_i\right)}{g(\theta_i)}, \qquad \theta_i \sim g(\theta). \end{aligned} \tag{6.10}$$

Note that if $g(\theta) = p(\theta)$, the importance sampler reduces to the brute force integration shown in Equation 6.9. Also note that if $g(\theta) = p(\theta \mid y)$, a single draw suffices to determine $p(y)$ exactly.

Importance sampling was invented by Stan Ulam and John von Neumann. Here we use it to estimate the marginal likelihood by repeatedly drawing samples and averaging—the samples are, however, not drawn from the prior (as per Equation 6.9, the brute force method), but instead they are drawn from some convenient density $g(\theta)$ (as per Equation 6.10; Andrieu, De Freitas, Doucet, & Jordan, 2003; Hammersley & Handscomb, 1964). The parameters in MPT models are constrained to the unit interval, and therefore the family of Beta distributions is a natural candidate for $g(\theta)$. The middle panel of Figure 6.4 shows an importance density (dashed line) for MPT parameter c in the no-conflict model for the data from Wagenaar and Boer (1987). This importance density is a Beta distribution that was fit to the posterior distribution for c using the method of moments. The importance density provides a good description of the posterior (the dashed line tracks the posterior almost perfectly) and therefore is more efficient than the brute force method illustrated in the left panel of Figure 6.4, which uses the prior as the importance density. Unfortunately, Beta distributions do not always fit MPT parameters so well; specifically, the Beta importance density may sometimes have tails that are thinner than the posterior, and this increases the variability of the marginal likelihood estimate. To increase robustness and ensure that the importance density has relatively fat tails, we can use a Beta mixture, shown in the right panel of Figure 6.4. The Beta mixture consists of a uniform prior component (i.e., the Beta$(1, 1)$ prior as in the left panel) and a Beta posterior component (i.e., a Beta distribution fit to the posterior, as in the middle panel). In this example, the mixture weight for the uniform component is $w = 0.2$. Small mixture weights retain the efficiency of the Beta posterior approach but avoid the extra variability due to thin tails. It is possible to increase efficiency further by specifying a multivariate importance density, but the present univariate approach is intuitive, easy to implement, and appears to work well in practice. The accuracy of the estimate can be confirmed by increasing the number of draws from the importance density, and by varying the w parameter.

Box 6.4 Importance sampling for MPT models using the Beta mixture method.

In sum, when the importance density equals the prior we have brute force integration, and when it equals the posterior we have a zero-variance estimator. However, knowledge of the posterior implies knowledge of its normalizing constant (i.e., the marginal likelihood), and this is exactly the quantity we wish to determine. In practice then, we want to use an importance density that is similar to the posterior, is easy to evaluate, and is easy to draw samples from. In addition, we want to use an importance density with tails that are not thinner than those of the posterior; thin tails cause the estimate to have high variance. These desiderata are met by the *Beta mixture* importance density described in Box 6.4: a mixture between a Beta$(1, 1)$ density and a Beta density that provides a close fit to the posterior distribution. Here we use a series of univariate Beta mixtures, one for each separate parameter, but acknowledge that a multivariate importance density is potentially even more efficient as it accommodates correlations between the parameters.

In our application to MPT models, we assume that all model parameters have uniform Beta$(1, 1)$ priors. For most MPT models, this assumption is fairly uncontroversial. The uniform priors can be thought of as a default choice; in the presence of strong prior knowledge one can substitute more informative priors. The uniform priors yield a default Bayes factor that can be a reference point for an analysis with more informative priors, if such an analysis is desired.

Before turning to the results of the Bayes factor model comparison, we first inspect the posterior distributions. The posterior distributions were approximated using Markov chain Monte Carlo sampling implemented in JAGS (Plummer, 2003) and WinBUGS (Lunn et al., 2012).[3] All code

[3] The second author used WinBUGS, the first and third authors used JAGS.

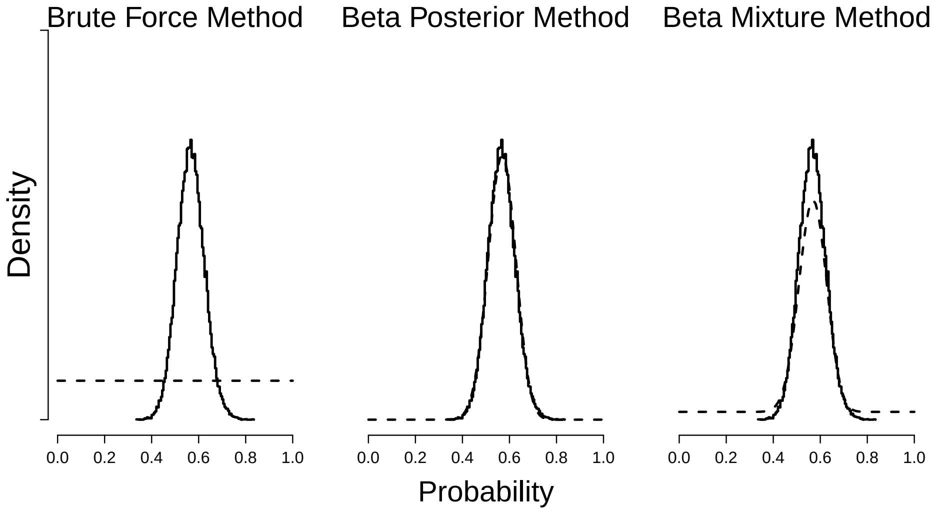


Figure 6.4 Three different importance sampling densities (dashed lines) for the posterior distribution (solid lines) of the c parameter in the no-conflict model as applied to the data from Wagenaar and Boer (1987). Left panel: a uniform Beta importance density (i.e., the brute force method); middle panel: a Beta posterior importance density (i.e., a Beta distribution that provides the best fit to the posterior); right panel: a Beta mixture importance density (i.e., a mixture of the uniform Beta density and the Beta posterior density, with a mixture weight $w = 0.2$ on the uniform component).

is available at `http://www.ejwagenmakers.com/papers.html`. Convergence was confirmed by visual inspection and the $\hat{R}$ statistic (Gelman & Rubin, 1992). The top panel of Figure 6.5 shows the posterior distributions for the no-conflict model. Although there is slightly more certainty about parameter p than there is about parameters q and c, the posterior distributions for all three parameters are relatively wide considering that they are based on data from as many as 562 participants.

The middle panel of Figure 6.5 shows the posterior distributions for the destructive-updating model. It is important to realize that when $d = 0$ (i.e., no destruction of the earlier memory), the destructive-updating model reduces to the no-conflict model. Compared to the no-conflict model, parameters p, q, and c show relatively little change. The posterior distribution for d is very wide, indicating considerable uncertainty about its true value. A frequentist point-estimate yields $\hat{d} = 0$ (Wagenaar & Boer, 1987; see also Table 6.1), but this does not convey the fact that this estimate is highly uncertain.

The lower panel of Figure 6.5 shows the posterior distributions for the coexistence model. When $s = 0$ (i.e., no suppression of the earlier memory), the coexistence model reduces to the no-conflict model. Compared to the no-conflict model and the destructive-updating model, parameters p, q, and c again show relatively little change. The posterior distribution for s is very wide, indicating considerable uncertainty about its true value.

The fact that the no-conflict model is nested under both the destructive-updating model and the no-conflict model allows us to inspect the extra parameters d and s and conclude that we have not learned very much about their true values. This suggests that, despite having tested 562 participants, the data do not firmly support one model over the other. We will now see how Bayes

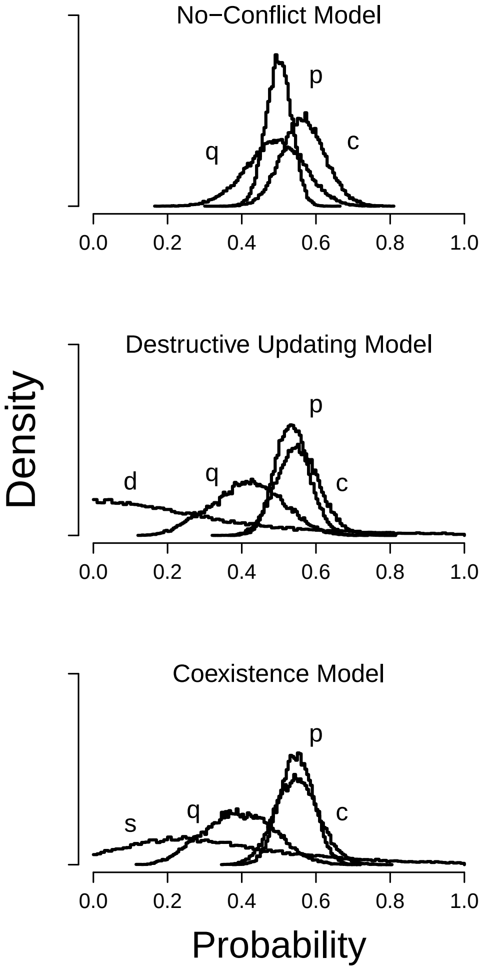


Figure 6.5 Posterior distributions for the parameters of the no-conflict MPT model, the destructive updating MPT model, and the coexistence MPT model, as applied to the data from Wagenaar and Boer (1987).

factors can make this intuitive judgment more precise.

We applied the Beta mixture importance sampling method to estimate marginal likelihoods for the three models under consideration. The results were confirmed by varying the mixture weight w, by independent implementations by the authors, and by comparison to the Savage-Dickey density ratio test presented later. Table 6.5 shows the results.

From the marginal likelihoods and the Jeffreys weights, we can derive the Bayes factors for the pair-wise comparisons; the Bayes factor is 2.77 in favor of the no-conflict model over the destructive updating model, the Bayes factor is 1.39 in favor of the coexistence model over the no-conflict model, and the Bayes factor is 3.86 in favor of the coexistence model over the destructive updating model. The first two of these Bayes factors are anecdotal or “not worth more than a bare mention” (Jeffreys, 1961), and the third one just makes the criterion for “moderate” evidence, although any enthusiasm about this level of evidence should be tempered by the realization that Jeffreys himself described a Bayes factor as high as 5.33 as “odds that would interest a gambler, but would be hardly worth more than a passing mention in a scientific paper” (Jeffreys, 1961, pp.

Table 6.5 Bayesian Evidence, Jeffreys Weights, and Pairwise Bayes Factors Computed From the Jeffreys Weights or Through the Savage-Dickey Density Ratio for the Wagenaar and Boer (1987) MPT Models.

	Bayesian	Jeffreys	Bayes factor (Savage-Dickey)		
	evidence	weight	over NCM	over DUM	over CXM
No-conflict model (NCM)	-30.55	0.36	1	2.77 (2.81)	0.72 (0.80)
Destructive updating model (DUM)	-31.57	0.13	0.36 (0.36)	1	0.26 (0.28*)
Coexistence model (CXM)	-30.22	0.51	1.39 (1.25)	3.86 (3.51*)	1

Note. * Derived through transitivity: 2.81 × 1/0.80 = 3.51.

256-257). In other words, the Bayes factors are consistent with the intuitive visual assessment of the posterior distributions: the data do not allow us to draw strong conclusions.

We should stress that Bayes factors apply to a comparison of any two models, regardless of whether or not they are structurally related or *nested*, so that one model is a special, simplified version of a larger, encompassing model. As is true for the information criteria and minimum description length methods, Bayes factors can be used to compare structurally very different models, such as for example REM (Shiffrin & Steyvers, 1997) versus ACT-R (Anderson et al., 2004), or the diffusion model (Ratcliff, 1978) versus the linear ballistic accumulator model (Brown & Heathcote, 2008). In other words, Bayes factors can be applied to nested and non-nested models alike. For the models under consideration, however, there exists a nested structure that allows one to obtain the Bayes factor through a computational shortcut.

Specifically, consider first the comparison between the no-conflict model $\mathcal{M}_{\text{NCM}}$ and the destructive updating model $\mathcal{M}_{\text{DUM}}$. As shown above, we can obtain the Bayes factor for $\mathcal{M}_{\text{NCM}}$ versus $\mathcal{M}_{\text{DUM}}$ by computing the marginal likelihoods using importance sampling. However, because the models are nested we can also obtain the Bayes factor by considering only $\mathcal{M}_{\text{DUM}}$, and dividing the posterior ordinate at $d = 0$ by the prior ordinate at $d = 0$. This surprising result was first published by Dickey and Lientz (1970), who attributed it to Leonard J. "Jimmie" Savage. The result is now generally known as the *Savage-Dickey density ratio* (e.g., Dickey, 1971; for extensions and generalizations, see Chen, 2005; Verdinelli & Wasserman, 1995; Wetzels, Grasman, & Wagenmakers, 2010; for an introduction for psychologists, see Wagenmakers et al., 2010; a short mathematical proof is presented in O'Hagan & Forster, 2004, pp. 174-177). Thus, we can exploit the fact that $\mathcal{M}_{\text{NCM}}$ is nested in $\mathcal{M}_{\text{DUM}}$ and use the Savage-Dickey density ratio to obtain the Bayes factor:

$$BF_{\text{NCM,DUM}} = \frac{m(y \mid \mathcal{M}_{\text{NCM}})}{m(y \mid \mathcal{M}_{\text{DUM}})} = \frac{p(d = 0 \mid y, \mathcal{M}_{\text{DUM}})}{p(d = 0 \mid \mathcal{M}_{\text{DUM}})}. \tag{6.11}$$

The Savage-Dickey density ratio test is visualized in Figure 6.6; the posterior ordinate at $d = 0$ is higher than the prior ordinate at $d = 0$, indicating that the data have increased the plausibility that d equals 0. This means that the data support $\mathcal{M}_{\text{NCM}}$ over $\mathcal{M}_{\text{DUM}}$. The prior ordinate equals 1, and hence $BF_{\text{NCM,DUM}}$ simply equals the posterior ordinate at $d = 0$. A nonparametric density estimator (Stone, Hansen, Kooperberg, & Truong, 1997) that respects the bound at 0 yields an estimate of 2.81. This estimate is close to 2.77, the estimate from the importance sampling approach.

The Savage-Dickey density ratio test can be applied similarly to the comparison between the no-conflict model $\mathcal{M}_{\text{NCM}}$ versus the coexistence model $\mathcal{M}_{\text{CXM}}$, where the critical test is at $s = 0$. Here the posterior ordinate is estimated to be 0.80, and hence the Bayes factor for $\mathcal{M}_{\text{CXM}}$ over

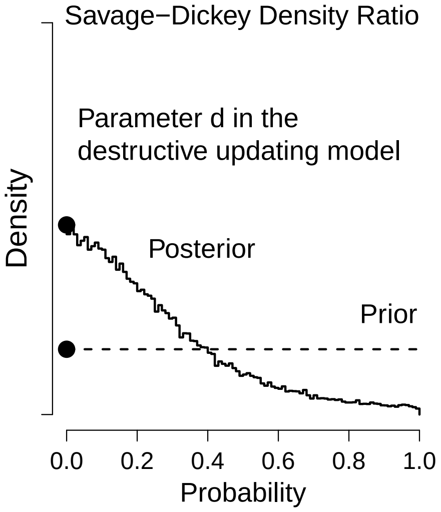


Figure 6.6 Illustration of the Savage-Dickey density ration test. The dashed and solid lines show the prior and the posterior distribution for parameter d in the destructive updating model. The black dots indicate the height of the prior and the posterior distributions at $d = 0$.

$\mathcal{M}_{\text{NCM}}$ equals $1/0.80 = 1.25$, close to the Bayes factor obtained through importance sampling, $BF_{\text{CXM,NCM}} = 1.39$.

With these two Bayes factors in hand, we can immediately derive the Bayes factor for the comparison between the destructive updating model $\mathcal{M}_{\text{DUM}}$ versus the coexistence model $\mathcal{M}_{\text{CXM}}$ through transitivity, that is, $BF_{\text{CXM,DUM}} = BF_{\text{NCM,DUM}} \times BF_{\text{CXM,NCM}}$. Alternatively, we can also obtain $BF_{\text{CXM,DUM}}$ by directly comparing the posterior density for $d = 0$ against that for $s = 0$:

$$\begin{aligned} BF_{\text{CXM,DUM}} &= BF_{\text{NCM,DUM}} \times BF_{\text{CXM,NCM}} \\ &= \frac{p(d=0 \mid y, \mathcal{M}_{\text{DUM}})}{p(d=0 \mid \mathcal{M}_{\text{DUM}})} \times \frac{p(s=0 \mid \mathcal{M}_{\text{CXM}})}{p(s=0 \mid y, \mathcal{M}_{\text{CXM}})} \\ &= \frac{p(d=0 \mid y, \mathcal{M}_{\text{DUM}})}{p(s=0 \mid y, \mathcal{M}_{\text{CXM}})}, \end{aligned} \tag{6.12}$$

where the second step is allowed because we have assigned uniform priors to both d and s, so that $p(d = 0 \mid \mathcal{M}_{\text{DUM}}) = p(s = 0 \mid \mathcal{M}_{\text{CXM}})$. Hence, the Savage-Dickey estimate for the Bayes factor between the two non-nested models $\mathcal{M}_{\text{DUM}}$ and $\mathcal{M}_{\text{CXM}}$ equals the ratio of the posterior ordinates at $d = 0$ and $s = 0$, resulting in the estimate $BF_{\text{CXM,DUM}} = 3.51$, close to the importance sampling result of 3.86.

Comparison of Model Comparisons

We have now implemented and performed a variety of model comparison methods for the three competing MPT models introduced by Wagenaar and Boer (1987): we computed and interpreted the Akaike information criteria (AIC), Bayesian information criteria (BIC), the Fisher information approximation of the minimum description length principle (FIA), and two computational implementations of the Bayes factor (BF).

The general tenor across most of the model comparison exercises has been that the data do not convincingly support one particular model. However, the destructive updating model is consistently ranked the worst of the set. Looking at the parameter estimates, it is not difficult to see why this is so: the d parameter of the destructive updating model (i.e., the probability of destroying memory through updating) is estimated at 0, thereby reducing the destructive updating model to the no-conflict model, and yielding an identical fit to the data (as can be seen in the likelihood column of Table 6.2). Since the no-conflict model counts as a special case of the destructive updating model, it is by necessity less complex from a model selection point of view—the d parameter is an unnecessary entity, the inclusion of which is not warranted by the data. This is reflected in the inferior performance of the destructive updating model according to all measures of generalizability.

The difference between the no-conflict model and the coexistence model is less clear-cut. Following AIC, the two models are virtually indistinguishable—compared to the coexistence model, the no-conflict model sacrifices one unit of log-likelihood for two units of complexity (one parameter). As a result, both models perform equally well under the AIC measure. Under the BIC measure, however, the penalty for the number of free parameters is more substantial, and here the no-conflict model trades a unit of log likelihood for $\log(N) = 6.33$ units of complexity, outdistancing both the destructive updating model and the coexistence model. The BIC is the exception in clearly preferring the no-conflict model over the coexistence model. The MDL, like the AIC, would have us hedge on the discriminability of the no-conflict model and the coexistence model.

The BF, finally, allows us to express evidence for the models using standard probability theory. Between any two models, the BF tells us how much the balance of evidence has shifted due to the data. Using two methods of computing the BF, we determined that the odds of the coexistence model over the destructive updating model almost quadrupled ($BF_{\text{CXM,DUM}} \approx 3.86$), but the odds of the coexistence model over the no-conflict model barely shifted at all ($BF_{\text{CXM,NCM}} \approx 1.39$). Finally, we can use the same principles of probability to compute Jeffreys weights, which express, for each model under consideration, the probability that it is true, assuming prior indifference. Furthermore, we can easily recompute the probabilities in case we wish to express a prior preference between the candidate models (for example, we might use the prior to express a preference for sparsity, as was originally proposed by Jeffreys, 1961).

6.5 Concluding Comments

Model comparison methods need to implement the principle of parsimony: goodness-of-fit has to be discounted to the extent that it was accomplished by a model that is overly complex. Many methods of model comparison exist (Myung et al., 2000; Wagenmakers & Waldorp, 2006), and our selective review focused on methods that are popular, easy-to-compute approximations (i.e., AIC and BIC) and methods that are difficult-to-compute "ideal" solutions (i.e., minimum description length and Bayes factors). We applied these model comparison methods to the scenario of three competing MPT models introduced by Wagenaar and Boer (1987). Despite collecting data from 562 participants, the model comparison methods indicate that the data are somewhat ambiguous; at any rate, the data do not support the destructive updating model. This echoes the conclusions drawn by Wagenaar and Boer (1987).

It is important to note that the model comparison methods discusses in this chapter can be applied regardless of whether the models are nested. This is not just a practical nicety; it also means that the methods are based on principles that transcend the details of a specific model implementation. In our opinion, a method of inference that is necessarily limited to the comparison of nested models is incomplete at best and misleading at worst. It is also important to realize that model comparison methods are *relative* indices of model adequacy; when, say, the Bayes factor expresses an extreme preference for model A over model B, this does not mean that model A fits the data at all well. Because it is a mistake to base inference on a model that fails to describe the data, a complete inference methodology features both relative and absolute indices of model adequacy. For the MPT models under consideration here, Wagenaar and Boer (1987) reported that the no-conflict model provided "an almost perfect fit" to the data.[4]

The example MPT scenario considered here was relatively straightforward. More complicated MPT models contain order-restrictions, feature individual differences embedded in a hierarchical framework (Klauer, 2010; Matzke et al., in press), or contain a mixture-model representation with different latent classes of participants (for application to other models, see Frühwirth-Schnatter, 2006; Scheibehenne, Rieskamp, & Wagenmakers, 2013). In theory, it is relatively easy to derive Bayes factors for these more complicated models. In practice, however, Bayes factors for complicated models may require the use of numerical techniques more involved than importance sampling. Nevertheless, for standard MPT models, the Beta mixture importance sampler appears to be a convenient and reliable tool to obtain Bayes factors. We hope that this methodology will facilitate the principled comparison of MPT models in future applications.

[4] We confirmed the high quality of fit in a Bayesian framework using posterior predictives (Gelman & Hill, 2007), results not reported here.

Part III

Correlations, Partial Correlations, and Mediation

Chapter 7

The Issue of Power in the Identification of “g” with Lower-Order Factors

This chapter has been published as:
Dora Matzke, Conor V. Dolan, and Dylan Molenaar (2010).
The issue of power in the identification of “g” with lower-order factors.
Intelligence, 38, 336-344.[1]

Abstract

In higher-order factor models, general intelligence (g) is often found to correlate perfectly with lower-order common factors, suggesting that g and some well-defined cognitive ability, such as working memory, may be identical. However, the results of studies that addressed the equivalence of g and lower-order factors are inconsistent. We suggest that this inconsistency may partly be attributable to the lack of statistical power to detect the distinctiveness of the two factors. The present study therefore investigated the power to reject the hypothesis that g and a lower-order factor are perfectly correlated using artificial datasets, based on realistic parameter values and on the results of selected publications. The results of the power analyses indicated that power was substantially influenced by the effect size and the number and the reliability of the indicators. The examination of published studies revealed that most case studies that reported a perfect correlation between g and a lower-order factor were underpowered, with power coefficients rarely exceeding 0.30. We conclude the paper by emphasizing the importance of considering power in the context of identifying g with lower-order factors.

7.1 Introduction

The positive intercorrelation among scores on cognitive ability tests is a well-established phenomenon, which is often explained by positing a general intelligence factor (g). The g-factor and the positive manifold of correlations may be viewed as synonymous, i.e., the positive manifold guarantees a dominant factor in principal component analyses (Basilevsky, 1983). However, we

[1] The final publication is available at `http://dx.doi.org/10.1016/j.intell.2010.02.001`.

view the g-factor as a strong hypothesis, as the positive manifold may be attributable to causes other than a general factor (Bartholomew, Deary, & Lawn, 2009; van der Maas et al., 2006).

The g-factor is supposed to reflect the operation of a process that is common to all cognitive tasks (Jensen, 1998). There is, however, considerable disagreement as to what the nature of this general process might be. For instance, some researchers argued that g is in large part a reflection of frontal lobe functions such as inhibitory and control processes (e.g., J. Duncan, Burgess, & Emslie, 1995; Embretson, 1995; Dempster, 1991), whereas others stressed the importance of the speed of information processing (e.g., Demetriou et al., 2005; Kail & Salthouse, 1994; Jensen, 1998), or the efficiency of working memory (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002).

One source of information concerning the nature of g is higher-order factor modeling, in which g features as the highest order factor (Jensen, 1998). In these models, g is often found to correlate perfectly with a lower-order common factor, suggesting that g and some well-defined broad cognitive ability (represented by the lower-order factor) may be identical. However, the results of such studies are inconsistent. First, g has been found to be perfectly correlated with a wide variety of cognitive abilities, such as fluid reasoning (Gustafsson, 1984; Undheim & Gustafsson, 1987; W. Johnson & Bouchard, 2005), nonverbal reasoning (Dunham, McIntosh, & Gridley, 2002), perceptual reasoning (W. Johnson & Bouchard, 2005), verbal/mathematical ability (Stauffer, Ree, & Carretta, 1996), and working memory (Colom, Rebollo, Palacios, Juan-Espinosa, & Kyllonen, 2004; Stauffer et al., 1996; Kyllonen & Christal, 1990; Kyllonen, 1993; Colom, Escorial, Shih, & Privado, 2007). The interpretation of such a perfect correlation varies from study to study: some discuss the theoretical implications (e.g., Gustafsson, 1984; Undheim & Gustafsson, 1987), whereas others merely note the results, but do not interpret them theoretically (e.g., W. Johnson & Bouchard, 2005). Second, several studies have failed to support the equivalence of general intelligence and the proposed cognitive abilities. For example, there is considerable evidence that g is highly correlated with both fluid reasoning and working memory, but cannot be considered identical with either of these constructs (e.g., Bickley, Keith, & Wolfle, 1995; Ackerman, Beier, & Boyle, 2002, 2005; Kane, Hambrick, & Conway, 2005).

We suggest that the inconsistency in the results of studies that have addressed the equivalence of g and lower-order factors may in part be attributable to a lack of statistical power. In an underpowered study, the probability of rejecting the hypothesis that two highly correlated (e.g., 0.8 or 0.9) factors are in fact perfectly correlated is low. In such studies, one should be reticent to attach too great a meaning to the supposedly perfect correlation between g and the lower-order common factor. Despite the theoretical importance of the issue of the exact nature of g, we are unaware of any study that has addressed the question of power in the context of identifying g with abilities represented by lower-order factors.

The goal of the present paper is therefore to study the power to correctly reject the hypothesis of perfectly correlated factors in situations where g and the lower-order factor of interest are strongly, but not perfectly, correlated. To this end, we investigated the power to detect a less than perfect correlation using exact population (summary) statistics, which we constructed on the basis of realistic parameter values and the results of selected publications that reported a perfect correlation between g and a lower-order common factor. We focused exclusively on hierarchical factor models, with g as the single highest order factor, i.e., with g at the apex of the hierarchy (see Jensen, 1998).

The outline of this article is as follows. In the first section, we introduce the concept of statistical power and present a brief overview of power calculations in the context of maximum likelihood estimation. In the second section, we describe the study in which we established the power to correctly reject the equivalence of two related factors under a variety of circumstances using realistic parameter values based on the literature. In the third section, we investigate the power of five published studies by reconstructing the original factor models using reported parameter values. We conclude the paper with a discussion.

7.2 Statistical Power and the Log-Likelihood Difference Test

The concept of statistical power plays an important role in formal statistical testing. Statistical power represents the probability of rejecting a hypothesis, given that it is false (e.g., J. Cohen, 1988, 1992; Kraemer & Thiemann, 1989). We show how power calculations are carried out within the common factor model. In this presentation, it may be useful to note that this approach to power calculation within models for covariance structures is conceptually the same as that to power calculation within a more basic statistical test like the independent samples t test. In case of a t test, power represents the probability of correctly rejecting the null hypothesis (H_0; no difference between the two samples) in favor of an alternative hypothesis (H_A; a difference between the two samples) using the t statistic. Here, we are concerned with the power to reject a parsimonious model (M_A) in favor of a more complex model (M_B) using the T_{diff} statistic (see below). We now present the details of this approach.

In covariance structure modeling, which includes higher-order factor modeling, M_B concerns the true model and M_A concerns the false model (Satorra & Saris, 1985; Saris & Satorra, 1993). Here M_A is a special case of M_B, in that M_A can be derived from M_B by the imposition of constraints on the parameters (e.g., by fixing a given parameter in M_B to zero): M_A is thus nested under M_B. In the present situation, M_B is the model which includes a less than perfect correlation between g and a first-order factor and M_A is the model in which the correlation is fixed to equal one, by the imposition of an appropriate constraint.[2] As shown in Table 7.1, the present article is thus concerned with the power to correctly reject M_A with perfectly correlated factors in favor of M_B with correlated, but distinct, factors. For any given statistical test, power is a function of the sample size, the Type I error probability (α), and the effect size, i.e., the discrepancy between the value of the parameter(s) of interest under M_B (correlation of say 0.8) and M_A (correlation of 1.00). A power of 0.80 is generally considered adequate: i.e., by consensus, a value lower than 0.80 implies an unacceptable risk of a Type II error. Power higher than 0.80 may, ceteris paribus, require unrealistically large sample size (J. Cohen, 1992).

Table 7.1 Probabilities of Correct and Incorrect Decisions in Hypothesis Testing in the Context of Covariance Structure Modeling

		Statistical decision	
		Reject M_A	Accept M_A
True state of the world	M_A is true ($r = 1$)	Type I error (α)	1-α
	M_B is true ($r < 1$)	Power (1-β)	Type II error (β)

Note. r = correlation between g and the lower-order factor of interest; M_B = model where $r < 1$; M_A = model where $r = 1$.

[2] Given the positive manifold, and its implication that the common factors are positively correlated (Krijnen, 2004), we do not consider the possibility of a perfect negative correlation.

In the setting of maximum likelihood estimation, the tenability of the parameter constraints associated with the M_A (i.e., the constraints that renders the correlation perfect) can be determined using a number of asymptotically equivalent statistical tests (Azzelini, 1996). Here we focus on the log-likelihood difference test, which can be calculated as

$$T_{diff} = T_A - T_B, \tag{7.1}$$

where T_A and T_B are the likelihood ratios of the M_A and M_B models, respectively.

To determine the power of the test, one has to consider the distribution of the test statistic T_{diff}. If the two factors of interest were truly perfectly correlated (i.e., if the M_A model was true), T_{diff} would asymptotically follow a central χ^2 distribution,

$$T_{diff} \sim \chi^2(df_{diff}, \lambda = 0), \tag{7.2}$$

where df_{diff} is the difference in the number of estimated parameters under the M_B and the M_A models and λ is the non-centrality parameter, which equals zero here.

If the correlation between the two factors of interest was truly less than one (i.e., if the M_B model was true), T_{diff} would follow the non-central χ^2 distribution (Satorra & Saris, 1985; Saris & Satorra, 1993),

$$T_{diff} \sim \chi^2(df_{diff}, \lambda > 0). \tag{7.3}$$

To put the non-central χ^2 distribution at use, one has to determine its shape. The shape of the non-central χ^2 distribution depends on df_{diff} and the non-centrality parameter λ. To obtain a numerical estimate of λ, one first assigns plausible values to the parameters of M_B and calculates the associated population covariance matrix. The parameters of M_B are chosen such that (inter alia) the correlation between g and the lower-order factor of interest is less than one (say 0.8 or 0.9). As the population covariance matrix is generated according to M_B (i.e., true model), it obtained features as the true population covariance matrix. Second, one expresses M_A by assigning plausible values to its parameters. The parameters of M_A are chosen such that the correlation between g and the lower-order factor of interest equals one. Note that both M_B and M_A must be fully specified; *all* model parameters must be assigned plausible values.[3] Third, one chooses a realistic, but arbitrary, sample size N, fits the M_B and the M_A models to the population covariance matrix and establishes the values of T_{diff} and df_{diff}. Note that the likelihood ratio associated with M_B (T_B) is zero, because M_B is the true model, i.e., the model used to obtain the population covariance matrix. The likelihood ratio associated with M_A (T_A) is greater than zero, as M_A is false and does not fit the population covariance matrix. The value of T_{diff} (Equation 7.1) equals the non-centrality parameter λ (Satorra & Saris, 1985; Saris & Satorra, 1993).

Once the value of λ is obtained, one can calculate the power of the test as follows. Choose the Type I error rate (α) and the associated critical value (C) based on the central χ^2 distribution. Next, determine the Type II error rate (β) by calculating the probability of observing a value of the test statistic T_{diff} that is smaller than the chosen critical value C, given the non-central χ^2 distribution,

$$\beta = P[\chi^2(df_{diff}, \lambda) < C]. \tag{7.4}$$

[3] See Hancock (2006) for a simplified approach to power calculation that does not require the specification of all parameter values.

The power of the test is then given by $1 - \beta$ (see Table 7.1). Note that this approach to power calculation does not rely on Monte Carlo simulations, rather it uses analytic methods to determine the probability to correctly reject M_A.

Thus the steps towards analytical power calculation are 1) choose the parameter values of M_B (i.e., the true model in which the correlation between g and the first-order factor is less than perfect) and calculate the associated population covariance matrix; 2) choose an arbitrary sample size N; 3) fit M_B to obtain T_B, which should equal zero (a useful check); 4) fit M_A model (i.e., the false model in which the correlation between g and the first-order factor equals one) to obtain T_A, which will assume a value greater than zero; 5) calculate λ and df_{diff}; 6) choose the Type I error rate (α) and calculate the associated critical value C; 7) calculate the power given, λ, N and α. As noted below, the power for other values of N can be calculated easily, i.e., does not require refitting the model.

7.3 Power Analysis

The objective of the power study was to establish the power to correctly reject the M_A of perfectly correlated factors under a variety of circumstances using realistic parameter values. The power to detect the distinctiveness of g and a lower-order factor depends on a number of aspects of the factor model, including the number and the reliability (i.e., explained variance in the factor model) of the indicators, and the effect size (i.e., the discrepancy between the correlations of g and the lower-order factor under M_B and M_A). Therefore, we systematically manipulated these features of the factor models to study their influence on power.

Design

We focused exclusively on hierarchical factor models with five first-order factors and with g as a single second-order factor. First, we specified M_B in which the correlations between g and all first-order factors were less than one. The parameter values of the models were chosen to span the range of values reported in published studies. In each model, the correlations (i.e., standardized second-order factor loadings) between g and the first-order factors were set to 0.95, 0.90, 0.85, 0.80, and 0.75, with corresponding explained variances of 0.90, 0.81, 0.72, 0.64, and 0.56.[4] To study the effects of the number and the reliability of the indicators, the M_B models were created by systematically manipulating these aspects of the models. The M_B models featured either two, three, or four indicators per first-order factor, where the reliabilities of the indicators were either relatively high (ranging from 0.55 to 0.80) or relatively low (ranging from 0.20 to 0.45). This design resulted in $3 \times 2 = 6$ different M_B models. Figure 7.1 presents an example of a hierarchical factor model used in the study. Note that each first-order factor had the same number of indicators and that the residuals of the indicators as well as the residuals of the first-order factors were uncorrelated.

Second, we specified the M_A models that implied perfect correlations between g and a given first-order factor. Each M_B model had five corresponding M_A models, where each M_A model was created by constraining the correlation between g and one of the five first-order factors to one. This design resulted in $6 \times 5 = 30$ different M_A models. Because each first-order factor correlated differently with g, this approach allowed us to investigate the influence of various effect sizes, ranging from 0.05 to 0.25.

[4] The explained variances of the first-order factors are given by the squares of the standardized second-order factor loadings. For example, the explained variance of η_1 in Figure 7.1 is given by $0.95^2 = 0.9$.

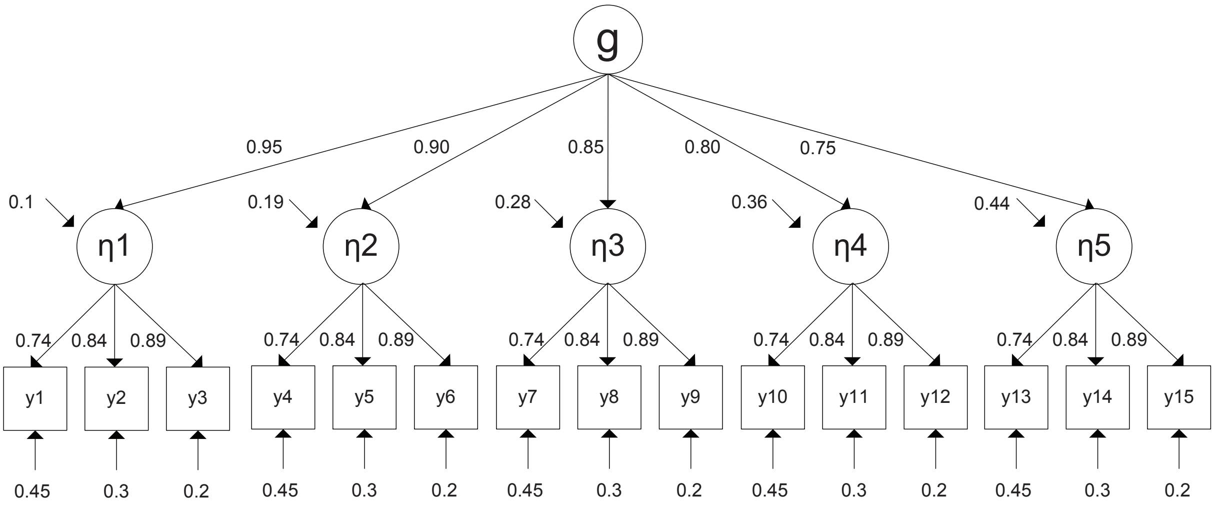


Figure 7.1 Example of a M_B model used in the study. The model features three relatively reliable indicators per first-order factor η. The values corresponding to the arrows beneath the indicators and the values corresponding to the arrows left above the first-order factors represent residual variances. Standardized parameters are shown.

Third, we obtained the non-centrality parameter λ for each M_A. To this end, we first calculated the population covariance matrices associated with the six M_B models. Next, we fitted each M_A to the data generated according to its corresponding M_B, using an arbitrary sample size of $N = 200$. To obtain perfectly correlated factors, each M_A was fitted by constraining the residual variance (i.e., unexplained variance) of the first-order factor of interest to equal zero (van der Sluis, Dolan, & Stoel, 2005). Discarding subject indices, let

$$\eta_i = \gamma_i g + \zeta_i \tag{7.5}$$

denote the regression of the i^{th} first-order factor (η) on g, with ζ_i representing the residual. Scaling the variance of g to equal one, the correlation between η_i and g equals

$$\rho_{\eta_i g} = \frac{\gamma_i}{\sqrt{\gamma_i^2 + \sigma_{\zeta_i}^2}}. \tag{7.6}$$

If $\sigma_{\zeta_i}^2$ is constrained to equal zero, then clearly the correlation equals

$$\rho_{\eta_i g} = \frac{\gamma_i}{\sqrt{\gamma_i^2}} = 1. \tag{7.7}$$

The goodness-of-fit statistics of the M_A models (T_A) were used as approximations for the non-centrality parameters.

Finally, we calculated the power to reject each M_A using the obtained non-centrality parameter and $df_{diff} = 1$. Note that because the M_A models were obtained by constraints on the variance components, the Type I error probability (α) was doubled to correct for the violation of the admissible parameter space (e.g, Dominicus, Skrondal, Gjessing, Pedersen, & Palmgren, 2006; Stoel, Garre, Dolan, & van den Wittenboer, 2006). The power was therefore computed given $\alpha = 0.05 \times 2 = 0.1$. Note also that the results reported in the following section are based on non-centrality parameters computed with $N = 200$. However, the non-centrality parameters and the power can easily be computed for any other sample size using:

$$\lambda_{new} = \left(\frac{\lambda_{original}}{200}\right) \times N_{new}, \tag{7.8}$$

where $\lambda_{original}$ is the value of the non-centrality parameter reported in the present article (based on $N = 200$) and N_{new} is the new sample size of interest. The models were fitted using LISREL (Jöreskog & Sörbom, 2001). The data generation and the power calculations were carried out using the R package (R Development Core Team, 2006). The R code for the power calculation is presented in Appendix D.1.

Results

Table 7.2 summarizes the results of the power calculations. Over the M_A models that we considered, power varied from 0.17 to 1. In general, the three manipulations (i.e., the number and the reliability of the indicators and the effect size) all influenced the power to detect the distinctiveness of g and the lower-order factors.

We first consider the effects of the reliability of the indicators. The results indicated that power increased as the reliability of the indicators increased. The increase in power was, however, more pronounced when the number of indicators and the effect sizes were relatively low. For low effect sizes, power coefficients varied by as much as 0.6 across the two reliability conditions. This difference was reduced for higher effect sizes, especially when the number of indicators was relatively high. With respect to the sample size, the results followed the same pattern. The sample size required to reach sufficient power (i.e., 0.80) was substantially lower for reliable indicators than for unreliable indicators; increases in the reliability of the indicators resulted in an average decrease of 90% in the necessary sample size.

Turning to the effects of the number of indicators, the results indicated that for reliable indicators power was high regardless of the number of indicators. For unreliable indicators, power increased as the number of indicators increased, with power coefficients varying by about 0.15 over the levels of this factor. Similarly, for reliable indicators, the necessary sample size was relatively low regardless of the number of indicators. For unreliable indicators, an increase in the number of indicators was accompanied with an average decrease of 45% in the required sample size.

Lastly, with respect to the influence of the effect size, the results indicated that for reliable indicators power was high regardless of the value of the effect size. For unreliable indicators, power increased as the effect size increased. Note, however, that the increase in power was generally more pronounced —reaching nearly 0.40— for lower effect sizes, especially when the number of indicators was high. Similarly, for reliable indicators, the necessary sample size was relatively low regardless of the value of the effect size. For unreliable indicators, the necessary sample size decreased as the effect size increased. This decrease varied between 70% and 30% over the levels of this factor.

To summarize, the results of the analyses indicated that power was substantially influenced by the effect size and the number and the reliability of the indicators. For reliable indicators (i.e., from 0.55 to 0.80), power was high regardless of the effect size and the number of indicators. For less reliable indicators (i.e., from 0.20 to 0.45), power increased as the effect size and the number of indicators increased, with a corresponding decrease in the sample size required to reach sufficient power.

Table 7.2 Power to Detect the Distinctiveness of g and the Lower-Order Factors.

		2 indicators		3 indicators		4 indicators	
Effect size		Reliable	Unreliable	Reliable	Unreliable	Reliable	Unreliable
		(0.75)	(0.38)	(0.68)	(0.32)	(0.70)	(0.33)
0.05	Power at N=200	0.77	0.17	0.87	0.21	0.96	0.31
	N at power=0.8	221	3016	160	1846	106	974
	λ	5.60	0.41	7.76	0.67	11.72	1.27
0.10	Power at N=200	1.00	0.33	1.00	0.47	1.00	0.69
	N at power=0.8	66	853	46	509	30	268
	λ	18.76	1.45	26.98	2.43	41.24	4.62
0.15	Power at N=200	1.00	0.55	1.00	0.74	1.00	0.94
	N at power=0.8	35	401	23	237	15	125
	λ	36.23	3.09	54.39	5.22	84.20	9.97
0.20	Power at N=200	1.00	0.72	1.00	0.90	1.00	0.99
	N at power=0.8	24	249	16	146	10	77
	λ	52.84	4.98	82.42	8.51	129.61	16.23
0.25	Power at N=200	1.00	0.85	1.00	0.97	1.00	1.00
	N at power=0.8	18	172	12	100	7	53
	λ	69.28	7.23	111.77	12.45	179.93	23.74

Note. The effect size reflects the value of the discrepancy between the correlations of g and the lower-order factor under the M_A (i.e., correlation of one) and the M_B. The average reliability of the indicators is shown in brackets. The reliabilities of the indicators in the M_B models were chosen as follows. In the two M_B models that featured two indicators per first-order factor, the reliabilities of the indicators were set to 0.7 and 0.8 in the reliable condition and to 0.3 and 0.45 in the unreliable condition. In the two M_B models that featured three indicators per first-order factor, the reliabilities of the indicators were set to 0.55, 0.7 and 0.8 in the reliable condition and to 0.2, 0.3 and 0.45 in the unreliable condition. In the two M_B models that featured four indicators per first-order factor, the reliabilities of the indicators were set to 0.55, 0.7, 0.75 and 0.8 in the reliable condition and to 0.2, 0.3, 0.35 and 0.45 in the unreliable condition.

7.4 Power Analysis of Selected Case Studies

In this section, we investigate the approximate power of five published studies in order to highlight the importance of power in the context of identifying g with lower-order factors. We reconstructed the original factor models using reported parameter values and determined the power to reject the equivalence of g and the proposed lower-order factors. We focused on studies that used hierarchical factor models —with g as a second or third-order factor— and reported a (near) perfect correlation between g and a lower-order factor.

Design

The selected case studies and some details of the investigated factor models are listed in Table 7.3. It is important to note that some of these studies (e.g., Gustafsson, 1984; Undheim & Gustafsson, 1987) formally tested the presence of a perfect correlation and clearly emphasized the equivalence of g and the proposed lower-order factors, whereas others (e.g., W. Johnson & Bouchard, 2005) merely reported the perfect correlation between the two factors and did not draw further conclusions about their equivalence. Nevertheless, the results of these publications provided interesting case studies to investigate the power to detect the distinctiveness of highly correlated factors.

The power analyses of the case studies were conducted as follows. First, we reconstructed the M_B models and the corresponding population covariance matrices using the original factor structures and the exact parameter values reported in the articles.[5] Although we attempted to

[5] In a few instances, our approximations of the original factor models have failed to converge. In these cases, we

approximate the original factor models as closely as possible, we did not allow for cross-factor loadings and residual correlations. Further, if the reported correlation between g and the lower-order factor of interest equaled one —either because it was fixed to one or estimated to be one— we set the standardized factor loading between the two factors to 0.95. Note also that in some models the residual variance of the lower-order factor of interest was negative (i.e., Heywood case; Heywood, 1931). Although a Haywood case raises important questions related to model misspecification and/or over-parameterization (Jöreskog & Sörbom, 1988), the issue of the adequacy of these models is beyond the scope of the present article and will not be considered further. For the purposes of the present investigation, if the reported residual variance of a lower-order factor was negative, we set the standardized residual variance to 0.10, corresponding to a standardized factor loading of 0.95. Next, we specified the M_A models of perfect correlation by constraining the residual variances of the lower-order factors of interest to equal zero. The M_A models were then fitted to the generated datasets using the original sample sizes. Lastly, we computed the power to reject the various M_A models using the obtained non-centrality parameters, $df_{diff} = 1$, and $\alpha = 0.05 \times 2 = 0.1$.

Results

The results of the case studies are shown in Table 7.3. Across the various models, power varied from 0.135 to 0.99. In most cases, however, power did not exceed 0.3, indicating that the power to reject the equivalence of g and a given lower-order factor was generally very low. In the light of the results reported above, this result was not unexpected. Most case studies featured extremely low effect sizes and factor models with only a few (two or three) relatively unreliable indicators per first-order factor. Also in line with our results, power was substantially higher for studies that used relatively reliable or a large number of indicators, reaching 0.6 for Gustafsson's (1984) model and exceeding 0.9 for the first model of W. Johnson and Bouchard (2005). In summary, the results suggested that the selected case studies, with a very few exceptions, were underpowered to detect the distinctiveness of g and the proposed lower-order factors.

7.5 General Discussion

The goal of this study was to determine the power to reject the hypothesis that g and a lower-order factor are perfectly correlated, given that the correlation is relatively high, but lower than one. First, we established the power under a variety of realistic circumstances using artificial datasets. Second, we investigated the power of five published studies by reconstructing the original factor models using reported parameter values.

The results of our power analyses revealed that power was substantially influenced by the effect size and the number and the reliability of the indicators. For highly reliable indicators, power was high regardless of the effect size and the number of indicators. For less reliable indicators, power increased as the effect size and the number of indicators increased. In the light of these results, the ideal dataset to investigate the equivalence of g and a lower-order factor would feature a relatively large number (i.e., three or four indicators per first-order factor) of reliable (i.e., from 0.55 to 0.80) indicators. Note, however, that Dolan (2000; see Jensen & Reynolds, 1982 for the summary statistics) found the mean reliability of the indicators of the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) to equal approximately 0.44 (SD = 0.15). Similarly, Dolan and Hamaker (2001; see Naglieri & Jensen, 1985 for the summary statistics) reported that

have slightly adjusted the factor structure or the parameter values of the models to assure convergence.

Table 7.3 Power and the Details of the Factor Models Used in the Case Studies

Article		Lower-order factor of interest	Mean reliability of indicators	Number of indicators	Effect size	N	λ	Power	N at power=0.80
Colom et al. (2004)	Model 1 (p.284)[a]	Working memory	0.37	12 (3)[b]	0.05	198	0.24	0.14	5101
	Model 2 (p.285)	Working memory	0.36	15 (3)	0.05	203	1.20	0.29	1046
	Model 3 (p.286)	Working memory	0.39	15 (3)	0.07	193	0.69	0.21	2010
Dunham et al. (2002)	Model 1 (p.159)	Nonverbal reasoning	0.57	6 (2)	0.05	130	0.24	0.14	3349
	Model 2 (p.159)[c]	Nonverbal reasoning	0.49	9 (2)	0.05	130	0.40	0.17	2010
	Model 3 (p.159)[c]	Memory	0.49	9 (2)	0.05	130	0.21	0.14	3828
Gustafsson (1984)	Model 1 (p.192)	Fluid reasoning	0.67	18 (2)	0.05	981	3.60	0.60	1685
W. Johnson and Bouchard (2005)	Model 1 (p.403)	Fluid reasoning	0.47	42 (7)	0.05	436	22.01	0.99	123
	Model 2 (p.408)	Perceptual reasoning	0.52	42 (5)	0.01	436	0.30	0.15	8985
Undheim and Gustafsson (1987)	Model 1 (p.155)	Fluid reasoning	0.52	26 (3)	0.05	144	0.40	0.17	2226
	Model 2 (p.157)	Fluid reasoning	0.53	10 (3)	0.05	144	0.25	0.14	3561
	Model 3 (p.161)	Fluid reasoning	0.49	28 (3)	0.05	149	0.69	0.21	1336
	Model 4 (p.163)	Fluid reasoning	0.54	13 (3)	0.05	149	2.50	0.48	369
	Model 5 (p.166)	Fluid reasoning	0.51	18 (3)	0.03	148	0.36	0.16	2542

Note. [a] The page number of the original factor model is shown in brackets. [b] The average number of indicators per first-order factor is shown in brackets. [c] The original factor model featured perfect correlations between *g* and both the nonverbal reasoning and memory factors. In the present analysis, we examined the power to reject the equivalence of *g* and the two first-order factors using two separate factor models.

the mean reliability of the indicators of the WISC-R and the Kaufman Assessment Battery for Children (K-ABC, Kaufman & Kaufman, 1983) equaled approximately 0.45 (SD = 0.18).

Our examination of published studies revealed that most of our case studies, which reported a perfect correlation between g and a lower-order factor, were underpowered, with power coefficient rarely exceeding 0.3. Consistent with our previous results, power was substantially higher for studies that used relatively reliable (i.e., Gustafsson, 1984) or a large number of indicators (i.e., W. Johnson & Bouchard, 2005). In the light of these findings, we recommend that one consider the issue of power before concluding that g and a given lower-order factor are perfectly correlated. In a study designed to address the possibly perfect relationship between g and a lower-order factor, one would ideally conduct power calculations beforehand. However, in a study in which one encounters a (possibly) unexpected perfect correlation, post-hoc power calculations can be useful.

With respect to the correlations between g and the proposed lower-order factors, there is little doubt that these may be quite large. Such correlations do certainly require an explanation. We believe that the high correlations often found between g and lower-order factors may partly result from using the same instrument to measure the common factors. Specifically, using the same instrument (e.g., WISC-R) to define the lower-order factors may introduce variance that is attributable to the particular measurement instrument (i.e., method variance; Campbell & Fiske, 1959), which in turn may increase the correlations between the measures and ultimately the correlation between g and the lower-order factors. For instance, two tests of a given construct, which employ the same method (e.g., paper and pencil), are likely to correlate higher than two tests that employ different methods (e.g., paper and pencil vs. experimental task). Also, the high correlations may partly be attributable to sampling fluctuations or may result from using heterogeneous samples to assess the equivalence of g and the proposed lower-order factors. For instance, IQ test scores are generally more highly correlated in a sample of participants with a wide age range (say 18 to 76 years of age) than in a sample with a smaller age range (18 to 36 years of age). Lastly, the reported high correlations may result from disattenuation effects associated with the unreliability of the composite scores derived from the aggregation of subtests loading on the lower-order factors (Gignac, 2007).

Nevertheless, it is possible that g may indeed be identical to a certain lower-order factor. However, given the very mixed results and the lack of power of most published studies, we are reluctant to accept this. We point out that if such an identity did truly exist, it would imply, by the application of Occam's razor, the demise of g as a causal factor in the study of individual differences in cognitive abilities: why entertain the notion of a single, essentially ill-defined higher order factor, if it is in fact identical to a well-defined lower-order factor (say, working memory)?

We also note that the focus on individual differences, which characterizes studies of g, has its inherent interpretational limitations: the presence of a perfect correlation between two variables is not sufficient to conclude that the variables are identical or that they share a common causal substrate. Suppose, for the sake of argument, that in a sample of children, who vary sufficiently in age, an appropriate statistical test reveals that the correlation between height and weight is equal to one. The perfect correlation between height and weight does obviously not imply that the two variables are identical nor that they necessarily share a common causal substrate.

In conclusion, the goal of the present study was to highlight the importance of considering power in the context of identifying g with lower-order factors. Our results provide useful guidelines on the ideal dataset and the necessary sample size required to reach sufficient power in a variety of realistic situations. Furthermore, the procedure used here to investigate power is easy to implement and the R code presented in Appendix D.1 provides a helpful tool to establish the power and the necessary sample size in the particular situation at hand. As pointed out above, the failure to do so might result in mistakenly concluding that g and a lower-order factor are perfectly correlated and therefore —at least from the perspective of individual differences— can be considered identical.

Chapter 8

Accounting for Measurement Error and the Attenuation of the Correlation: A Bayesian Approach

This chapter is currently in preparation as:
Dora Matzke, Alexander Ly, Ravi Selker, Wouter D. Weeda, Benjamin Scheibehenne, Michael D. Lee, and Eric-Jan Wagenmakers
Accounting for measurement error and the attenuation of the correlation: A Bayesian approach.

Abstract

The Pearson product-moment correlation coefficient can be severely underestimated when the observations are subject to measurement noise. Various approaches exist to correct the estimation of the correlation in the presence of measurement error, but none are routinely applied in psychological research. Here we outline a Bayesian correction method for the attenuation of correlations proposed by Behseta et al. (2009) that is conceptually straightforward and easy to apply. We illustrate the Bayesian correction with two empirical data sets; in each data set, we first estimate posterior distributions for the uncorrected and corrected correlation coefficient and then compute Bayes factors to quantify the evidence that the data provide for the presence of an association. We demonstrate that correcting for measurement error can substantially increase the correlation between noisy observations.

8.1 Introduction

The Pearson product-moment correlation coefficient is a well-known and frequently used measure to assess the linear relationship between two variables. Its popularity in psychological research is illustrated by the fact that we found that 42% of the 67 articles in the 2012 volume of the *Journal of Experimental Psychology: General* (JEP:G) report at least one Pearson correlation coefficient, with 152 correlations in total, and an average of 5.43 reported correlations per article. Also well-known, at least among statisticians, is that measurement error decreases the observed correlation between the variables (e.g., Charles, 2005; Spearman, 1904).

Although it is generally recognized that most —if not all— psychological constructs are measured only imperfectly, few researchers seem to acknowledge explicitly that the observed correlation often underestimates the true correlation between two variables. This neglect is especially puzzling because various approaches are now available to correct the correlation for the measurement error that affects the observations. Attempts to remedy the problem of the attenuation of the correlation date back to Spearman (1904), who proposed to correct the measured correlation using the reliability with which the observations were obtained. Spearman's method, however, suffers from a number of shortcomings. First, Spearman's correction formula can produce corrected correlation coefficients in excess of 1.00. Second, Spearman's method assumes homogenous error variances, an assumption that is likely to be violated in many real-life applications.

As an alternative, Behseta et al. (2009) proposed a Bayesian correction method that does not suffer from the above limitations. Contrary to Spearman's (1904) formula, the Bayesian method yields corrected correlations that are naturally bounded by -1.00 and 1.00 and is not limited to situations with homogenous error variances. Behseta et al.'s approach is conceptually straightforward and their simulations demonstrated that it is superior to Spearman's method in terms of the accuracy of the recovered corrected correlation.

Despite the availability of methods to correct the correlation coefficient for attenuation —be it Spearman's (1904) traditional method or Behseta et al.'s (2009) Bayesian approach— psychologists seldom attempt to adjust their correlations for measurement noise. Indeed, out of the 28 JEP:G articles in the 2012 volume which reported one or more correlations, only one acknowledged the deleterious effects of measurement error and corrected the observed correlation. This situation is unfortunate; as demonstrated by Behseta et al. —and as we will demonstrate again shorty— correcting the correlation for attenuation may substantially increase the association between noisy observations. Of course, correction is only possible if the magnitude of the measurement error is known. Luckily, our JEP:G literature review suggests that the uncertainty of the observations may be often estimated from the data, for example, when each observation is derived as the average of multiple trials in a repeated measures design. Specifically, we found that for 25% (38/152) of the reported correlations, the measurement error could have been estimated and corrected for. For 42% (16/38) of these correlations, correction was possible for both variables; for the remaining 58% (22/38), correction was possible for only one of the variables.

The goal of the present article is therefore to facilitate the correction of attenuated correlations with Behseta et al.'s (2009) Bayesian approach. To this end, we will first illustrate the consequences of measurement error for the computation of the correlation and present Spearman's (1904) traditional attenuation formula. We will then describe Behseta et al.'s Bayesian correction method in detail. Finally, we will illustrate the use of the Bayesian correction with two empirical data sets: one focusing on the correlation between parameters of cumulative prospect theory (Tversky & Kahneman, 1992), the other focusing on the correlation between general intelligence and the drift rate parameter of the Ratcliff diffusion model (Ratcliff, 1978; Wagenmakers, 2009).

8.2 Attenuation of the Correlation and Spearman's Correction

In this section, we first show why the presence of measurement error decreases the observed correlation between two variables. We then discuss Spearman's (1904) method for correcting the attenuation. Let θ and β be two independent random variables and let $\hat{\theta}$ and $\hat{\beta}$ be the observed, noise-contaminated measurements:

$$\hat{\theta} = \theta + \epsilon_{\theta} \tag{8.1}$$

and

$$\hat{\beta} = \beta + \epsilon_\beta, \tag{8.2}$$

where ϵ_θ and ϵ_β are the measurement errors associated with θ and β, respectively. The correlation between the *unobserved* variables θ and β is given by

$$\text{Cor}(\theta, \beta) = \frac{\text{Cov}(\theta, \beta)}{\sqrt{\text{Var}(\theta) \times \text{Var}(\beta)}}. \tag{8.3}$$

Assuming that the measurement errors are uncorrelated with θ and β and with each other, the correlation between the *observed* variables $\hat{\theta}$ and $\hat{\beta}$ is given by

$$\text{Cor}(\hat{\theta}, \hat{\beta}) = \text{Cor}(\theta + \epsilon_\theta, \beta + \epsilon_\beta) = \frac{\text{Cov}(\theta, \beta)}{\sqrt{(\text{Var}(\theta) + \text{Var}(\epsilon_\theta)) \times (\text{Var}(\beta) + \text{Var}(\epsilon_\beta))}}. \tag{8.4}$$

It immediately follows from Equation 8.3 and Equation 8.4 that the observed correlation $\text{Cor}(\hat{\theta}, \hat{\beta})$ is always lower than the unobserved true correlation $\text{Cor}(\theta, \beta)$.

To remedy the problem of attenuation, Spearman (1904) proposed to correct the observed correlation coefficient using the reliabilities of the measurements:

$$r_{\theta\beta} = \frac{r_{\hat{\theta}\hat{\beta}}}{\sqrt{r_{\hat{\theta}\hat{\theta}}} \times \sqrt{r_{\hat{\beta}\hat{\beta}}}}, \tag{8.5}$$

where $r_{\theta\beta}$ is the corrected sample correlation coefficient, $r_{\hat{\theta}\hat{\beta}}$ is the observed sample Pearson correlation coefficient, and $r_{\hat{\theta}\hat{\theta}}$ and $r_{\hat{\beta}\hat{\beta}}$ are the reliabilities of $\hat{\theta}$ and $\hat{\beta}$, respectively. The reliabilities can be computed using the sample variances, $s^2_{\hat{\theta}}$ and $s^2_{\hat{\beta}}$, and the measurement error variances, $\sigma^2_{\epsilon_\theta}$ and $\sigma^2_{\epsilon_\beta}$, as follows:

$$r_{\hat{\theta}\hat{\theta}} = \frac{s^2_{\hat{\theta}} - s^2_{\epsilon_\theta}}{s^2_{\hat{\theta}}} \tag{8.6}$$

and

$$r_{\hat{\beta}\hat{\beta}} = \frac{s^2_{\hat{\beta}} - s^2_{\epsilon_\beta}}{s^2_{\hat{\beta}}}. \tag{8.7}$$

Confidence intervals for the corrected correlation coefficient $r_{\theta\beta}$ are outlined, for example, in Charles (2005) and Winnie and Belfry (1982).

Spearman's (1904) correction for measurement error is related to errors-in-variables models. Errors-in-variables models (e.g., Buonaccorsi, 2010; Cheng & Van Ness, 1999; Fuller, 1987; for Bayesian solutions, see Congdon, 2006; Gilks et al., 1996; Gustafson, 2004; Lunn et al., 2012) are extensions of standard regression models that —similar to Spearman's method— aim at correcting the bias in parameter estimates that results from measurement error. If the criterion and the response variables are both assumed to be measured with noise, Spearman's method and the correction within standard linear regression models result in the same disattenuation (see Behseta et al., 2009, Appendix). Ratcliff and Strayer (2014) outline an alternative method to deal with the adverse consequences of measurement error using Monte Carlo simulations. Cole and Preacher (2014) discuss solutions to deal with measurement uncertainty in the context of path analysis.

Example

Consider the following example. A researcher sets out to investigate the association between excitability (θ) and depression (β); she hypothesizes that people with increased responsiveness to threatening stimuli are likely to report more depressive symptoms. The researcher measures excitability using mean response time (RT) to pictures of threatening images and measures depression using a standard depression questionnaire. The researcher then correlates the observed mean RTs ($\hat{\theta}$) and depression scores ($\hat{\beta}$) of the $i = 1, ..., 10$ participants and obtains a Pearson correlation coefficient $r_{\hat{\theta}\hat{\beta}}$ of -0.57. The mean RTs and depression scores of the 10 hypothetical participants are reported in Table 8.1.

Aware of the fact that neither excitability nor depression is measured with perfect reliability, the researcher decides to use Spearman's (1904) method to correct the correlation coefficient for the unreliability of the observations. Assuming that the reliability of mean RT $r_{\hat{\theta}\hat{\theta}}$ is .65 and the reliability of the depression questionnaire $r_{\hat{\beta}\hat{\beta}}$ is .39, she obtains the following corrected correlation coefficient using Spearman's formula:

$$r_{\theta\beta} = \frac{-0.57}{\sqrt{0.65} \times \sqrt{0.39}} = -1.13. \tag{8.8}$$

Table 8.1 Data for the Hypothetical Experiment on the Relationship between Mean RT and Depression.

Participant i	Mean RT ($\hat{\theta}$)	Depression score ($\hat{\beta}$)
1	435	27
2	491	24
3	448	33
4	363	16
5	402	8
6	390	19
7	394	18
8	375	12
9	468	18
10	428	25
Observed variance (s^2)	1732.04	54.67
Error variance (s^2_ϵ)	605.73	33.56
Reliability	.65	.39

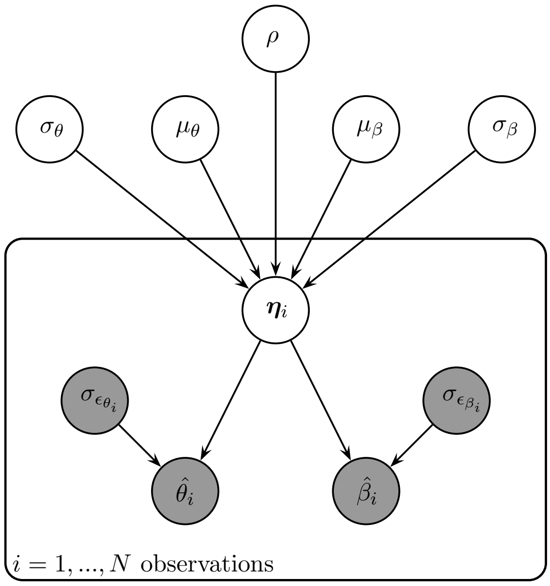
Note. The reliabilities are computed with the observed variances ($s^2_{\hat{\theta}}$ and $s^2_{\hat{\beta}}$) and the error variances ($s^2_{\epsilon_\theta}$ and $s^2_{\epsilon_\beta}$) using Equation 8.6 and Equation 8.7.

Our hypothetical experiment illustrates two shortcomings of Spearman's (1904) correction for attenuation. First, Spearman's method can produce corrected correlation coefficients in excess of 1.00 or -1.00, implying —as shown in Equation 8.3 and Equation 8.4— that true score variance is larger than the total observed variance (i.e., true score variance + error variance; see also Muchinsky, 1996). Second, Spearman's correction assumes homogenous error variances. Consider, however, our illustrative RT experiment; mean RT is unlikely to be measured with the same precision for each participant. In fact, in most psychological investigations, error variance is likely to differ across the observations, as the observations usually represent different participants.

In the next section, we describe Behseta et al.'s (2009) Bayesian correction for attenuation, an alternative to Spearman's (1904) method that does not suffer from the above mentioned shortcomings: The Bayesian method yields corrected correlations that are bounded by -1 and 1 and can be applied to situations with heterogeneous error variances. Behseta et al.'s approach is conceptually straightforward and brings along the benefits of Bayesian modeling, such as easy-to-use statistical software and a coherent inferential framework.

8.3 Bayesian Correction for the Attenuation of the Correlation

Behseta et al.'s (2009) correction for the attenuation of the correlation is based on Bayesian multilevel (or hierarchical) modeling (e.g., Farrell & Ludwig, 2008; Gelman & Hill, 2007; M. D. Lee, 2011; Matzke & Wagenmakers, 2009; Rouder et al., 2005) and estimates the posterior distribution of the corrected correlation coefficient using Markov chain Monte Carlo sampling (MCMC; Gamerman & Lopes, 2006; Gilks et al., 1996). As Behseta et al. showed through a series of simulation studies, the Bayesian procedure is superior to Spearman's (1904) method in terms of the accuracy of the recovered corrected correlation and the coverage of the confidence interval, especially when the assumption of normality is violated. In this section, we describe Behseta et al.'s Bayesian approach in detail. The graphical representation of the Bayesian correction method is shown in Figure 8.1.



$$\mu_\theta, \mu_\beta \sim \text{Normal}(0, 1000)$$

$$\sigma_\theta, \sigma_\beta \sim \text{Uniform}(0, 100)$$

$$\rho \sim \text{Uniform}(-1, 1)$$

$$\boldsymbol{\eta}_i \sim \text{MultivariateNormal}\Big((\mu_\theta, \mu_\beta)\,, \begin{bmatrix} \sigma_\theta^2 & \rho\sigma_\theta\sigma_\beta \\ \rho\sigma_\theta\sigma_\beta & \sigma_\beta^2 \end{bmatrix}\Big)$$

$$\hat{\theta}_i \sim \text{Normal}(\eta_{1_i}, \sigma^2_{\epsilon_{\theta i}})$$

$$\hat{\beta}_i \sim \text{Normal}(\eta_{2_i}, \sigma^2_{\epsilon_{\beta i}})$$

Figure 8.1 *Graphical model for Behseta et al.'s (2009) Bayesian correction for the attenuation of the correlation.* Observed variables are represented by shaded nodes and unobserved variables are represented by unshaded nodes. The graph structure indicates dependencies between the nodes (e.g., M. D. Lee, 2008). The normal distributions are parameterized in terms of the variance σ^2. $\eta_{1_i} = \theta_i$; $\eta_{2_i} = \beta_i$.

Level I: Modeling the Observed Data

The shaded nodes in the horizontal panel in Figure 8.1 represent the observed variables in the model. As before, let θ and β represent the true values, $\hat{\theta}$ and $\hat{\beta}$ the corresponding observed values, and ϵ_θ and ϵ_β the errors associated with θ and β, respectively. For each observation i, $i = 1, ..., N$, $\hat{\theta}_i$ and $\hat{\beta}_i$ are given by

$$\hat{\theta}_i = \theta_i + \epsilon_{\theta_i} \tag{8.9}$$

and

$$\hat{\beta}_i = \beta_i + \epsilon_{\beta_i}. \tag{8.10}$$

The error terms ϵ_{θ_i} and ϵ_{β_i} are assumed to be drawn from independent zero-centered normal distributions with variance $\sigma_{\epsilon_i^2}$:

$$\epsilon_{\theta_i} \sim \text{Normal}(0, \sigma^2_{\epsilon_{\theta_i}}) \tag{8.11}$$

and

$$\epsilon_{\beta_i} \sim \text{Normal}(0, \sigma^2_{\epsilon_{\beta_i}}). \tag{8.12}$$

The error variances are assumed to be known a priori or are estimated from data. Note that contrary to Spearman's (1904) correction, the Bayesian approach does not assume homogenous error variances across the N observations —each observation i has its own error variance.

Level II: Modeling Unobserved Means, Variances, and Correlations

The unshaded nodes in Figure 8.1 represent the unobserved variables in the model. For each observation i, $i = 1, ..., N$, $\boldsymbol{\eta}_i = (\theta_i, \beta_i)$ is assumed to follow a bivariate normal distribution, with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\eta}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{8.13}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_\theta \\ \mu_\beta \end{pmatrix} \tag{8.14}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_\theta & \rho\sigma_\theta\sigma_\beta \\ \rho\sigma_\theta\sigma_\beta & \sigma^2_\beta \end{pmatrix}. \tag{8.15}$$

Here ρ is the corrected correlation between θ and β —the correlation of interest that is uncontaminated by measurement error.

The means $\boldsymbol{\mu}$ and the three elements (i.e., σ_θ, σ_β, and ρ) of the variance-covariance matrix $\boldsymbol{\Sigma}$ are estimated from the data and require prior distributions. In the present article, we will use the following prior set-up:

$$\mu_\theta, \mu_\beta \sim \text{Normal}(0, 1000), \tag{8.16}$$

$$\sigma_\theta, \sigma_\beta \sim \text{Uniform}(0, 100), \tag{8.17}$$

and

$$\rho \sim \text{Uniform}(-1, 1). \tag{8.18}$$

Note that Behseta et al. (2009) used a slightly different prior set-up, where $\boldsymbol{\Sigma}$ was assumed to follow an Inverse-Wishart distribution. We chose to model the individual components in $\boldsymbol{\Sigma}$ rather than $\boldsymbol{\Sigma}$ itself because this provides an intuitive specification that allows users to adapt the range of the uniform prior for σ_θ and σ_β to the measurement scale of their variables in a straightforward manner.

Bayesian Parameter Estimation

In Bayesian parameter estimation, the prior distributions for the model parameters are updated by the incoming data to obtain posterior distributions. The posterior distribution often cannot be derived analytically; rather it must be approximated using numerical sampling techniques, such as MCMC sampling (Gamerman & Lopes, 2006; Gilks et al., 1996). The posterior distribution quantifies the uncertainty of the parameter estimates. The central tendency of the posterior, such as the mean or median, is often used as a point estimate. The dispersion of the posterior, such as the standard deviation or the percentiles, quantifies the precision of the parameter estimate: the larger the dispersion, the greater the uncertainty in the estimated parameter. For example, the 95% Bayesian credible interval for the corrected correlation ρ ranges from the 2.5^{th} to the 97.5^{th} percentile of its posterior distribution, indicating that we can be 95% confident that the true value of ρ lies within this range.

Parameter estimation for the present approach may proceed using standard Bayesian statistical software, such as WinBUGS (Bayesian inference Using Gibbs Sampling for Windows; Lunn et al., 2012; for an introduction for psychologists, see Kruschke, 2010b, and M. D. Lee & Wagenmakers, 2013). The WinBUGS script is presented in Appendix E.1. Note that the WinBUGS script requires minimal, if any, modification to run under OpenBUGS (Lunn et al., 2009) or JAGS (Plummer, 2003, 2013). The Stan project (Stan Development Team, 2012) provides yet another alternative to obtain posterior distributions for the parameters.

Bayesian Hypothesis Testing

Once the model parameters are estimated, we can formally assess the presence of an association using Bayes hypothesis testing. Throughout the article, we will rely on the Bayes factor —a popular Bayesian model selection measure— to quantify the probability of the data under the null hypothesis (H_0: $\rho = 0$) relative to the probability of the data under the alternative hypothesis (H_1: $\rho \neq 0$). For instance, BF_{01} of 10 indicates that the data are 10 times more likely under the null hypothesis than under the alternative hypothesis. Alternatively, BF_{01} of $\frac{1}{10}$ indicates that the data are 10 times more likely under the alternative hypothesis than under the null hypothesis (e.g., Jeffreys, 1961; Kass & Raftery, 1995).

We will compute two-sided Bayes factors using the Savage-Dickey density ratio method (e.g., Dickey & Lientz, 1970; Wagenmakers et al., 2010; Wetzels, Grasman, & Wagenmakers, 2010), assuming a uniform prior distribution for the correlation ρ parameter. The Savage-Dickey density ratio is an intuitive and flexible approach for the computation of Bayes factors in nested model comparison. Applied to the present situation, BF_{01} is given by the ratio of the height of the posterior and the prior distribution of ρ under the alternative hypothesis at $\rho = 0$. The height of the prior distribution is calculated by evaluating the uniform probability density function on -1.00 and 1.00 at $\rho = 0$; the height of the uniform prior distribution at $\rho = 0$ equals $\frac{1}{2}$. The height of the posterior distribution is calculated as follows. We first obtain samples from the posterior distribution of ρ using WinBUGS. We then fit to the posterior samples a scaled beta distribution with parameters α and β. Lastly, we evaluate the height of the scaled beta distribution at $\rho = 0$ using the obtained α and β parameters. One-sided (i.e., order-restricted) Bayes factors will be computed as recommended by Morey and Wagenmakers (2014), namely by correcting the two-sided Bayes factor using the proportion of posterior samples that is consistent with the order-restriction. The R script (R Core Team, 2012) for the computation of the Bayes factor is available in the supplemental materials at `http://dora.erbe-matzke.com/publications.html`.

8.4 Empirical Examples

In this section, we illustrate the use of Behseta et al.'s (2009) Bayesian correction for attenuation with two empirical data sets. In the first example, we assessed the correlation between parameters of cumulative prospect theory (CPT; Tversky & Kahneman, 1992) measured at two different time points. In the second example, we assessed the correlation between general intelligence and the drift rate parameter of the Ratcliff diffusion model (Ratcliff, 1978). In order to apply the graphical model shown in Figure 8.1, we first estimated posterior distributions for the CPT and diffusion model parameters for each participant separately. We then computed posterior distributions for the uncorrected[1] and the corrected correlation coefficients using the mean of the posterior distribution of the individual model parameters as point estimate.[2] In the corrected analysis, we used the variance of the posterior distribution of the individual model parameters as estimate for the participant-specific error variance. Finally, we formally assessed the presence of an association using Bayes hypothesis testing.

Example 1: Inference for the Correlation between Parameters of CPT

As our first example, we computed the uncorrected and corrected correlation in a data set obtained from a decision making experiment reported in Glöckner and Pachur (2012). The 64 participants were instructed to choose between monetary gambles in two experimental sessions. The two sessions were separated by one week and each featured 138 two-outcome gambles. The observed choice data were modeled using cumulative prospect theory (CPT; Tversky & Kahneman, 1992). CPT has a number of free parameters that reflect specific individual differences. Here we focus on the δ parameter that governs how individual decision makers weight the probability information of the gambles: higher values of δ indicate high risk aversion, whereas lower values of δ indicate less risk aversion.

As in other models in the decision making literature, the CPT parameters are assumed to be relatively stable across short periods of time. Here we therefore examined the association between the δ parameter measured at the two experimental sessions. The CPT was fit to the data of each individual participant, separately for the two measurement occasions. Model parameters were obtained using Bayesian parameter estimation with JAGS (Plummer, 2013), by adapting an existing model by Nilsson et al. (2011).[3] The prior for δ was uninformative across possible

[1] In the uncorrected Bayesian analysis, the bivariate normal distribution in Equation 8.13 was placed directly on the observed data.

[2] As we will discuss later, in the Bayesian framework, we are not limited to the two-step procedure outlined in this section; Bayesian hierarchical modeling allows for the *simultaneous* estimation of the individual parameters and the group-level means and covariances.

[3] The CPT account of performance in the Glöckner and Pachur (2012) data set is merely an illustration; we do not suggest that the CPT with the present parameter setting provides the best, or even an adequate, description

parameter values found in previous research but excluded theoretically implausible values.

We computed posterior distributions for the uncorrected and corrected correlation between the δ_1 and the δ_2 parameters. We treated the mean of the posterior distribution of the individual δ_1 and δ_2 parameters as observed data. In the corrected analysis, we used the variance of the posterior distribution of the individual δ_1 and δ_2 parameters as estimate for the participant-specific error variance. Lastly, we computed one-sided Bayes factors for the uncorrected and corrected correlation to quantify the evidence that the data provide for H_0 ($\rho = 0$) relative to H_1 ($\rho > 0$).

Results

The results are shown in Figure 8.2. The top-row panels show scatterplots of the observed $\hat{\delta}_{i,1}$ and $\hat{\delta}_{i,2}$ parameters and the standard deviation of the measurement errors (i.e., $\sigma_{\epsilon_{\delta_{i,1}}}$ and $\sigma_{\epsilon_{\delta_{i,2}}}$). The bottom left panel shows a scatterplot of the mean of the posterior distribution of the corrected δ parameters (i.e., $\boldsymbol{\eta}_i = (\delta_{i,1}, \delta_{i,2})$) and the corresponding posterior standard deviations. The bottom right panel shows the posterior distribution of the uncorrected and the corrected correlation (i.e, ρ).

As shown in the upper left panel of Figure 8.2, the uncorrected Pearson correlation between the observed $\hat{\delta}_1$ and $\hat{\delta}_2$ parameters is .62. If we take into account the uncertainty of the observations, the correlation increases substantially: The bottom left panel shows that the posterior means of the corrected $\delta_{i,1}$ and $\delta_{i,2}$ parameters are associated very highly; the bottom right panel shows that the posterior distribution of the corrected ρ parameter is shifted to the right relative to the posterior of the uncorrected correlation. In fact, after correcting for the noise in $\hat{\delta}_1$ and $\hat{\delta}_2$, the mean of the posterior distribution of ρ increases from .61 to .92.

One-sided Bayes factors indicate decisive evidence (Jeffreys, 1961) for the presence of an association for the corrected as well as the uncorrected ρ parameter; in both cases, the data are more than 4,000,000 times more likely under H_1 than under H_0. This result is visually evident from the fact that the posterior distributions are located away from zero such that their height at $\rho = 0$ is all but negligible.

Note that the dramatic increase in the correlation observed in the present data set is not unusual. Figure 8.3 shows the results of a simulation study where we investigated the magnitude of the expected attenuation for different values of the latent correlation in a parameter setting that resembles the one found in the present data set. We conducted five sets of simulations, each with a different true “latent” correlation: .92 (i.e., the posterior mean of the corrected ρ in the present data set), .21, 0, -.21, and -.92. For each set of simulations, we generated 1,000 synthetic data sets with $N = 64$, using the error variances $\sigma^2_{\epsilon_{\delta_{i,1}}}$ and $\sigma^2_{\epsilon_{\delta_{i,2}}}$ obtained from fitting the CPT to the data, and the posterior mean of the μ_{δ_1}, μ_{δ_2}, σ_{δ_1}, and σ_{δ_2} parameters estimated with the Bayesian correction method.[4] In each synthetic data set, we then computed the attenuated “observed” correlation. The gray violin plots in Figure 8.3 show the distribution of the predicted attenuated correlations for the five values of the latent correlation.

Two results are noteworthy. First, all else being equal, the larger the absolute value of the true latent correlation, the larger the attenuation. This relationship is also evident from Equation 8.3 and Equation 8.4. Second, the observed attenuated correlation in the empirical data (i.e., .62; horizontal dashed line) is slightly higher than expected, but is well within the 2.5^{th} and 97.5^{th} percentile of the attenuated correlations predicted by the model with the present parameter setting.

of the data of the individual participants. Note also that Glöckner and Pachur reported the results from fitting a slightly different model than the one used in the present article.

[4] Note that this procedure is not the same as the posterior predictive assessment of model fit (e.g., Gelman & Hill, 2007; Gelman et al., 1996).

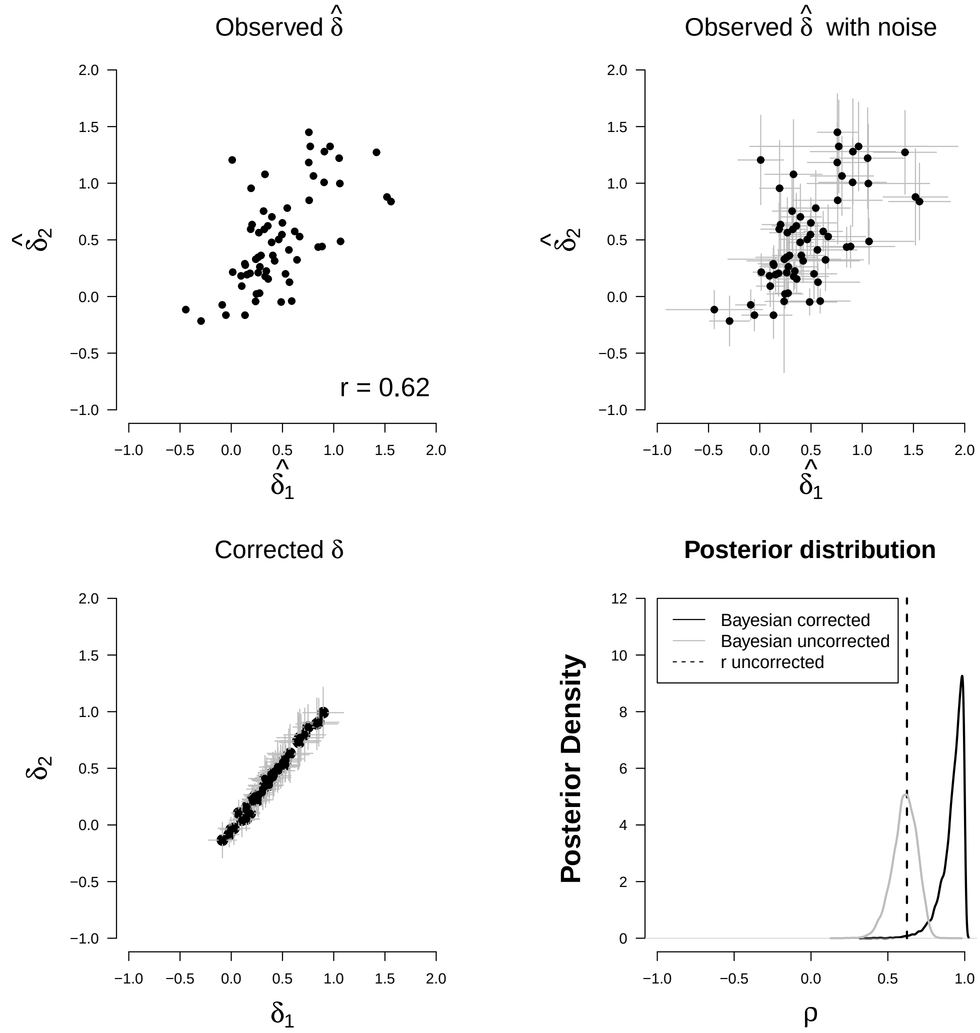


Figure 8.2 *Corrected and uncorrected correlation between parameters of cumulative prospect theory (CPT).* The top left panel shows a scatterplot of the observed $\hat{\delta}_{i,1}$ and $\hat{\delta}_{i,2}$ parameters. The top right panel shows a scatterplot of the observed $\hat{\delta}_{i,1}$ and $\hat{\delta}_{i,2}$ parameters and the standard deviation of the measurement errors (i.e., $\sigma_{\epsilon_{\delta_{i,1}}}$ and $\sigma_{\epsilon_{\delta_{i,2}}}$; gray lines). The $\hat{\delta}_{i,1}$, $\hat{\delta}_{i,2}$, $\sigma_{\epsilon_{\delta_{i,1}}}$, and $\sigma_{\epsilon_{\delta_{i,2}}}$ values were obtained from fitting the CPT to the data. The bottom left panel shows a scatterplot of the posterior mean of the corrected $\delta_{i,1}$ and $\delta_{i,2}$ parameters (i.e., $\boldsymbol{\eta}_i = (\delta_{i,1}, \delta_{i,2})$). The gray lines show the standard deviation of the posterior distribution of the parameters. The bottom right panel shows the posterior distribution of the uncorrected (gray density line) and the corrected correlation (i.e, ρ; black density line). The dashed vertical line shows the Pearson correlation coefficient computed with the observed $\hat{\delta}_{i,1}$ and $\hat{\delta}_{i,2}$ parameters. r = Pearson correlation coefficient.

In sum, correcting the correlation for measurement noise resulted in a dramatic increase in the correlation between the CPT parameters; the mean of the posterior distribution of the correlation parameter increased from .62 to .92. Despite this increase, the Bayes factor indicated decisive evidence for the presence of an association in the corrected as well as the uncorrected analysis.

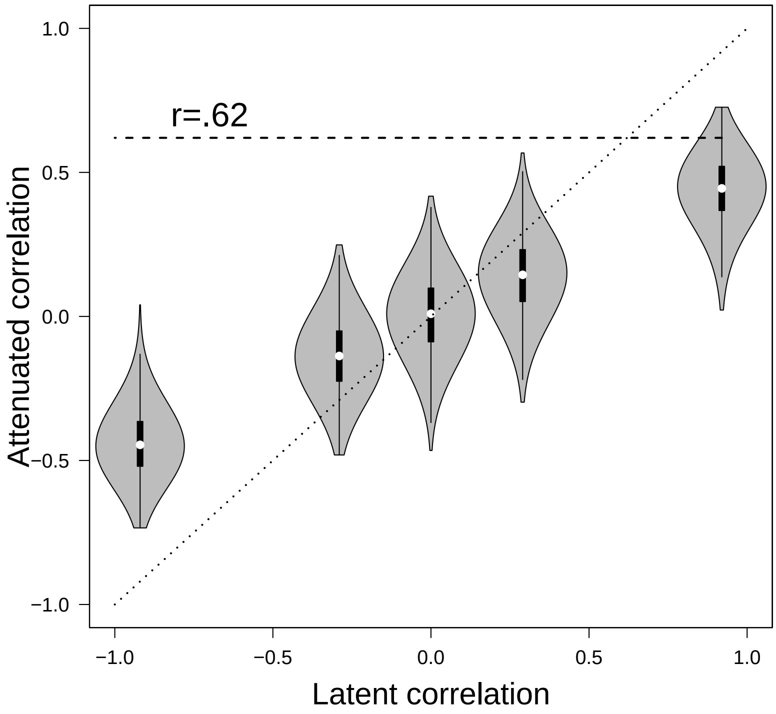


Figure 8.3 *Expected attenuation of the correlation between the δ parameters of cumulative prospect theory (CPT).* The figure shows that the attenuation of the correlation increases with the absolute value of the latent correlation. The gray violin plots show the distribution of the predicted attenuated correlations for five values of the latent correlation. Violin plots (e.g., Hintze & Nelson, 1998) combine information available from density plots with information about summary statistics in the form of box plots. The black boxplot in each violin plot ranges from the 25^{th} to the 75^{th} percentile of the attenuated correlations predicted by the model, where the white circle represents the median of the predictions. The predicted correlations were generated using the error variances $\sigma^2_{\epsilon_{\delta_{i,1}}}$ and $\sigma^2_{\epsilon_{\delta_{i,2}}}$ obtained from fitting the CPT to the data, and the posterior mean of the μ_{δ_1}, μ_{δ_2}, σ_{δ_1}, and σ_{δ_2} parameters estimated with the Bayesian correction method. The dashed line shows the observed attenuated correlation in the empirical data. r = Pearson correlation coefficient.

Example 2: Inference for the Correlation between General intelligence and Diffusion Model Drift Rate

As a second example, we computed the uncorrected and corrected correlation in a data set collected by Weeda and Verouden (unpublished data). The data set featured response times (RT) and accuracy data from 51 participants performing a two-choice RT task. The stimuli were borrowed from the π-paradigm (Vickers, Nettelbeck, & Willson, 1972; Jensen, 1998) and consisted of a series of configurations, each with one horizontal and two vertical lines (i.e.; two legs) that together formed the letter π, with one of the vertical lines longer than the other. The task was to indicate by means of a button press whether the left or the right leg of the π was longer. Task difficulty was manipulated on three levels —easy, medium, and difficult— by varying the difference between the length of the two legs.

The RT and accuracy data were modeled with the Ratcliff diffusion model (Ratcliff, 1978; Wagenmakers, 2009). The diffusion model provides a theoretical account of performance in speeded two-choice tasks. The four key parameters of the diffusion model correspond to well-defined psychological processes (Ratcliff & McKoon, 2008; Voss et al., 2004), such as response caution (a), a priori bias (z), the time taken up by processes unrelated to decision making (e.g., encoding and motor processes; T_{er}), and —the parameter of interest in the present article— the rate of information accumulation drift rate (v).[5]

The drift rate parameter of the diffusion model has been repeatedly associated with higher cognitive functions and reasoning (i.e., Ratcliff, Schmiedek, & McKoon, 2008; Ratcliff, Thapar, & McKoon, 2010; Schmiedek et al., 2007; van Ravenzwaaij, Brown, & Wagenmakers, 2011), and this is why we focus here on the correlation between drift rate and general intelligence. The four key diffusion model parameters were estimated from the RT and accuracy data for each participant separately using Bayesian parameter estimation with the diffusion model JAGS module (Wabersich & Vandekerckhove, 2014). We used uninformative prior distributions based on parameter values reported in Matzke and Wagenmakers (2009). As drift rate is known to decrease with increasing task difficulty (e.g., Ratcliff & McKoon, 2008), we used the following order-restriction: $v_{difficult} < v_{medium} < v_{easy}$. The remaining parameters were constrained to be equal across the conditions, and we set $z = \frac{a}{2}$.[6] General intelligence was measured by the total score of the 20-min version of the Raven Progressive Matrices Test (Hamel & Schmittmann, 2006; Raven, Raven, & Court, 1998).

We computed posterior distributions for the uncorrected and corrected correlation between the mean of the drift rate parameters across the three task difficulty conditions ($\bar{v}$) and the Raven total score (g). For mean drift rate $\bar{v}$, we treated the mean of the posterior distribution of the individual $\bar{v}$ parameters as observed data.[7] In the corrected analysis, we used the variance of the posterior distribution of the individual $\bar{v}$ parameters as estimate for the participant-specific error variance. For the Raven total score g, we assumed homogenous error variance and —for illustrative purposes— investigated how the extent of the correction varies as a function of the amount of measurement noise assumed in the data. Specifically, we examined three scenarios: we assumed that 5%, 25%, and 55% of the total variance in Raven scores is attributable to measurement error, corresponding to excellent, acceptable, and very poor reliability, respectively. Lastly, we computed one-sided Bayes factors for the uncorrected and corrected correlation to quantify the evidence that the data provide for H_0 ($\rho = 0$) relative to H_1 ($\rho > 0$) under each scenario.

[5] In addition to these key parameters, the diffusion model also features parameters that describe the trial-to-trial variability of the key parameters.

[6] The diffusion model account of performance in the Weeda and Verouden data set (unpublished data) is merely an illustration; we do not suggest that the diffusion model with the present parameter constraints provides the best, or even an adequate, description of the data of the individual participants.

[7] Note that the scale of both drift rate and general intelligence are bounded: $\bar{v}$ can take on values between 0 and 5.86 (i.e., prior range) and the Raven total score can take on values between 0 and 36. The use of the bivariate normal group-level distribution shown in Figure 8.1 is therefore theoretically unjustified. As a solution, we may transform the individual $\bar{v}$ parameters and the Raven scores g to the real line using a probit transformation. Additional analyses not reported here confirmed that using the transformed $\bar{v}$ and g values yields results that are very similar to the ones obtained using the untransformed drift rates and Raven scores. For simplicity, in the present article, we will report the results of modeling the untransformed $\bar{v}$ and Raven g values.

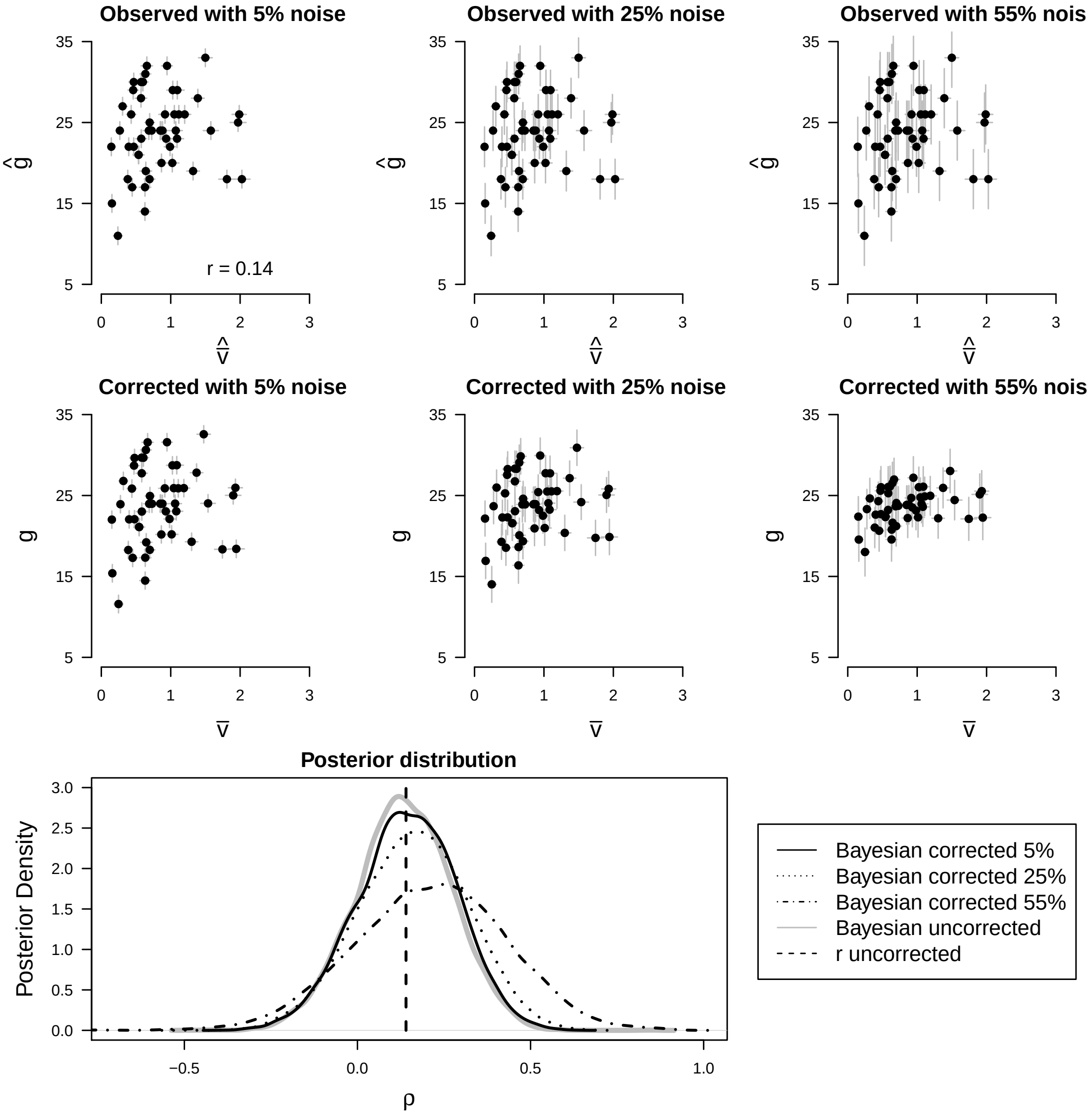


Figure 8.4 *Corrected and uncorrected correlation between mean drift rate $\bar{v}$ and Raven total score g.* The top panels show scatterplots of the observed mean drift rates $\hat{\bar{v}}_i$ and Raven total scores $\hat{g}_i$ with the standard deviation of the measurement errors (i.e., $\sigma_{\epsilon_{\bar{v}_i}}$ and σ_{ϵ_g}; gray lines) for the 5%, 25%, and 55% measurement noise scenarios. The $\hat{\bar{v}}_i$ and the $\sigma_{\epsilon_{\bar{v}_i}}$ values were obtained from fitting the diffusion model to the data. The middle panels show scatterplots of the posterior mean of the corrected $\bar{v}_i$ and g_i parameters (i.e., $\boldsymbol{\eta}_i = (\bar{v}_i, g_i)$). The gray lines show the standard deviation of the posterior distribution of the parameters. The bottom panel shows the posterior distribution of the uncorrected (gray density line) and the corrected correlations (i.e, ρ) for the 5% (solid black density line), 25% (dotted black density line), and the 55% (dotted-dashed black density line) measurement noise scenarios. The dashed vertical line shows the Pearson correlation coefficient computed with the observed mean drift rates $\hat{\bar{v}}_i$ and Raven total scores $\hat{g}_i$. r = Pearson correlation coefficient.

Results

The results are shown in Figure 8.4. The top-row panels show scatterplots of the observed mean drift rates $\hat{\bar{v}}$ and Raven total scores $\hat{g}$ with the standard deviation of the measurement errors (i.e., $\sigma_{\epsilon_{\bar{v}_i}}$ and σ_{ϵ_g}) for the 5%, 25%, and 55% noise scenarios. The middle-row panels show scatterplots of the mean of the posterior distribution of the corrected $\bar{v}$ and g parameters (i.e., $\boldsymbol{\eta}_i = (\bar{v}_i, g_i)$) and the corresponding posterior standard deviations for each scenario. The bottom panel shows the posterior distribution of the uncorrected and the corrected correlation (i.e, ρ) for each scenario.

As shown in the upper left panel of Figure 8.4, the uncorrected Pearson correlation between the observed $\hat{\bar{v}}$ parameters and the observed Raven scores $\hat{g}$ is .14. As expected, the magnitude of the correction for attenuation increases with increasing error variance: The middle-row panels show that the association between the posterior means of the corrected $\bar{v}_i$ and g_i parameters becomes stronger; the bottom panel shows that the mean of the posterior distribution of the corrected ρ parameter is progressively shifted to higher values. Note, however, that the correction is modest even if we assume that the Raven total score is an extremely unreliable indicator of general intelligence. The mean of the posterior distribution of ρ equals .13 in the uncorrected analysis, .14 in the corrected analysis with 5% noise, .16 in the corrected analysis with 25% noise, and .21 in the corrected analysis with 55% noise. Note also that the posterior of ρ tends to be quite spread out, a tendency that becomes more pronounced with increasing error variance.

One-sided Bayes factors indicate evidence *against* the presence of an association between mean drift rate and Raven total score. The evidence, however, is "worth no more than a bare mention" (Jeffreys, 1961). The BF_{01} decreases from 2.13 in the uncorrected analysis to 2.00 in the corrected analysis with 5% noise, to 1.75 in the corrected analysis with 25% noise, and to 1.32 in the corrected analysis with 55% noise. Even with extremely unreliable Raven scores, the data are thus still more likely to have occurred under H_0 than under H_A. Note however that BF_{01} of 1.32 —or even BF_{01} of 2.13— constitutes almost perfectly ambiguous evidence, indicating that the data are not sufficiently diagnostic to discriminate between H_0 and H_A. Inspection of the posterior distribution of the ρ parameters suggests a similar conclusion: ρ is estimated quite imprecisely (i.e., the posteriors are spread out) in all four analyses.

Figure 8.5 shows the results of a simulation study where we investigated the magnitude of the expected attenuation for different values of the latent correlation in a parameter setting that resembles the one found in the present data set. Throughout the simulations, we assumed that 55% of the total variance of the Raven scores is attributable to error variance. We conducted five sets of simulations, each with a different value of the true "latent" correlation: .92, .21 (i.e., the posterior mean of the corrected ρ in the present data set), 0, -.21, and -.92. For each set of simulations, we generated 1,000 synthetic data sets with $N = 51$, using the $\sigma^2_{\epsilon_{\bar{v}_i}}$ parameters obtained from fitting the diffusion model to the data, and the posterior means of the $\mu_{\bar{v}}$, μ_g, $\sigma_{\bar{v}}$, and σ_g parameters estimated with the Bayesian correction method. In each synthetic data set, we then computed the attenuated "observed" correlation. The gray violin plots in Figure 8.5 show the distribution of the predicted attenuated correlations for the five values of the latent correlation.

As pointed out earlier, all else being equal, the larger the absolute value of the true latent correlation, the larger the attenuation. Moreover, considering the relatively low corrected correlation in the present data set, an attenuation of only $.21 - .14 = .07$ is perfectly reasonable. In fact, the median of the attenuated correlations predicted by the model with the present parameter setting very closely approximates the observed attenuated correlation in the empirical data (i.e., .14; horizontal dashed line).

In sum, correcting for measurement noise resulted in negligible increase in the correlation between drift rate and general intelligence; even with unrealistically unreliable Raven scores, the posterior mean of the correlation parameter increased from .13 only to .21. Regardless of the type of analysis (uncorrected or uncorrected) and regardless of the magnitude of the error variance, the Bayes factor indicated evidence against the presence of an association between drift rate and general intelligence. The evidence for the null hypothesis was, however, only anecdotal, a result that is attributable to the substantial uncertainty in the estimated correlation parameters.

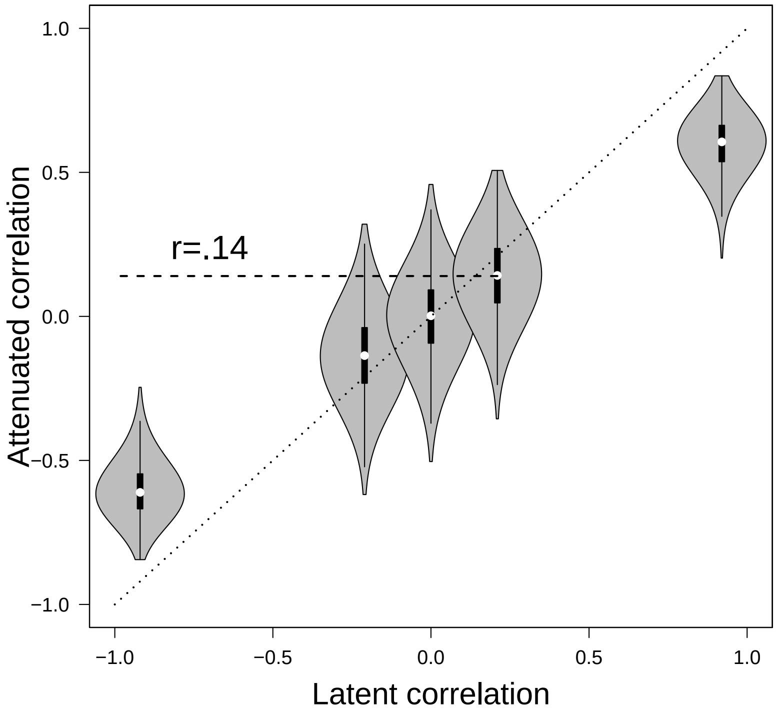


Figure 8.5 *Expected attenuation of the correlation between mean drift rate and general intelligence.* The figure shows that the attenuation of the correlation increases with the absolute value of the latent correlation. The gray violin plots show the distribution of the predicted attenuated correlations for five values of the latent correlation. The black boxplot in each violin plot ranges from the 25^{th} to the 75^{th} percentile of the attenuated correlations predicted by the model, where the white circle represents the median of the predictions. The predicted correlations were generated assuming 55% error variance in the Rave scores, using the $\sigma^2_{\epsilon_{\bar{v}_i}}$ parameters obtained from fitting the diffusion model to the data, and the posterior means of the $\mu_{\bar{v}}$, μ_g, $\sigma_{\bar{v}}$, and σ_g parameters estimated with the Bayesian correction method. The dashed line shows the observed attenuated correlation in the empirical data. r = Pearson correlation coefficient.

8.5 Discussion

Although various approaches are available to correct the correlation in the presence of measurement error, such corrections are presently the exception rather than the rule in experimental psychology. The goal of the present paper was therefore to demonstrate the application of the Bayesian correction of attenuated correlations (Behseta et al., 2009). We illustrated the use of the Bayesian method with two empirical data sets; in each data set, we first estimated posterior distributions for the uncorrected and corrected correlation and then computed Bayes factors to quantify the evidence that the data provide for the presence of the association.

In our first example, we computed the uncorrected and corrected correlation between parameters of cumulative prospect theory (Tversky & Kahneman, 1992) and demonstrated that correcting for measurement noise can result in a dramatic increase in the correlation: the mean of the posterior distribution of the correlation parameter increased from .61 to .92. The Bayes factor indicated decisive evidence for the presence of an association in the corrected as well as the uncorrected analysis. In our second example, we assessed the correlation between general intelligence and the drift rate parameter of the diffusion model (Ratcliff, 1978; Wagenmakers, 2009), where we examined three scenarios: we assumed that 5%, 25%, and 55% of the total variance in Raven scores is attributable to measurement error, corresponding to excellent, acceptable, and very poor reliability, respectively. Correcting for measurement noise resulted in negligible increase in the correlation; even with extremely unreliable Raven scores, the posterior mean of the correlation parameter increased from .13 to only .21. In all analyses, we obtained anecdotal evidence against the presence of an association between drift rate and general intelligence, a result that is attributable to the substantial uncertainty in the estimated correlation parameters.

Behseta et al.'s (2009) Bayesian correction for attenuation is easy-to-use and conceptually straightforward. In fact, the present approach can be viewed as a simple Bayesian structural equation model with two latent variables, each with a single indicator (see, for example, S.-Y. Lee, 2007; Song & Lee, 2012). The original formulation of the Bayesian correction method relies on a slightly different prior set-up than the one used in the present article. Specifically, Behseta et al. used an Inverse-Wishart distribution to model the variance-covariance matrix of the corrected observations, whereas we chose to model the standard deviations and the correlation separately using uniform distributions. We feel that the present specification is more intuitive and allows users to adapt the range of the uniform prior for the standard deviations to the measurement scale of their variables in a straightforward manner. Note also that Bayesian parameter estimation is insensitive to the choice of the prior as long as sufficiently informative data are available (e.g., Edwards et al., 1963; Gill, 2002; M. D. Lee & Wagenmakers, 2013).

Correcting the correlation for measurement noise is, of course, impossible unless the error variance of the observations is known or can be estimated from the data. Our investigation of the extent of the correction as a function of the amount of measurement noise in the Raven scores served merely as an illustration. In real-life applications, the magnitude of the error variance should not be cherry-picked to obtain the desired (higher) correlation; rather it should be estimated from the data. Bayesian inference is particularly suitable for modeling measurement error because the resulting posterior distributions can be automatically used to quantify the uncertainty of the parameter estimates.[8] Accordingly, in our two examples, we treated the mean of the posterior distribution of the CPT and diffusion model parameters as observed data and used the variance of the posterior distributions as estimate for the participant-specific error variance. Note also that within the Bayesian framework, we are not limited to the two-step procedure illustrated in the present article; Bayesian hierarchical modeling allows for the *simultaneous* estimation of the individual parameters and the group-level means and covariances, where the correlation is automatically adjusted for the uncertainty of the individual estimates. The Bayesian estimation of covariance structures is illustrated, for example, in Gelman and Hill (2007), Klauer (2010), Matzke et al. (in press), Rouder et al. (2008), and Rouder et al. (2007).

[8] Naturally, using the variance of the posterior distribution as estimate for the error variance is only sensible if the posteriors are approximately normally distributed.

Our literature review showed that nearly 50% of the articles published in the 2012 volume of the *Journal of Experimental Psychology: General* reported at least one Pearson product-moment correlation coefficient. Despite the wide-spread use of correlations, most researchers do not acknowledge explicitly that the observed correlation often underestimates the true correlation if the variables are measured with noise. Here we illustrated the use of a Bayesian correction procedure and showed that its application can dramatically increase the estimated correlation. Of course, estimating the uncertainty of the observations is not always feasible. Also, our simulations confirmed that for relatively low true correlations, the correction is likely to have only a negligible effect. We nevertheless urge researchers to carefully consider the issue of the attenuation and whenever possible correct the observed correlation for the uncertainty of the measurements.

Chapter 9

A Default Bayesian Hypothesis Test for Mediation

A modified version of this chapter is currently in press as:
Michèle B. Nuijten, Ruud Wetzels, Dora Matzke, Conor V. Dolan,
and Eric-Jan Wagenmakers (in press).
A default Bayesian hypothesis test for mediation. *Behavior Research Methods*.[1]

Abstract

In order to quantify the relationship between multiple variables, researchers often carry out a mediation analysis. In such an analysis, a mediator (e.g., knowledge of a healthy diet) transmits the effect from an independent variable (e.g., classroom instruction on a healthy diet) to a dependent variable (e.g., consumption of fruits and vegetables). Almost all mediation analyses in psychology use frequentist estimation and hypothesis testing techniques. A recent exception is Yuan and MacKinnon (2009), who outlined a Bayesian parameter estimation procedure for mediation analysis. Here we complete the Bayesian alternative to frequentist mediation analysis by specifying a default Bayesian hypothesis test based on the Jeffreys-Zellner-Siow approach. We further extend this default Bayesian test by allowing a comparison to directional or one-sided alternatives, using Markov chain Monte Carlo techniques implemented in JAGS. All Bayesian tests are implemented in the `R` package `BayesMed`.

9.1 Introduction

Mediated relationships are central to the theory and practice of psychology. In the prototypical scenario, a mediator (M, e.g., knowledge of a healthy diet) transmits the effect from an independent variable (X, e.g., classroom instruction on a healthy diet) to a dependent variable (Y, e.g., consumption of fruits and vegetables). Other examples arise in social psychology, where attitudes (X) cause intentions (M), and these intentions affect behavior (Y; MacKinnon, Fairchild, & Fritz,

[1] The final publication is available at `http://link.springer.com/article/10.3758/s13428-014-0470-2`.

2007). To quantify such relationships between mediator, independent variable, and dependent variable, researchers often use a toolbox of popular statistical methods collectively known as mediation analysis.

The currently available tools for mediation analyses are almost exclusively based on classical or frequentist statistics, featuring concepts such as confidence intervals and p values. Recently, Yuan and MacKinnon (2009) proposed an alternative, Bayesian mediation analysis that allows researchers to obtain a posterior distribution (and associated credible interval) for the mediated effect. This posterior distribution quantifies the uncertainty about the strength of the mediated effect under the assumption that the effect does not equal zero. This approach constitutes a valuable addition to the toolbox of mediation methods, but it specifically concerns parameter estimation and not hypothesis testing. As Yuan and MacKinnon (2009) state in their conclusion: "One important topic we have not covered in this article is hypothesis testing (...) Strict Bayesian hypothesis testing is based on Bayes factor, which is essentially the odds of the null hypothesis being true versus the alternative hypothesis being true, conditional on the observed data. The use of Bayesian hypothesis testing (...) would be a reasonable future research topic in Bayesian mediation analysis."

Hence, the goal of this paper is to add another statistical method to the toolbox of mediation analysis, namely the Bayes factor hypothesis test alluded to by Yuan and MacKinnon (2009). In the development of this test we have assumed a default specification of prior distributions based on the Jeffreys-Zellner-Siow framework (Liang et al., 2008), promoted in psychology by Jeff Rouder, Richard Morey, and colleagues (Rouder et al., 2009; Rouder, Morey, Speckman, & Province, 2012; Rouder & Morey, 2012) as well as ourselves (Wetzels et al., 2009, 2012; Wetzels & Wagenmakers, 2012). In our opinion, the default specification of prior distributions is useful because it provides a reference analysis that can be carried out regardless of subjective considerations about the topic at hand. Of course, researchers who have prior knowledge may wish to incorporate that knowledge into the models to devise a more informative test (e.g., Armstrong & Dienes, 2013; Dienes, 2011; Guo, Li, Yang, & Dienes, 2013). Here we focus solely on the default test as it pertains to the prototypical, single-level scenario of three variables.

The outline of this paper is as follows. First, we briefly discuss the conventional frequentist tests and the existing Bayesian mediation analysis proposed by Yuan and MacKinnon (2009). We then explain Bayesian hypothesis testing in general and introduce our default Bayesian hypothesis test for mediation. We illustrate the performance of our test with a simulation study and an example of a psychological study. Finally, we discuss software in which we implemented the Bayesian methods for mediation analysis: the `R` package `BayesMed`.

9.2 Frequentist Mediation Analysis

Consider a relation between an independent variable X and a dependent variable Y (see Figure 9.1, panel (a)). In a linear regression equation, such a relation can be represented as follows:

$$Y_i = \beta_{0(1)} + \tau X_i + \epsilon_{(1)}, \tag{9.1}$$

where subscript i identifies the participant, τ represents the relation between the independent variable X and the dependent variable Y, $\beta_{0(1)}$ is the intercept, and $\epsilon_{(1)}$ is the residual. The effect of X on Y, path τ, is called the total effect.

The relation between X and Y can be mediated by mediating variable M, which means that a change in X leads to a change in M, which then leads to a change in Y (see Figure 9.1, panel (b) and (c)). The resulting mediation model can be represented by the following set of linear regression equations:

$$Y_i = \beta_{0(2)} + \tau' X_i + \beta M_i + \epsilon_{(2)}, \tag{9.2}$$

$$M_i = \beta_{0(3)} + \alpha X_i + \epsilon_{(3)}, \tag{9.3}$$

where τ' represents the relation between X and Y after adjusting for the effects of the mediator M, α represents the relation between X and M, and β represents the relation between M and Y. Furthermore, $\epsilon_{(1)}$, $\epsilon_{(2)}$, and $\epsilon_{(3)}$ are assumed to be conditionally normally distributed, independent, homoskedastic residuals. Throughout the remainder of this paper, we focus on the standardized mediation model (i.e., a model in which the variables are standardized), and refer to the regression coefficients α, β, and τ' as paths.

The product of α and β is the indirect effect, or the mediated effect, assuming that α and β are independent. The remaining direct effect of X on Y is denoted with τ'. If the mediated effect differs from zero and τ' equals zero, the effect of X on Y is completely mediated by M (see Figure 9.1 panel (c)). If τ' has a value other than zero, the relationship between X and Y is only partially mediated by M (see Figure 9.1 panel (b)).

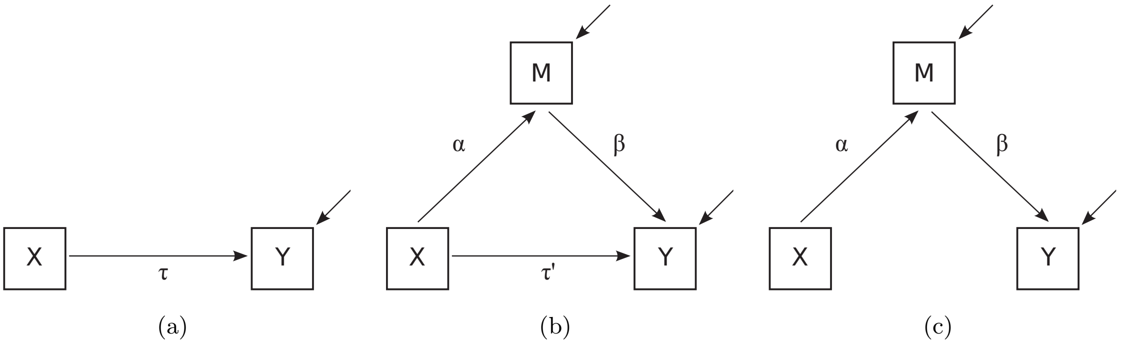


Figure 9.1 Diagram of the standard mediation model. Panel (a) shows a direct relation between X and Y, panel (b) shows partial mediation, and panel (c) shows full mediation. Diagonal arrows indicate that the graphical node is perturbed by an error term.

A popular method to test for mediation is to test paths α and β simultaneously. The estimated indirect effect $\hat{\alpha}\hat{\beta}$ is divided by its standard error and the resulting Z statistic is compared to the standard normal distribution to assess whether the effect is significantly different from zero, in which case the null hypothesis of no mediation can be rejected.

Several approaches are available for calculating the standard error of $\hat{\alpha}\hat{\beta}$, but the one used in the Sobel test (Sobel, 1982) is the most commonly reported:

$$\hat{\sigma}_{\hat{\alpha}\hat{\beta}} = \sqrt{\hat{\beta}^2 \hat{\sigma}^2_{\alpha} + \hat{\alpha}^2 \hat{\sigma}^2_{\beta}}, \tag{9.4}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the point estimates of the regression coefficients of the mediated effect, and $\hat{\sigma}_\alpha$ and $\hat{\sigma}_\beta$ their standard errors. The 95% confidence interval for the mediated effect is then given by $\hat{\alpha}\hat{\beta} \pm 1.96 \times \hat{\sigma}_{\hat{\alpha}\hat{\beta}}$.

One problem with the Sobel test is that it assumes a symmetrical sampling distribution for the mediated effect, whereas in reality this distribution is skewed (MacKinnon, Lockwood, & Hoffman, 1998). Consequently, the Sobel test has relatively low power (MacKinnon, Warsi, & Dwyer, 1995). A solution to this problem is to construct a confidence interval that takes the asymmetry of the distribution into account (see e.g., the product method of MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002) or the profile likelihood method (see Venzon & Moolgavkar, 1988).

Our goal here is not to argue against frequentist statistics in general, or p values in particular — for this, we refer the interested reader to the following articles and references therein: Berger and Delampady (1987); Berger and Wolpert (1988); Dienes (2011); Edwards et al. (1963); O'Hagan and Forster (2004); Rouder et al. (2012); Sellke, Bayarri, and Berger (2001); Wagenmakers (2007); Wetzels et al. (2011). Instead, our goal is to outline an additional Bayesian tool that can be used for mediation analysis. The availability of multiple tools is useful, not just because different situations may require different tools, but also because they allow a robustness check; if different tools yield opposing conclusions the careful researcher does well to report the results from both tests, indicating that the data are ambiguous in the sense that the conclusion depends on the analysis method at hand.

9.3 An Alternative: Bayesian Estimation

Our end goal is to propose a Bayesian alternative for the frequentist mediation test. Below we consider the Bayesian treatment of the mediation model in detail, but first we briefly discuss Bayesian inference in general terms. In the Bayesian framework, uncertainty is quantified by probability. Prior beliefs about parameters are formalized by prior probability distributions which are updated by the observed data to result in posterior beliefs or posterior distributions (Dienes, 2008; M. D. Lee & Wagenmakers, 2013; Kruschke, 2010b; O'Hagan & Forster, 2004).

The Bayesian updating process proceeds as follows. First, before observing the data under consideration, the Bayesian statistician assigns a probability distribution to one or more model parameters θ based on her prior knowledge — hence, this distribution is known as the prior probability distribution or simply "the prior", denoted $p(\theta)$. Next, one observes data $\mathbf{D}$, and the statistical model can be used to calculate the associated probability of $\mathbf{D}$ occurring under specific values of θ, a quantity known as the likelihood, denoted $p(\mathbf{D} \mid \theta)$. The prior distribution $p(\theta)$ is then updated to the posterior distribution $p(\theta \mid \mathbf{D})$ according to Bayes' rule:

$$p(\theta \mid \mathbf{D}) = \frac{p(\mathbf{D} \mid \theta)p(\theta)}{p(\mathbf{D})}. \tag{9.5}$$

Note that the marginal likelihood $p(\mathbf{D}) = \int p(\mathbf{D} \mid \theta)p(\theta)\,\mathrm{d}\theta$ functions as a normalizing constant that ensures that the posterior distribution will integrate to one. Because the normalizing constant does not contain θ it is not important for parameter estimation, and Equation 9.5 is often written as follows:

$$p(\theta \mid \mathbf{D}) \propto p(\mathbf{D} \mid \theta)p(\theta), \tag{9.6}$$

or in words:

$$\text{Posterior Distribution} \propto \text{Likelihood} \times \text{Prior Distribution},$$

where $\propto$ means "proportional to".

In a Bayesian mediation analysis the above updating principle can be used to transition from prior to posterior distributions for parameters α, β, and τ', as proposed by Yuan and MacKinnon (2009). Their method allows the user to determine the posterior distribution of the indirect effect

$\alpha\beta$, together with a 95% credible interval. This interval has the intuitive interpretation that we can be 95% confident that the true value of $\alpha\beta$ resides within this interval.

The approach of Yuan and MacKinnon (2009) is appropriate when estimating the size of the mediated effect. However, in experimental psychology the research question is often framed in terms of model selection or hypothesis testing, that is, the researcher seeks to answer the question: "does the effect exist?". Parameter estimation and model selection have different aims and, depending on the situation at hand, one procedure may be more appropriate than the other. We contend that there are situations where a hypothesis test is scientifically useful (e.g., Iverson, Wagenmakers, & Lee, 2010; Rouder et al., 2009) and in what follows we proceed to outline a default Bayesian hypothesis test for mediation. In order to keep this article self-contained, we will first introduce the principles of Bayesian hypothesis testing (Hoijtink, Klugkist, & Boelen, 2008; Myung & Pitt, 1997; Vandekerckhove, Matzke, & Wagenmakers, 2013; Wagenmakers et al., 2010).

9.4 Bayesian Hypothesis Testing

A Bayesian hypothesis test is a model selection procedure with two models or hypotheses. Assume two competing models or hypotheses, $\mathcal{M}_0$ and $\mathcal{M}_1$, with respective a priori plausibility $p(\mathcal{M}_0)$ and $p(\mathcal{M}_1) = 1 - p(\mathcal{M}_0)$. Differences in prior plausibility are often subjective but can be used to formalize the idea that extraordinary claims require extraordinary evidence (M. D. Lee & Wagenmakers, 2013, Chapter 7). The ratio $p(\mathcal{M}_1)/p(\mathcal{M}_0)$ is known as the prior model odds. The data update the prior model odds to arrive at the posterior model odds, $p(\mathcal{M}_1 \mid \mathbf{D})/p(\mathcal{M}_0 \mid \mathbf{D})$, as follows:

$$\frac{p(\mathcal{M}_1 \mid \mathbf{D})}{p(\mathcal{M}_0 \mid \mathbf{D})} = \frac{p(\mathbf{D} \mid \mathcal{M}_1)}{p(\mathbf{D} \mid \mathcal{M}_0)} \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}, \tag{9.7}$$

or in words:

$$\text{Posterior Model Odds} = \text{Bayes Factor} \times \text{Prior Model Odds.}$$

Equation 9.7 shows that the change in model odds brought about by the data is given by the so-called Bayes factor (Jeffreys, 1961), which is the ratio of marginal likelihoods (i.e., normalizing constants in Equation 9.5):

$$BF_{10} = \frac{p(\mathbf{D} \mid \mathcal{M}_1)}{p(\mathbf{D} \mid \mathcal{M}_0)}. \tag{9.8}$$

The Bayes factor quantifies the weight of evidence for $\mathcal{M}_1$ versus $\mathcal{M}_0$ that is provided by the data and as such it represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648) and "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378).

A $BF_{10} > 1$ indicates that the data are more likely under $\mathcal{M}_1$, and a $BF_{10} < 1$ indicates that the data are more likely under $\mathcal{M}_0$. For example, when $BF_{10} = .08$ the observed data are 12.5 times more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$ (i.e., $BF_{01} = 1/BF_{10} = 1/.08 = 12.5$). Note that the Bayes factor allows researchers to quantify evidence in favor of the null hypothesis.

Even though the default Bayes factor has an unambiguous and continuous scale, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 9.1. We replaced the labels "worth no more than a bare mention" with "anecdotal", "decisive" with "extreme", and "substantial" with "moderate". These labels facilitate scientific communication but should be considered only as an approximate descriptive articulation of different standards of evidence. Under equal prior odds, Bayes factors can be converted to posterior probabilities $p\left(\mathcal{M}_1 \mid D\right) = BF_{10}/\left(BF_{10}+1\right)$. This means that, for example, $BF_{10} = 2$ translates to $p\left(\mathcal{M}_1 \mid D\right) = 2/3$.

Table 9.1 Evidence Categories for the Bayes Factor BF_{10} (Jeffreys, 1961)

.

Bayes factor BF_{10}			Interpretation
	>	100	Extreme evidence for $\mathcal{M}_1$
30	–	100	Very strong evidence for $\mathcal{M}_1$
10	–	30	Strong evidence for $\mathcal{M}_1$
3	–	10	Moderate evidence for $\mathcal{M}_1$
1	–	3	Anecdotal evidence for $\mathcal{M}_1$
	1		No evidence
1/3	–	1	Anecdotal evidence for $\mathcal{M}_0$
1/10	–	1/3	Moderate evidence for $\mathcal{M}_0$
1/30	–	1/10	Strong evidence for $\mathcal{M}_0$
1/100	–	1/30	Very strong evidence for $\mathcal{M}_0$
	<	1/100	Extreme evidence for $\mathcal{M}_0$

Note. We replaced the labels “Not worth more than a bare mention” with “Anecdotal”, “Decisive” with “Extreme”, and “Substantial” with “Moderate”.

9.5 Bayesian Hypothesis Test for Mediation

The Bayesian hypothesis test for mediation contrasts the following two models:

$$\begin{aligned} \mathcal{M}_0 &: \alpha\beta = 0, \\ \mathcal{M}_1 &: \alpha\beta \neq 0. \end{aligned} \tag{9.9}$$

Observe that $\mathcal{M}_1$ entails that both $\alpha \neq 0$ and $\beta \neq 0$, so that BF_{10} can be obtained by combining the evidence for the presence of the two paths. Furthermore, note that in the standardized model, path α equals the correlation r_{XM}, and path β equals the partial correlation $r_{MY|X}$. This means that we can use the existing default Bayesian hypothesis tests for correlation and partial correlation (Wetzels & Wagenmakers, 2012) and combine the evidence for the presence of the separate paths to yield the overall Bayes factor for mediation.

The Default JZS Prior

The construction of good default priors is an active area of research in Bayesian statistics (e.g., Consonni, Forster, & La Rocca, 2013; Overstall & Forster, 2010). Most work in this area has been done in the context of linear regression. It is therefore advantageous to formulate the tests for correlation and partial correlation in terms of linear regression, so that existing developments for the selection of default priors can be brought to bear.

A popular default prior for linear regression is Zellner’s g prior, which includes a normal distribution on the regression coefficients $\boldsymbol{\alpha}$, Jeffreys’ prior on the precision ϕ (i.e., a prior that is invariant under transformation; Jeffreys, 1961), and a uniform prior on the intercept β_0:

$$\begin{aligned} p(\boldsymbol{\alpha} \mid \phi, g, \mathbf{X}) &\sim N(0, \frac{g}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}), \\ p(\phi) &\propto \frac{1}{\phi}, \\ p(\beta_0) &\propto 1, \end{aligned} \tag{9.10}$$

where $\mathbf{X}$ denotes the matrix of predictor variables and the precision ϕ is the inverse of the variance. The coefficient g is a scaling factor and controls the weight of the prior relative to the weight of the data. For example, if $g = 1$, the prior has exactly as much weight as the data, and if $g = 10$, the prior has one tenth of the weight of the data. A popular default choice is $g = n$, the unit information prior, where the prior has as much influence as a single observation (Kass & Wasserman, 1995) and the behavior of the test becomes similar to that of BIC (G. Schwarz, 1978).

However, Liang et al. (2008) showed that the above specification yields a bound on the Bayes factor, even when there is overwhelming information supporting $\mathcal{M}_1$. This "information paradox" can be overcome by assigning the regression coefficients a Cauchy prior instead of a normal prior (Zellner & Siow, 1980). Equivalently, this can be accomplished by assigning g from Equation 9.10 an Inverse-Gamma$(1/2, n/2)$ prior:

$$\begin{aligned} p(\boldsymbol{\alpha} \mid \phi, g, \mathbf{X}) &\sim N(0, \frac{g}{\phi}(\mathbf{X}^T\mathbf{X})^{-1}), \\ p(g) &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{(-3/2)} e^{-n/(2g)}, \\ p(\phi) &\propto \frac{1}{\phi}. \end{aligned} \tag{9.11}$$

The above specification is known as the Jeffreys-Zellner-Siow or JZS prior. The JZS prior was adopted by Wetzels and Wagenmakers (2012) for the default tests of correlation and partial correlation, and the same tests are used here to compute the Bayes factor for mediation. It should be stressed, however, that the framework is general and allows researchers to add substantive knowledge about the topic under study by changing the prior distributions (e.g., Armstrong & Dienes, 2013; Dienes, 2011; Guo et al., 2013). With the JZS tests for correlation and partial correlation in hand, we created the default Bayesian hypothesis test for mediation in three steps as described in the next paragraphs.

Step 1: Evidence for Path α

The first step in the hypothesis test for mediation is to establish the Bayes factor for a correlation between X and M, path α (see Figure 9.1). This test can be formulated as a comparison between two linear models:

$$\begin{aligned} \mathcal{M}_0 &: M = \beta_0 + \epsilon, \\ \mathcal{M}_1 &: M = \beta_0 + \alpha X + \epsilon, \end{aligned} \tag{9.12}$$

where ϵ is the normally distributed error term. The default JZS Bayes factor quantifies the extent to which the data support $\mathcal{M}_1$ with path α versus $\mathcal{M}_0$ without path α, as follows (Wetzels & Wagenmakers, 2012):

$$\begin{aligned} BF_{10} &= BF_{\alpha} \\ &= \frac{P(\mathbf{D} \mid \mathcal{M}_1)}{P(\mathbf{D} \mid \mathcal{M}_0)} \\ &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} \times \int_0^{\infty} (1+g)^{(n-2)/2} \times [1 + (1-r^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} \, \mathrm{d}g, \end{aligned} \tag{9.13}$$

where n is the number of observations and r is the sample correlation.

For the proposed mediation test, we have to multiply the posterior probabilities of paths α and β, as both independent paths need to be present for mediation to hold. Hence we need to convert the Bayes factor for path α to a posterior probability. Under the assumption of equal prior odds this conversion is straightforward:

$$p(\alpha \neq 0 \mid \mathbf{D}) = \frac{BF_{\alpha}}{BF_{\alpha} + 1}. \tag{9.14}$$

Step 2: Evidence for Path β

The second step in the hypothesis test for mediation is to establish the Bayes factor for a unique correlation between M and Y (without any influence from X), path β (see Figure 9.1). Again, this test can be formulated as a comparison between two linear models:

$$\begin{aligned} \mathcal{M}_0 &: Y = \beta_0 + \tau X + \epsilon, \\ \mathcal{M}_1 &: Y = \beta_0 + \tau' X + \beta M + \epsilon, \end{aligned} \tag{9.15}$$

where ϵ is the normally distributed error term. The default JZS Bayes factor quantifies the extent to which the data support $\mathcal{M}_1$ with path β versus $\mathcal{M}_0$ without path β, as in a test for partial correlation (Wetzels & Wagenmakers, 2012):

$$\begin{aligned} BF_{10} &= BF_{\beta} \\ &= \frac{P(\mathbf{D} \mid \mathcal{M}_1)}{P(\mathbf{D} \mid \mathcal{M}_0)} \\ &= \frac{\int_0^{\infty} (1+g)^{(n-1-p_1)/2} \times [1 + (1-r_1^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} \, \mathrm{d}g}{\int_0^{\infty} (1+g)^{(n-1-p_0)/2} \times [1 + (1-r_0^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} \, \mathrm{d}g}, \end{aligned} \tag{9.16}$$

where n is the number of observations, r_1^2 and r_0^2 represent the explained variance of $\mathcal{M}_1$ and $\mathcal{M}_0$, respectively, and $p_1 = 2$ and $p_0 = 1$ are the number of regression coefficients or paths in $\mathcal{M}_1$ and $\mathcal{M}_0$, respectively. As before, we can convert the Bayes factor for β to a posterior probability under the assumption of unit prior odds:

$$p(\beta \neq 0 \mid \mathbf{D}) = \frac{BF_{\beta}}{BF_{\beta} + 1}. \tag{9.17}$$

Step 3: Evidence for Mediation

The third step in the hypothesis test for mediation is to multiply the evidence for α with the evidence for β to obtain the overall evidence for mediation:

$$\text{Evidence for Mediation} = p(\alpha \neq 0 \mid \mathbf{D}) \times p(\beta \neq 0 \mid \mathbf{D}). \tag{9.18}$$

The resulting evidence for mediation is a posterior probability that ranges from zero when there is no evidence for mediation at all, to one when there is absolute certainty that mediation is present. We can also express the evidence for mediation as a Bayes factor through a simple transformation:

$$BF_{med} = \frac{\text{Evidence for Mediation}}{1 - \text{Evidence for Mediation}}, \tag{9.19}$$

where a $BF_{med} > 1$ indicates evidence for mediation, and $BF_{med} < 1$ indicates evidence against mediation.

Testing for Full or Partial Mediation

An optional fourth step in the hypothesis test for mediation is to assess the evidence for full versus partial mediation. The relation between X and Y is fully mediated by M when $\alpha\beta$ differs from zero and the direct path between X and Y, path τ', is zero. The evidence for τ' can be assessed with the JZS test for partial correlation as we did for path β (see Equation 9.16). Note however that the specification of the null model has changed:

$$\begin{aligned} \mathcal{M}_0 &: Y = \beta_0 + \beta M + \epsilon, \\ \mathcal{M}_1 &: Y = \beta_0 + \tau' X + \beta M + \epsilon. \end{aligned} \tag{9.20}$$

With this model specification, the default JZS Bayes factor quantifies the extent to which the data support $\mathcal{M}_1$ with path τ' versus $\mathcal{M}_0$ without path τ'. As before, the resulting JZS Bayes factor for τ' can be converted to a posterior probability:

$$p(\tau' \neq 0 \mid \mathbf{D}) = \frac{BF_{\tau'}}{BF_{\tau'} + 1}. \tag{9.21}$$

Together, the Bayes factor for τ' and the Bayes factor for mediation indicate whether mediation is full or partial: if the Bayes factor for mediation is substantially greater than one and the Bayes factor for τ' is substantially smaller than one, there is evidence for full mediation. On the other hand, if both the Bayes factor for mediation and the Bayes factor for τ' are substantially greater than one, there is evidence for partial mediation.

9.6 Simulation Study

In order to provide an indication of how the mediation test performs, we designed a simulation study. The goal of the simulation study was to confirm that the Bayes factor draws the correct conclusion: when mediation is present we expect BF_{med} to be higher than 1, when mediation is absent we expect BF_{med} to be lower than 1.

Creating the Data Sets

We assessed performance of the test in different scenarios. The parameters α and β could take the values 0, .30, and .70, τ was fixed to zero. We did not vary τ' since it has no influence on the Bayes factor for mediation, which only concerns the effect $\alpha\beta$. Furthermore, we chose four sample sizes: $N = 20, 40, 80$, and 160. The 3×3 parameter values combined with the four sample sizes resulted in 36 different scenarios. For each scenario, we created the corresponding covariance matrix of X, Y, and M, all with a variance of one. This standardization has no bearing on the results as they are scale free. We then used the covariance matrix to generate for each scenario N multivariate normally distributed values for X, M, and Y.

Results

Figure 9.2 shows the natural logarithm of the Bayes factors for mediation in the different scenarios. The different shades of grey of the panels show the strength of the mediation that governed the generated data: the darker the grey, the stronger the mediation. In the scenarios in which there was no mediation ($\alpha = 0$ and/or $\beta = 0$) the Bayes factors indicated moderate to very strong evidence for the null model, depending on the sample size. In the scenario of strong mediation ($\alpha = .7$ and $\beta = .7$) the Bayes factors quickly increase from anecdotal evidence ($N = 20$) to moderate evidence ($N = 40$) and further on to very strong and extreme evidence for mediation. In the scenarios of moderate mediation ($\alpha = .7$ and $\beta = .3$ and vice versa), the Bayes factors start to indicate evidence for mediation from sample sizes of around 60. In the scenario of weak mediation ($\alpha = .3$ and $\beta = .3$) the mediation is too weak for the proposed test to detect it with small sample sizes. In those scenarios the test only starts to indicate evidence for mediation from a sample size of around 80 onward. In summary, the proposed test can distinguish between no mediation and mediation, provided that effect size and sample size are sufficiently large.

Discussion

The results from the simulation study confirm that the JZS Bayesian hypothesis test for mediation performs as advertised: when mediation is absent, the test indicates moderate to strong evidence against mediation; when mediation is present, the test indicates evidence for mediation, provided that effect size and sample size are sufficiently large. As expected, the evidence for mediation increases with effect size and with sample size.

Even though the default test performs well in a qualitative sense, it has one shortcoming that remains to be addressed: with the proposed method it is not possible to perform a one-sided test. This is regrettable, because in many situations the researcher has a clear idea on the direction of the possible paths α, β, and τ'. In order to perform a one-sided Bayesian hypothesis test, the prior need to be restricted such that it assigns mass to only positive (or negative) values. This is not possible in the mediation test as outlined above.

9.7 Extension to One-Sided Tests

As mentioned above, our default prior on a regression coefficient is a Cauchy(0,1) distribution. This prior instantiates a two-sided test, as it represents the belief that the effect is just as likely to be positive than negative. In many situation, however, researchers have strong prior ideas about the direction of the effect (Hoijtink et al., 2008). In the Bayesian framework, such prior ideas are directly reflected in the prior distribution. More specifically, assume we expect path α to be greater

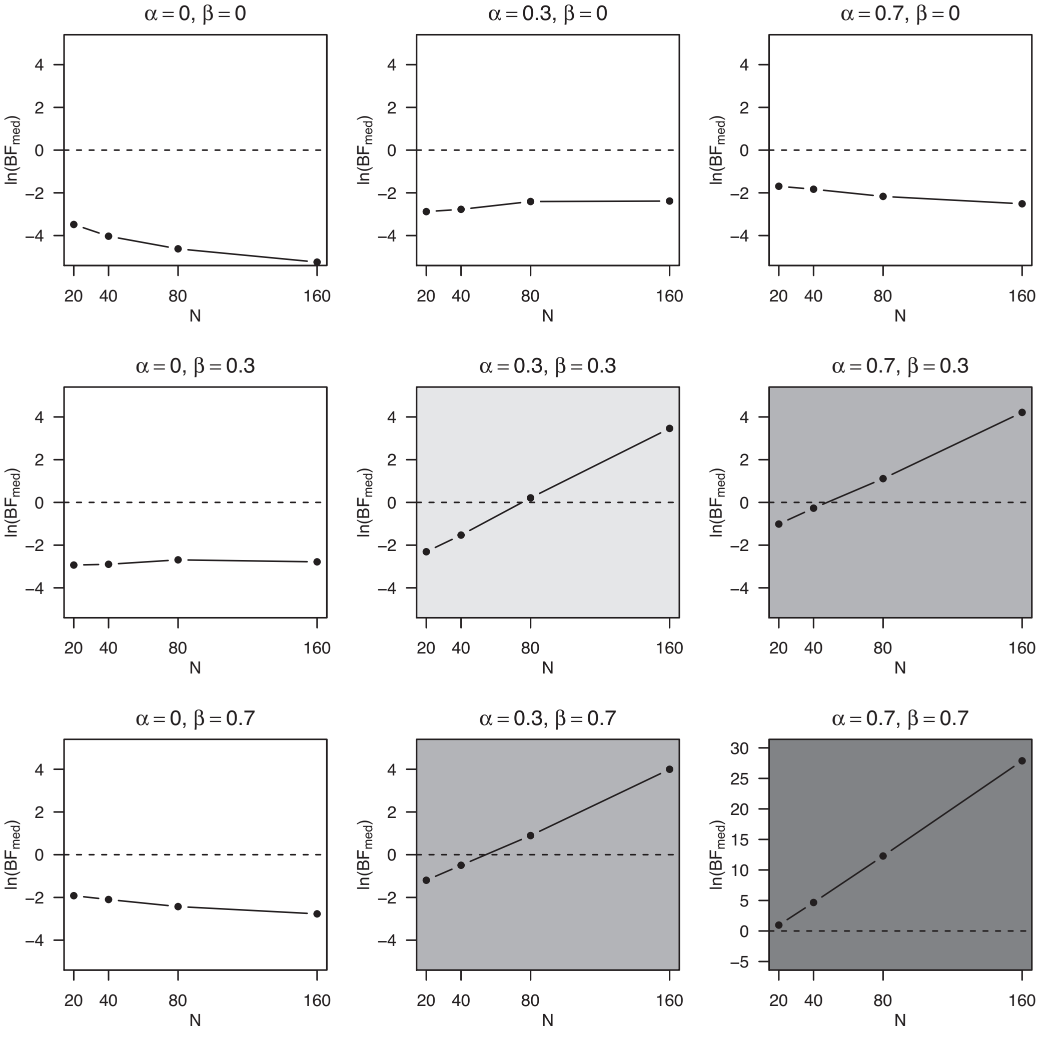


Figure 9.2 Performance of the default JZS Bayesian hypothesis test for mediation in different scenarios. Each panel shows the natural logarithm of the Bayes factor for mediation for different values of α and β and different sample sizes. The white panels correspond to scenarios in which there is no mediation, the grey panels to scenarios in which there is mediation. The darker the panel, the stronger the mediation that is present. The horizontal dotted line at zero indicates the boundary that separates evidence for the null model (below the line) and evidence for the mediation model (above the line). Note that the scaling in the scenario of strong mediation is different from the other scenarios to give a more adequate overview of the results.

than zero and we seek a test of this order-restricted hypothesis against the null hypothesis that α is zero. For this, we consider the following three hypotheses:

$$\begin{aligned} \mathcal{M}_0 &: \alpha = 0, \\ \mathcal{M}_1 &: \alpha \sim \text{Cauchy}(0,1), \\ \mathcal{M}_2 &: \alpha \sim \text{Cauchy}^{+}(0,1), \end{aligned} \tag{9.22}$$

where Cauchy$^{+}(0,1)$ indicates that α can only take values on the positive side of the Cauchy(0,1) distribution (i.e., it is a folded Cauchy distribution).

The test of interest features the comparison between the one-sided hypothesis $\mathcal{M}_2$ versus the null hypothesis $\mathcal{M}_0$, that is, we seek the Bayes factor BF_{20}. This Bayes factor can be derived in many ways, for instance using relatively straightforward techniques such as the Savage-Dickey density ratio (Dickey & Lientz, 1970; Wagenmakers et al., 2010; Wetzels, Grasman, & Wagenmakers, 2010) or relatively intricate techniques such as reversible jump MCMC (Green, 1995). Here we apply a different method that is possibly the most reliable and the least computationally expensive (Pericchi, Liu, & Torres, 2008; Morey & Wagenmakers, 2014). This method takes advantage of the fact that we can easily calculate the two-sided Bayes factor, BF_{10}. With this Bayes factor in hand, we only need to apply a simple correction to derive the desired one-sided Bayes factor BF_{10}. Specifically, note that the Bayes factor is transitive:

$$BF_{20} = BF_{21} \times BF_{10}, \tag{9.23}$$

which is immediately apparent from its expanded form

$$\frac{p(D \mid \mathcal{M}_2)}{p(D \mid \mathcal{M}_0)} = \frac{p(D \mid \mathcal{M}_2)}{p(D \mid \mathcal{M}_1)} \times \frac{p(D \mid \mathcal{M}_1)}{p(D \mid \mathcal{M}_0)}. \tag{9.24}$$

Thus, the desired one-sided test on α requires only BF_{21} and BF_{10}. We already have access to BF_{10}, and this leaves the calculation of BF_{21}, that is, the Bayes factor in favor of the order-restricted model $\mathcal{M}_2$ over the unrestricted model $\mathcal{M}_1$. As was shown by Klugkist et al. (2005), this Bayes factor equals the ratio of two probabilities that can be easily obtained: the first is the posterior probability that $\alpha > 0$, under the unrestricted model $\mathcal{M}_1$; the second is the prior probability that $\alpha > 0$, again under the unrestricted model $\mathcal{M}_1$. Formally:

$$BF_{21} = \frac{p(\alpha > 0 \mid \mathcal{M}_1, \mathbf{D})}{p(\alpha > 0 \mid \mathcal{M}_1)}. \tag{9.25}$$

Since the prior distribution is symmetric around zero, the denominator equals .5 and Equation 9.25 can be further simplified to:

$$BF_{21} = 2 \cdot p(\alpha > 0 \mid \mathcal{M}_1, \mathbf{D}) \tag{9.26}$$

One straightforward way to determine $p(\alpha > 0 \mid \mathcal{M}_1, \mathbf{D})$ is (1) to use a generic program for Bayesian inference such as WinBUGS, JAGS, or Stan; (2) implement $\mathcal{M}_1$ in the program and collect Markov chain Monte Carlo (MCMC) samples from the posterior distribution of α; (3) approximate $p(\alpha > 0 \mid \mathcal{M}_1, \mathbf{D})$ by the proportion of posterior MCMC samples for α that are greater than zero.[2]

In our implementation of the one-sided mediation tests, we make use of Equations 9.23 and 9.26. In order to obtain BF_{21}, we implemented the unrestricted models in JAGS (Plummer, 2009). The JAGS code itself is provided in Appendix F.1, and it allows researchers to adjust the prior

[2] The approximation can be made arbitrarily close by increasing the number of MCMC samples.

distributions if they so desire. We confirmed the correctness of our JAGS implementation by comparing the analytical results for the two-sided Bayes factor BF_{10} against the Savage-Dickey density ratio results based on the MCMC samples from JAGS (see Appendix F.2). Finally, note that our one-sided mediation test can incorporate order-restriction on any of the paths simultaneously.

9.8 Example: The Firefighter Data

To illustrate the workings of the various mediation tests, we will apply them to the same example data Yuan and MacKinnon (2009) used, concerning the PHLAME firefighter study (Elliot et al., 2007). In this study it was investigated whether the effect of a randomized exposure to one of three interventions (X) on the reported eating of fruits and vegetables (Y) was mediated by knowledge of the benefits of eating fruits and vegetables (M; see Equations 9.1, 9.2, and 9.3). The interventions were either a "team-centered peer-led curriculum" or "individual counseling using motivational interviewing techniques", both to promote a healthy lifestyle, or a control condition. The correlation matrix of the data is shown in Table 9.2.

Table 9.2 Correlation Matrix of the PHLAME Firefighter Data. $N = 354$.

	X	Y	M
X	1.00	0.08	0.18
Y	0.08	1.00	0.16
M	0.18	0.16	1.00

The Conventional Approach: The Frequentist Product Method

Yuan and MacKinnon (2009) first reported the results of the conventional frequentist product method mediation analysis (see Table 9.3). This method tests whether the indirect effect $\alpha\beta$ differs significantly from zero. The estimate for $\alpha\beta$ was .056 with a standard error of .026 (estimated with the Sobel method; Sobel, 1982), with the 95% confidence interval $(.013, .116)$ (MacKinnon, Lockwood, & Williams, 2004; the interval takes into account that $\alpha\beta$ is not normally distributed). Since the 95% confidence interval does not include zero, frequentist custom suggests that the test provides evidence that the effect of X on Y is mediated by M.

The Yuan and MacKinnon (2009) Approach: Bayesian Parameter Estimation

Next, Yuan and MacKinnon (2009) reported the results of their Bayesian mediation analysis, which is based on parameter estimation with noninformative priors (see Table 9.3). The mean of the posterior distribution of $\alpha\beta$ was .056 with a standard error of .027. The 95% credible interval for $\alpha\beta$ was $(.011, .118)$. These Bayesian estimates are numerically consistent with the frequentist results. It should be stressed, however, that the 95% confidence interval does not allow a test. As summarized by Berger (2006): "Bayesians cannot test precise hypotheses using confidence intervals. In classical statistics one frequently sees testing done by forming a confidence region for the parameter, and then rejecting a null value of the parameter if it does not lie in the confidence region. This is simply wrong if done in a Bayesian formulation (and if the null value of the parameter is believable as a hypothesis)." (p. 383; see also Lindley, 1957; Wagenmakers & Grünwald, 2006).

Table 9.3 Three Estimates of the Mediated Effect $\hat{\alpha}\hat{\beta}$ for the PHLAME Firefighter Data Set with Associated 95% Confidence/Credible Intervals.

	$\hat{\alpha}\hat{\beta}$	$CI_{95\%}$
Frequentist product method	.056	(.013, .116)
Yuan & MacKinnon (2009)	.056	(.011, .118)
Default Bayesian hypothesis test	.056	(.012, .116)

The Bayes Factor Approach: The Default Bayesian Hypothesis Test

We will now consider the results of the proposed Bayesian hypothesis test with the default JZS prior set-up. First we estimated the posterior distribution of $\alpha\beta$, using the method of Yuan and MacKinnon (2009) but now with the JZS prior instead of a noninformative prior (see Table 9.3). The resulting posterior distribution had a mean of .056 and a 95% credible interval of $(.012, .116)$. This is consistent with the results of both the frequentist test and the Bayesian mediation estimation routine of Yuan and MacKinnon (2009). As expected, the choice of the JZS prior set-up over a noninformative prior set-up does not much influence the results in term of parameter estimation.

The advantage of the JZS prior specification is that we can also formally test whether the effect differs from zero. Our analytical test indicates that the Bayes factor for path α is 10.06, which corresponds to a posterior probability of $10.06/(10.06 + 1) = .91$. The Bayes factor for path β is 2.68, which corresponds to a posterior probability of $2.68/(2.68 + 1) = .73$. If we multiply these posterior probabilities, we obtain the posterior probability for mediation: $.91 \times .73 = .66$. This posterior probability is easily converted to a Bayes factor: $.66/(1 - .66) = 1.94$. Hence, the data are about twice as likely under the model with mediation than under the model without mediation. In terms of Jeffreys' evidence categories this evidence is anecdotal or "not worth more than a bare mention".

It is also possible to include an order-restriction in the mediation model at hand. According to the theory, we expect a positive relation between the mediator "knowledge of the benefits of eating fruits and vegetables" and the dependent variable "the reported eating of fruits and vegetables", or in other words: we expect path β to be greater than zero. If we implement this order-restriction, our test indicates a new Bayes factor for path β of 5.33, with a corresponding posterior probability of $5.33/(5.33 + 1) = .84$. If we multiply the posterior probability of α with the new posterior probability of β, we obtain the new posterior probability of mediation: $.91 \times .84 = .76$, with a corresponding Bayes factor for mediation of $.76/(1 - .76) = 3.17$. With the imposed order restriction, the observed data are now about three times as likely under the mediation model than under the model without mediation, which according to the Jeffreys' evidence categories constitutes evidence for mediation on the border between "anecdotal" and "moderate".

9.9 R package: BayesMed

In order to make our default Bayesian hypothesis tests available, we built the `R` package `BayesMed` (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, in preparation). `R` is a free software environment for statistical computing and graphics (R Core Team, 2012) and can be easily downloaded and installed, which makes it a good platform for our test.

`BayesMed` includes both the basic test for mediation (`jzs_med`) and the accompanying tests for correlation (`jzs_cor`) and partial correlation (`jzs_partcor`), as well as the associated Savage-

Dickey density ratio versions (`jzs_medSD`, `jzs_corSD`, and `jzs_partcorSD`, respectively). Furthermore, we added the possibility to estimate the indirect effect $\alpha\beta$, based on the procedure outlined in Yuan and MacKinnon (2009), but with a JZS prior set-up. Finally, we also included the Firefighter data. The use of the tests and their options are described extensively in the help files within the package. Until the package is available on CRAN, it can be obtained from `https://github.com/MicheleNuijten/BayesMed`.

9.10 Concluding Comments

We have outlined a default Bayesian hypothesis test for mediation and presented an `R` package that allows it to be applied easily. This default test complements the earlier work by Yuan and MacKinnon (2009) on Bayesian estimation for mediation. In addition, we have extended the default tests by allowing more informative, one-sided alternatives to be tested as well. Nevertheless, our test constitutes only a first step. Avenues for further development include, but are not limited to, the following: (1) integrate the estimation and testing approaches by using the estimation outcomes from earlier work as a prior for the later test (Verhagen & Wagenmakers, 2014); (2) explore methods to incorporate substantive prior knowledge (e.g., Dienes, 2011); (3) extend the test to interval null hypotheses, that is, null hypotheses that are not defined by a point mass at zero, but instead by a practically meaningful area around zero (Morey & Rouder, 2011); and (4) generalize the methodology to more complex models such as hierarchical models or mixture models.

As for all Bayesian hypothesis tests that are based on Bayes factors, users need to realize that the test depends on the specification of the alternative hypothesis. In general, it is a good idea to conduct a sensitivity analysis and examine the extent to which the outcomes are qualitatively robust to alternative plausible prior specifications (e.g., Wagenmakers et al., 2011). Such sensitivity analyses are facilitated by our JAGS code presented in Appendix F.1.

In sum, we have provided a default Bayesian hypothesis test for mediation. This test allows users to quantify statistical evidence in favor of both the null hypothesis (i.e., no mediation) and the alternative hypothesis (i.e., full or partial mediation). The test also allows informative hypotheses to be tested in the form of order-restrictions. Several extension of the methodology are possible and await future implementation.

Part IV

Improving Research Practice

Chapter 10

Two Birds with One Stone: A Preregistered Adversarial Collaboration on Horizontal Eye Movements in Free Recall

This chapter has been submitted for publication as:
Dora Matzke, Sander Nieuwenhuis, Hedderik van Rijn, Heleen A. Slagter, Maurits W. van der Molen, and Eric-Jan Wagenmakers (2013).
Two birds with one stone: A preregistered adversarial collaboration on horizontal eye movements in free recall.

Abstract

A growing body of research suggests that horizontal saccadic eye movements facilitate the retrieval of episodic memories in free recall and recognition memory tasks. Nevertheless, a minority of studies have failed to replicate this effect. The present paper attempts to resolve the inconsistent results by introducing a novel variant of proponent-skeptic collaboration. The proposed approach combines the features of adversarial collaboration and purely confirmatory preregistered research. Prior to data collection, the adversaries reached consensus on an optimal research design, formulated their expectations, and agreed to submit the findings to an academic journal regardless of the outcome. To increase transparency and secure the purely confirmatory nature of the investigation, the two parties set up a publicly available adversarial collaboration agreement that detailed the proposed design and all foreseeable aspects of the data analysis. As anticipated by the skeptics, a series of Bayesian hypothesis tests indicated that horizontal eye movements did not improve free recall performance. The skeptics suggest that the non-replication may partly reflect the use of suboptimal and questionable research practices in earlier eye movement studies. The proponents counter this suggestion and use a p-curve analysis to argue that the effect of horizontal eye movements on explicit memory does not merely reflect selective reporting.

10.1 Introduction

Do horizontal saccades make it easier for people to retrieve events from memory? Past research seems to suggest that they do. A growing number of investigations report that only 30 seconds of horizontal saccadic eye movements can improve memory retrieval and boost performance in both recall and recognition tasks. A number of studies have, however, failed to replicate the seemingly well-established effect of horizontal eye movements on free recall performance.

Motivated by the inconsistent results, here we describe a purely confirmatory proponent-skeptic collaboration that focuses on the association between horizontal eye movements and episodic memory. Proponent-skeptic collaboration has been repeatedly advocated as a constructive method of scientific conflict resolution (Hofstee, 1984; Kahneman, 2003; Latham, Erez, & Locke, 1988; Mellers, Hertwig, & Kahneman, 2001). Moreover, we believe that adversarial collaborations —especially when coupled with the preregistration of the statistical analyses— may remedy a number of factors that contributed to the recent crisis of confidence in psychological research and may increase the transparency of scientific communication (see also Koole & Lakens, 2012; Wagenmakers et al., 2011).

10.2 Preregistered Adversarial Collaboration: A Confirmatory Proponent-Skeptic Investigation

Adversarial collaboration is a cooperative research effort that is undertaken by two (groups of) investigators who hold different views on a particular empirical question. The term adversarial collaboration was coined by Kahneman (2003, see also Latham et al., 1988), who —unsatisfied with the inefficiency of traditional reply-rejoinder disputes— urged the scientific community to engage in a "good-faith effort to conduct debates by carrying out joint research" (p. 729). The goal of an adversarial collaboration is to reach consensus on an experimental design and the corresponding testable hypotheses. To facilitate the interpretation of the results, the adversaries are required to formulate and document their expectations about the outcome of the study prior to data collection. Adversarial collaborations are often carried out under the guidance of a third-party researcher, the arbiter, who oversees the collaboration and acts as an impartial referee in case of disagreements (see also Mellers et al., 2001; Nier & Campbell, 2012). Although adversarial collaboration does not necessarily result in the complete resolution of the disagreement, it often leads to new testable hypotheses and is therefore likely to advance the debate.

Although the past two decades have witnessed a number of successful adversarial collaborations in various disciplines (e.g., Bateman, Kahneman, Munro, Starmer, & Sugden, 2005; Cadsby, Croson, Marks, & Maynes, 2008; Gilovich, Medvec, & Kahneman, 1998; Mellers et al., 2001; Schlitz, Wiseman, Watt, & Radin, 2006; Tetlock & Mitchell, 2009; Wiseman & Schlitz, 1997, 1998), this form of conflict resolution is unfortunately still the exception rather than the rule. The lack of adversarial collaboration is especially unfortunate in light of the recent "crisis of confidence" (Pashler & Wagenmakers, 2012, p. 528) in psychological research. The crisis is fueled by concerns about the replicability of key results (e.g., Hunter, 2001) and the widespread use of questionable research practices, such as the selective reporting of significant results (e.g., Simmons, Nelson, & Simonsohn, 2011). The controversy has drawn widespread public attention and triggered a broad range of responses. At one end of the spectrum, failures to replicate key studies in the psychological literature (e.g., Doyen, Klein, Pichon, & Cleeremans, 2012; Shanks et al., 2013) have prompted hostility and finger-pointing between research groups. At the other end of the spectrum, the dispute has given rise to valuable attempts to identify and remedy the factors that contributed to the development of the crisis. Although the proposed recommendations vary considerably in focus, they all emphasize the importance of increasing the transparency of scientific communication (Ioannidis, 2005; Koole & Lakens, 2012; Pashler & Harris, 2012; Wagenmakers et al., 2011, 2012).

Transparency should not only be a concern once the data have been collected; it has been suggested that researchers should commit themselves to the methods of data analysis prior to data collection (e.g., Wagenmakers et al., 2012; de Groot, 1961a, 1961b). Failure to do so may lure researchers into tailoring the analyses to patterns in the observed data in order to find statistically significant results (John, Loewenstein, & Prelec, 2012; Simmons et al., 2011). Fishing for significant results, however, invalidates the interpretation of Type I and Type II error rates and may lead to distorted statistical conclusions. In fact, Wagenmakers et al. (2012) argued that the widespread confusion between exploratory and confirmatory research is the main 'fairy-tale' factor in contemporary psychology. Wagenmakers et al. have therefore urged researchers to preregister their studies and publicly disclose prior to data collection which dependent variables they intend to measure and which statistical analyses they intend to conduct (see also Bakker, van Dijk, & Wicherts, 2012; Chambers, Munafo, & et al., 2013; de Groot, 1961a; Goldacre, 2009; Ioannidis, 2005; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Wagenmakers et al., 2011). The preregistration of experiments has been substantially simplified by the development of web-based research archives and data repositories such as the Open Science Framework (OSF; `http://openscienceframework.org`).

Here we advocate a hybrid variant of scientific conflict resolution that combines the features of adversarial collaboration (Kahneman, 2003) and preregistered confirmatory research (Wagenmakers et al., 2012). The proposed approach may not only assists the constructive resolution of scientific debates, but may also remedy a number of factors that contributed to the present crisis in psychology. We propose the following guidelines for preregistered proponent-skeptic collaborations (see also Mellers et al., 2001, and Hofstee, 1984, for suggestions for adversarial collaborations). First, the adversaries reach consensus on an optimal research design. This precaution eliminates the possibility of later disputes regarding the execution of the study. Second, the two parties formulate their hypotheses and expectations in advance. This precaution decreases the probability of the investigators falling prey to various cognitive biases, such as hindsight bias (i.e., judging an event as more predictable after it has occurred; Roese & Vohs, 2012) and confirmation bias (i.e., favoring information that confirms one's own hypotheses; Nickerson, 1998). Third, the adversaries agree to write a joint article and submit it to an academic journal regardless of the outcome of the study. This precaution may in the long term counteract publication bias and the file drawer problem (Rosenthal, 1979; Greenwald, 1975). Lastly, as the novel but crucial ingredient, the two parties set up an adversarial collaboration agreement. The agreement describes the proposed research design and all foreseeable aspects of the pre-processing and analysis of the data. This precaution secures the purely confirmatory nature of the investigation and increases the transparency of scientific communication.

The remainder of the article describes a joint investigation that focused on the effects of horizontal eye movements on episodic memory. We will first introduce the research area, motivate the reasons for the preregistered adversarial collaboration, and describe the proposed experimental design and the corresponding statistical analyses. We will then describe the methods of the study in more detail and present the results of the investigation. Lastly, the adversaries will present their own perspective on the results as well as on the process of the joint work.

10.3 Horizontal Eye Movements and Episodic Memory

Background and Motivation

Past research suggests that horizontal saccadic eye movements assist the consolidation and retrieval of memories. For instance, bilateral eye movements have been reported to decrease the severity of memory symptoms in eye-movement desensitization and reprocessing (EMDR, Shapiro, 1989), a well-known therapeutic approach for the treatment of post traumatic stress disorder (e.g., C. W. Lee & Cuijpers, 2013). During EMDR, the patient is required to recall the traumatic memory while performing horizontal eye movements. EMDR is argued to change the traumatic (sensory) memory into a more (verbal) declarative memory, while simultaneously reducing the patient's emotional arousal and avoidance.

As a result of the suggested association between eye movements and memory in clinical contexts, the past decades have witnessed a growing number of experimental studies on the effects of horizontal eye movements. Eye movement experiments typically employ either free recall or recognition memory paradigms and require participants to perform 30 seconds of horizontal eye movements immediately prior to the test phase. According to the alternating hemispheric activation hypothesis (Christman, Garvey, Propper, & Phaneuf, 2003; Propper & Christman, 2008), alternating horizontal eye movements result in the alternating activation of the two brain hemispheres. This activation pattern may lead to increased hemispheric communication, which in turn benefits the retrieval of memories. As strongly right-handed individuals show lower interhemispheric interaction than mixed- and left-handed individuals, the benefits of horizontal saccades are typically more pronounced for strongly right-handers (e.g., Brunyé, Mahoney, Augustyn, & Taylor, 2009; Lyle, Logan, & Roediger, 2008; Lyle, Hanaver-Torrez, Hackländer, & Edlin, 2012).

Consistent with the alternating hemispheric activation hypothesis, the majority of eye movement studies report that horizontal eye movements improve episodic memory retrieval compared to no eye movements, especially for strongly right-handed participants (e.g., Brunyé et al., 2009; Christman et al., 2003; Christman, Propper, & Dion, 2004; Lyle et al., 2008; Lyle & Osborn, 2011; Nieuwenhuis et al., 2013; Parker, Buckley, & Dagnall, 2009; Parker & Dagnall, 2007, 2010, 2012; Parker, Relph, & Dagnall, 2008). Similarly, various studies show that horizontal eye movements improve memory performance compared to vertical eye movements (e.g., Brunyé et al., 2009; Christman et al., 2003; Parker et al., 2009; Parker & Dagnall, 2007, 2012; Parker et al., 2008). The literature is, however, not entirely consistent. First, Lyle et al. (2008) reported that vertical eye movements –similar to horizontal eye movements– improve memory retrieval compared to no eye movements. Second, Samara, Elzinga, Slagter, and Nieuwenhuis (2011) found that the beneficial effect of horizontal eye movements was only present for the recall of emotional stimuli.

Motivated in part by the above mentioned inconsistencies, the skeptics (i.e., the first, third, and sixth author) have recently conducted two pilot studies in which they attempted to replicate the beneficial effect of horizontal eye movements on free recall. The skeptics compared the recall of emotional and neutral study words from Samara et al. (2011) after horizontal and vertical eye movements. In the first study, the skeptics tested 19 strongly right-handed participants in a within-subject design and found no difference in recall performance between the two eye movement conditions. In the second study, the skeptics tested 16 strongly right-handed participants in a between-subject design. In line with the first study, no differences were found between the horizontal and vertical eye movement condition. The skeptics were thus unable to replicate the beneficial effect of horizontal eye movements on free recall performance.

In light of the somewhat inconsistent results in the literature and the additional null results obtained in the two pilot studies, the skeptics invited the proponents (i.e., second and fourth author) to participate in the present adversarial collaboration. Prior to data collection, the adversaries appointed an impartial referee (i.e., the fifth author) and set up an adversarial collaboration agreement. The adversarial collaboration agreement was registered at the OSF before a single participant was tested. The preregistration and the agreement can be downloaded at `http://openscienceframework.org/project/LAyZm/`.

Proposed Experiment and Expectations

The proposed experiment was an attempt to establish whether horizontal eye movements improve episodic memory retrieval. The investigation followed a strictly confirmatory design and relied on preregistered statistical analyses. The adversaries agreed that the proposed design best reflected the prototypical experiment in the field, and that the results were potentially the most compelling to both skeptics and proponents.

Participants were presented with a list of neutral study words for a subsequent free recall test. Prior to recall, participants were requested to perform –depending on the experimental condition– either horizontal, or vertical, or no eye movements (i.e., looking at a central fixation point). The type of eye movement was thus manipulated between-subjects. As the effect of eye movement on episodic memory has been reported to be influenced by handedness, we tested only strongly right-handed individuals. The dependent variable of interest was the number of correctly recalled words.

The proponents expected horizontal eye movements to affect recall performance. Specifically, the proponents expected that the number of correctly recalled words (1) was higher in the horizontal than in the no eye movement condition, and (2) was higher in the horizontal than in the vertical eye movement condition. The proponents did not expect the number of correctly recalled words to differ between the vertical and the no eye movement condition. In contrast, the skeptics did not expect horizontal eye movements to affect recall performance. Specifically, the skeptics did not expect the number of correctly recalled words to differ between (1) the horizontal and no eye movement condition, (2) the horizontal and vertical eye movement condition, and (3) the vertical and no eye movement condition.

To demonstrate that the results are not contaminated by unintended peculiarities of the experimental setting, the skeptics and the proponents also attempted to replicate the well-established associative-priming effect using a lexical decision task (e.g., de Groot, 1984, 1987; Neely, 1976, 1977). The associative-priming task required participants to categorize letter strings as words or nonwords. Each target word was preceded by a prime word that was either an associate of the target (e.g., dog-cat) or was unrelated to the target (e.g., uncle-cat). The dependent variable of interest was the mean response time (RT) for correct responses to target words. Typically, mean correct RTs are shorter for target words preceded by related primes than for target words preceded by unrelated primes.

Data Analysis

In adversarial collaborations is it essential to be able to quantify evidence in favor of the null hypothesis. Moreover, it is desirable to collect data until the pattern of results is sufficiently clear. Neither requirement can be accomplished within the framework of frequentist statistics. The present experiment therefore relied on Bayesian hypothesis testing using the Bayes factor (e.g., Berger & Mortera, 1999; Edwards et al., 1963; Jeffreys, 1961; Kass & Raftery, 1995; Rouder et al., 2012, 2009; Wagenmakers, 2007; Wagenmakers et al., 2010, 2011, 2012; Wetzels et al., 2009).

The Bayes factor (BF_{01}) is a Bayesian model selection measure that quantifies the probability of the data under the null hypothesis (H_0) relative to the probability of the data under the alternative hypothesis (H_1).[1] For instance, $BF_{01} = 10$ indicates that the data are 10 times more likely under the null hypothesis than under the alternative hypothesis. Alternatively, $BF_{01} = \frac{1}{10}$ indicates that the data are 10 times more likely under the alternative hypothesis than under the null hypothesis. Within the framework of Bayesian inference, the intention with which the data are collected is irrelevant (Edwards et al., 1963); hence we can monitor the Bayes factor as the data are collected (i.e., sequential hypothesis testing), and may stop testing whenever the evidence is sufficiently compelling.

Accordingly, the adversaries set out to test at least 20 participants in each of the three eye movement conditions and agreed to stop testing whenever the Bayes factor for the horizontal eye movement vs. no eye movement condition comparison reflects 'strong' evidence for the null or the alternative hypothesis (see Jeffreys, 1961, for a classification scheme for the Bayes factor). Specifically, the two parties agreed to stop data collection whenever $BF_{01} > 10$ or $BF_{01} < .1$ for the horizontal vs. no eye movement condition comparison. The adversarial collaboration agreement contains the precise specification of the stopping rule.

Skeptics and proponents agreed to test the three hypotheses using default unpaired Bayesian t tests as specified by Wetzels et al. (2009). This test relies on the default Jeffreys-Zellner-Siow prior setting, the standard choice for model selection in regression models (Liang et al., 2008) and in the t test (Rouder et al., 2009; Wagenmakers et al., 2011, 2012). The test assumes a Cauchy distribution for the effect size under the alternative hypothesis with a location parameter of zero and a scale parameter of one (i.e., $\delta \sim \text{Cauchy}(0, 1)$). The Cauchy distribution resembles a standard normal distribution with relatively fat tails, reflecting lack of knowledge about the effect size in a particular paradigm. The Cauchy distribution has been proposed as an objective prior and results in a conservative test.

As the proponents had specific expectations about the direction of the effects (e.g., better recall in the horizontal than in the no eye movement condition), the adversaries used order-restricted (i.e, one-sided) t tests, resulting in a folded Cauchy distribution for effect size that is defined for positive numbers only (i.e., $\delta \sim \text{Cauchy}(0, 1)^+$). Note that neither party expected differences in recall performance between the vertical and the no eye movement condition. The adversaries nevertheless decided to use a one-sided t test because a few studies in the literature reported that —similar to horizontal eye movements— vertical eye movement may also improve episodic memory (e.g., Lyle et al., 2008). The adversaries tested the presence of the associative-priming effect using a one-sided paired-sample Bayesian t test as specified by Wetzels et al. (2009).

10.4 Methods

The detailed description of the materials and the procedures of the experiment is also available in the adversarial collaboration agreement.

Participants

Participants were recruited from the psychology student pool of the University of Amsterdam. The degree of handedness within this pool of subjects had been assessed with the Edinburgh Handedness Inventory (EHI; Oldfield, 1971) as part of an earlier test battery (i.e., the UvA "testweek").

[1] The subscript 01 in BF_{01} indicates that we compute the probability of the data under H_0 relative to the probability of the data under H_1. In contrast, the subscript 10 would indicate that we compute the probability of the data under H_1 relative to the probability of the data under H_0.

Handedness scores range from −100 (strongly left) to +100 (strongly right) in steps of 5. Individuals with EHI score equal to or above +80 were considered strongly right-handed and were approached to participate in the experiment.

Skeptics and proponents agreed to exclude the data of two participants: one participant was under the influence of drugs, whereas the other participant failed to provide a valid EHI score. The remaining 79 participants (17 male; mean age 21.22 years; mean EHI 95.06) had normal or corrected-to-normal vision, were native speakers of Dutch, and were not diagnosed with dyslexia. Participation was rewarded with course credits or with €10.

Tasks and Stimuli

Free Recall and Eye Movement Task

The study list for the free recall task consisted of a primacy buffer of three words, 72 experimental words, and a recency buffer of three words. The study words were neutral Dutch words that featured in Zeelenberg, Wagenmakers, and Rotteveel (2006).[2] Before the presentation of the first word, a fixation cross appeared in the middle of the screen for 3000 ms. The study words were then presented sequentially in black using lower-case 34 point Arial in the middle of a light-gray display for 2000 ms, with an inter-stimulus interval of 500 ms. The order of word presentation was randomized across participants.

The computerized eye movement task started with a central fixation cross presented against a light-gray display for 3000 ms. In the horizontal and vertical eye movement conditions, participants were instructed to follow a black circle with a diameter of approximately 4° visual angle with their eyes. The circle alternated between the left and right (horizontal eye movements) or between the top and bottom (vertical eye movements) portion of the display for 30 sec. As the circle changed position every 500 ms, participants performed two saccadic eye movements per second. The distance between the left and right position of the circle was the same as the distance between the top and bottom position, namely 27°. In the no eye movement condition, a colored circle was presented at the center of the display. The circle changed color every 500 ms, alternating between blue and red.

Associative-Priming Task

The stimulus pool consisted of 64 prime-word pairs and 64 prime-nonword pairs. The primes and the word targets were Dutch nouns, while the nonwords were pseudowords derived from Dutch nouns by changing one or two letters. The nonwords were generated using the Wuggy software (Keuleers & Brysbaert, 2010). In all prime-word pairs, the target word appeared as an associate of the prime in the Dutch word association norms (de Groot, 1980). The prime-word pairs were adopted from de Groot (1984, 1987). The primes for the prime-nonword pairs were unrelated to the prime-word pairs and to the nouns that were used to create the nonwords. One subset of the prime-nonword pairs was adopted from de Groot (1984), whereas the other subset was selected uniquely for the purpose of the present experiment.

The stimulus pool was used to create two lists that each contained 32 related prime-word pairs and 32 unrelated prime-word pairs. The unrelated word pairs were created by rearranging the primes and the word targets of 32 of the 64 related prime-word pairs. Each target word thus appeared in both lists, either as a target in a related prime-word pair or as a target in an unrelated prime-word pair. The length and frequency of the target words were equated across the related

[2] The stimulus words are available from the adversarial collaboration agreement.

and unrelated prime-word pairs in both lists. The associative strength of the related prime-word pairs was equated across the two lists. The same prime-nonword pairs were used across the two lists. Word length was equated across nonwords and the target words in the prime-word pairs.[3]

The two stimulus lists were counterbalanced across participants. The prime-word pairs and the prime-nonword pairs were presented sequentially on a computer screen. The order of stimulus presentation was randomized across participants. The stimuli were presented in black using lowercase 34 point Arial in the middle of a light-gray display. First, a fixation cross appeared on the screen for 1000 ms slightly above and left of the position of the to-be-presented prime, followed by a blank inter-stimulus interval of 20 ms. Next, the prime appeared in the middle of the screen for 400 ms, followed by a blank inter-stimulus interval of 40 ms. Next, the target appeared slightly below the position of the previously presented prime. The target remained on screen until the participant responded or until 2400 ms elapsed. Participants were instructed to press 'M' with their right index finger for 'word' responses and to press 'Z' with their left index finger for 'nonword' responses. Incorrect responses were followed by the message 'FOUT' (i.e., incorrect), responses slower than 1200 ms were followed by the message 'TE LANGZAAM' (i.e., too slow), and responses faster than 200 ms were followed by the message 'TE SNEL' (i.e., too fast). If the participant failed to respond within 2400 ms, 'TE LANGZAAM' appeared automatically on the screen and an error was recorded. The feedback was presented 20 ms after response/target offset, slightly below the position of the previous target. The feedback remained on the screen for 1200 ms. The feedback scheme was intended to promote accurate but fast responding. Following 1000 ms after a correct response or after the offset of an error message, the fixation cross reappeared on the screen.

The experimental stimuli were presented in four blocks of 32 prime-target pairs. A forced rest of 30 sec. separated the experimental blocks. The presentation of the 128 experimental prime-target pairs was preceded by a practice list of 32 prime-target pairs. The practice list consisted of 8 related prime-word pairs, 8 unrelated prime-word pairs, and 16 prime-nonword pairs, none of which were also present in the 128 experimental word-target pairs.

Procedure

Participants were tested individually. Participants were seated behind the computer screen and were given an explanation of the tasks. For the free recall test, participants were explicitly instructed to memorize the presented words for a subsequent memory test. For the eye movement task, participants were instructed to follow the circle with their eyes by making saccadic eye movements and to keep their head still. The experimenter carefully monitored participants' compliance with the instructions, including the eye movement behavior.

Participants were randomly assigned to the three eye movement conditions based on the order of arrival (i.e., Participant 1 was assigned to the horizontal eye movement condition, Participant 2 to the vertical eye movement condition, Participant 3 to the no eye movement condition, Participant 4 to the horizontal eye movement condition, etc.). Participants were then presented with the study list and performed —depending on the eye movement condition— horizontal, vertical, or no eye movements. Next, participants performed a 5-minute paper-and-pencil free recall test.

After a 10-minute break, participants carried out the associative-priming task. Instructions emphasized fast but accurate responding. Participants were instructed to pay attention to both letter strings (i.e., prime and target), but only respond to the second letter string (i.e., the target). The instructions did not mention the association between the related prime-word pairs. Lastly, participants completed an exit interview, inquiring about their age and gender. In addition, par-

[3] The associative-priming stimuli are available from the adversarial collaboration agreement.

ticipants were asked to indicate whether they were aware of the goal of the experiment, and if so, they were asked to describe what they thought the goal was.

10.5 Results

Confirmatory Analyses

Eye Movement Task

The free recall data are available on the OSF at `http://openscienceframework.org/project/pXT3M/`. Based on the exclusion criteria specified in the adversarial collaboration agreement, we excluded the free recall data of two participants who correctly described the goal of the experiment and four participants who recalled fewer than five items correctly. The analyses reported below are based on the data of 25 participants in the horizontal eye movement ($N_H = 25$), 24 participants in the vertical eye movement ($N_V = 24$), and 24 participants in the no eye movement condition ($N_F = 24$).

The left panel of Figure 10.1 shows the average number of correctly recalled experimental words in the three eye movement conditions; on average, participants in the horizontal eye movement condition recalled the fewest words and participants in the no eye movement condition recalled the most words. The average number of correctly recalled words was 10.88 (4.32) in the horizontal, 12.96 (5.89) in the vertical, and 15.29 (6.38) in the no eye movement condition. The right panel of Figure 10.1 shows the posterior distribution of each of the effect sizes. In Bayesian inference, the posterior distribution quantifies the uncertainty about an estimated parameter (i.e., effect size) conditional on the evidence provided by the data. The posterior distributions assign most mass to negative effect sizes. Thus, consistent with the observed data, the posterior distributions for the effect sizes indicate that participants recalled fewer words in the horizontal eye movement condition than either in the vertical or the no eye movement condition and that participants recalled fewer words in the vertical than in the no eye movement condition. Effect size is the largest for the horizontal vs. no eye movement comparison. The horizontal vs. vertical and the vertical vs. no eye movement comparisons resulted in smaller and nearly identical effect size estimates.

As Bayesian inference allows for sequential hypothesis testing, we computed the Bayes factor after each triad of participants. Figure 10.2 shows the results of the sequential analyses using one-sided unpaired Bayesian t tests under the assumption of equal variances. The sequential analysis plots show the log Bayes factor as a function of the number of participants per condition; log Bayes factors smaller than zero indicate evidence for the alternative hypothesis, whereas log Bayes factors higher than zero indicate evidence for the null hypothesis.

For all three hypotheses, the evidence in favor of the null hypothesis gradually increased as the data accumulated. After testing 73 participants, the Bayes factor indicated that the data are 15 times more likely under the null hypothesis of no difference between the horizontal and the no eye movement condition than under the alternative hypothesis ($BF_{01} = 15.39$).[4] Similarly, the Bayes factor indicated that the data are more than 10 times more likely under the null hypothesis of no difference between the horizontal and the vertical eye movement condition than under the alternative hypothesis ($BF_{01} = 10.12$). Lastly, the Bayes factor indicated that the data are more

[4] After five weeks of data collection, the BF_{01} was above 10 for the horizontal eye movements vs. no eye movement comparison. The adversaries, however, agreed to continue testing for an additional week in order to obtain compelling evidence also for the horizontal vs. vertical eye movements and the vertical vs. no eye movement comparisons. For the amendment to the adversarial collaboration agreement that documents this decision, see the OSF at `http://openscienceframework.org/project/pXT3M/`

than 9 times more likely under the null hypothesis of no difference between the vertical and the no eye movement condition than under the alternative hypothesis ($BF_{01} = 9.64$). As shown in the right panels of Figure 10.2, essentially the same results were obtained under the assumption of unequal variances. Unsurprisingly, the frequentist alternatives of the one-sided unpaired Bayesian t tests yielded non-significant results: $t(47) = -2.85$, $p > .99$ for the horizontal vs. no eye movement comparison, $t(47) = -1.41$, $p = .92$ for the horizontal vs. vertical comparison, and $t(46) = -1.32$, $p = .90$ for the vertical vs. no eye movement comparison, assuming equality of variances.

In sum, as anticipated by the skeptics, the Bayes factor indicated strong evidence in favor of the null hypothesis for the horizontal vs. no eye movement as well as the horizontal vs. vertical eye movement comparisons. As expected by both parties, the Bayes factor indicated substantial evidence in favor of the null hypothesis for the vertical vs. no eye movement comparisons.

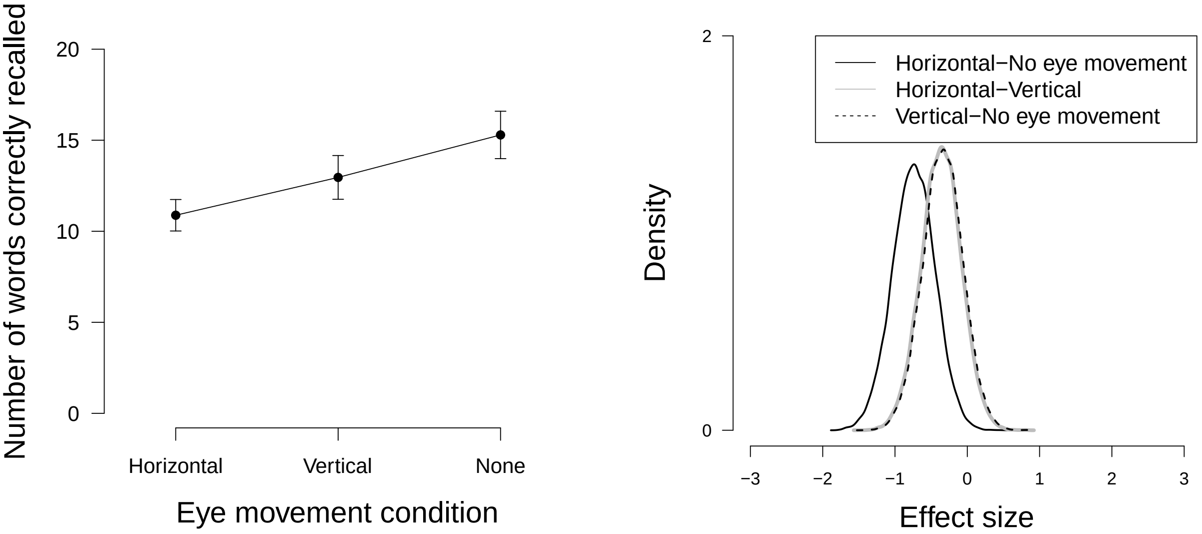


Figure 10.1 *Mean number of words recalled correctly and effect sizes in the three eye movement conditions.* The left panel shows the average number of experimental words recalled correctly in the three eye movement conditions. The error bars indicate the standard error. The right panel shows the posterior distribution of the estimated effect size for the horizontal−no eye movement comparison (solid black line), for the horizontal−vertical eye movement comparison (solid grey line), and for the vertical−no eye movement comparison (dashed line).

Associative-Priming Task

The priming data are available on the OSF at `http://openscienceframework.org/project/pXT3M/`. We used only correct RTs that were longer than 250 ms and shorter than 1500 ms, resulting in an average exclusion rate of 6.39%. Based on the exclusion criteria specified in the adversarial collaboration agreement, we excluded one participant with error rate higher than 20%. We excluded the data of one additional participant because of computer failure. The analysis reported below is based on 77 participants.

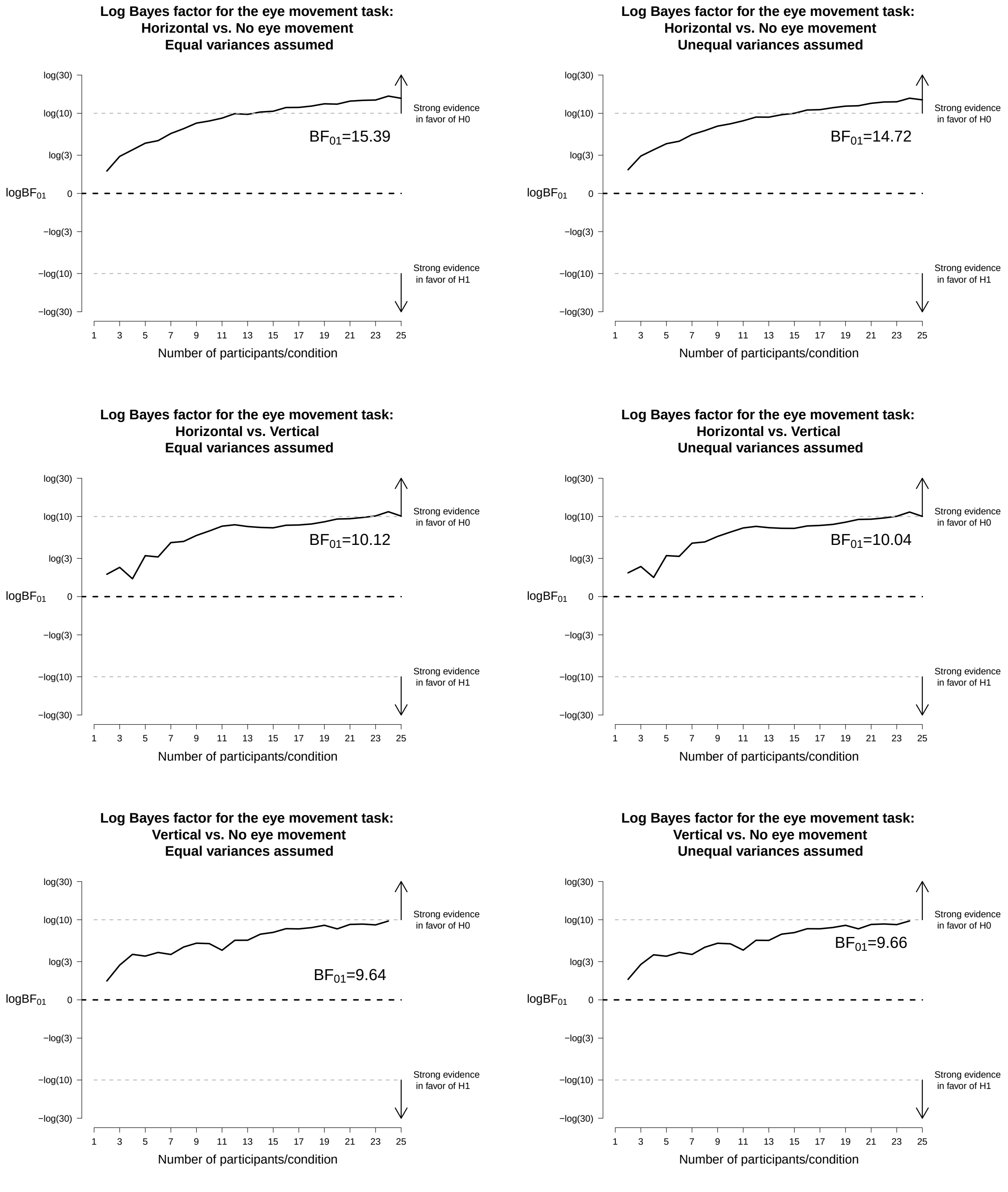


Figure 10.2 *Log Bayes factors for the comparison of the number of correctly recalled words between the horizontal, vertical, and no eye movement conditions.*

Figure 10.3 shows mean RT for the related and the unrelated prime-word pairs and the corresponding effect size. As expected, mean RTs for target words preceded by related primes (493.96 ms, sd = 66.44) were shorter than mean RTs for target words preceded by unrelated primes (527.06, sd = 66.35). The posterior distribution assigns most mass to large negative effect sizes. Figure 10.4 shows the results of the sequential analysis using Bayes factors from the default one-sided paired-sample Bayesian t test. As the data accumulated, the evidence for the alternative hypothesis gradually increased. After testing 77 participants, the Bayes factor indicated that the data are 528,848,417 times more likely under the alternative hypothesis than under the null hypothesis ($BF_{01} = 1.890901E - 09$). This result supports the adversaries' expectation and indicates extreme evidence for the presence of the associative-priming effect.

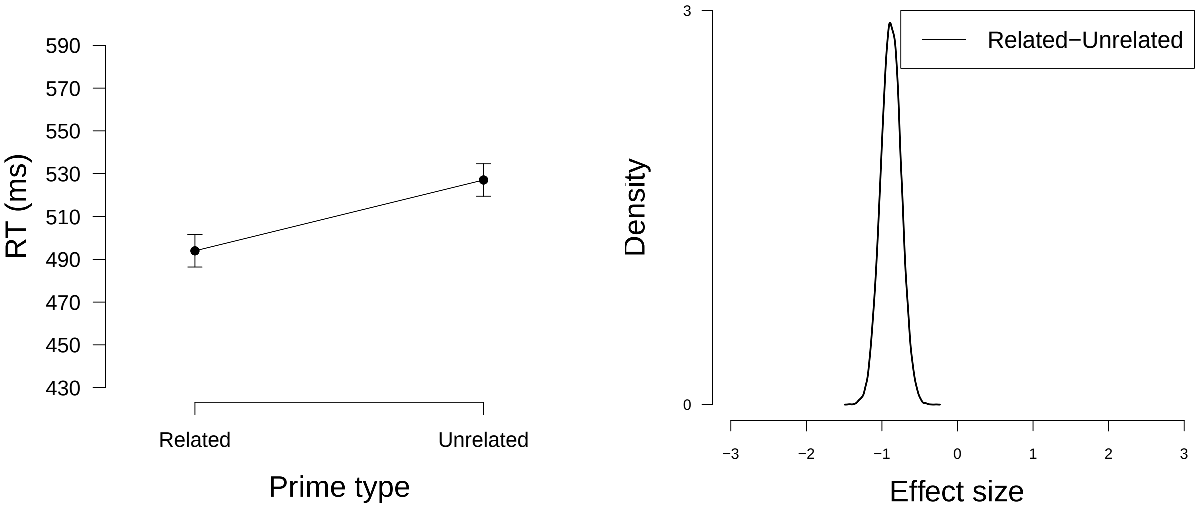


Figure 10.3 *Mean RT and effect size for the associative-priming task.* The left panel shows mean RT for the related and the unrelated prime-word pairs. The error bars indicate the standard error. The right panel shows the posterior distribution of the estimated effect size for the related−unrelated prime-word comparison.

Exploratory Analyses

This section presents the results of a series of analyses aimed at exploring the robustness of the conclusions with respect to the prior setting used for the analysis of the eye movement data. In order to minimize the role of subjective expectations, the confirmatory analyses assumed the default Cauchy$(0, 1)^+$ prior for effect size. The choice of the Cauchy prior may nevertheless be disputed; we might just as well have used a prior that is informed by the eye-movement literature or a prior that assumes smaller variability in effect size than the default Cauchy distribution. Especially the latter possibility warrants further investigation as Bayes factors are sensitive to the shape of the prior distribution (e.g., Bartlett, 1957; Liu & Aitkin, 2008; Vanpaemel, 2010). Specifically, wide prior distributions define highly complex models (i.e., models that can generate a wide range of predictions), resulting in Bayes factors that support the null hypothesis. Thus, highly uninformative prior distributions yield Bayes factors that lend infinite support for the null hypothesis (Jeffreys, 1961).

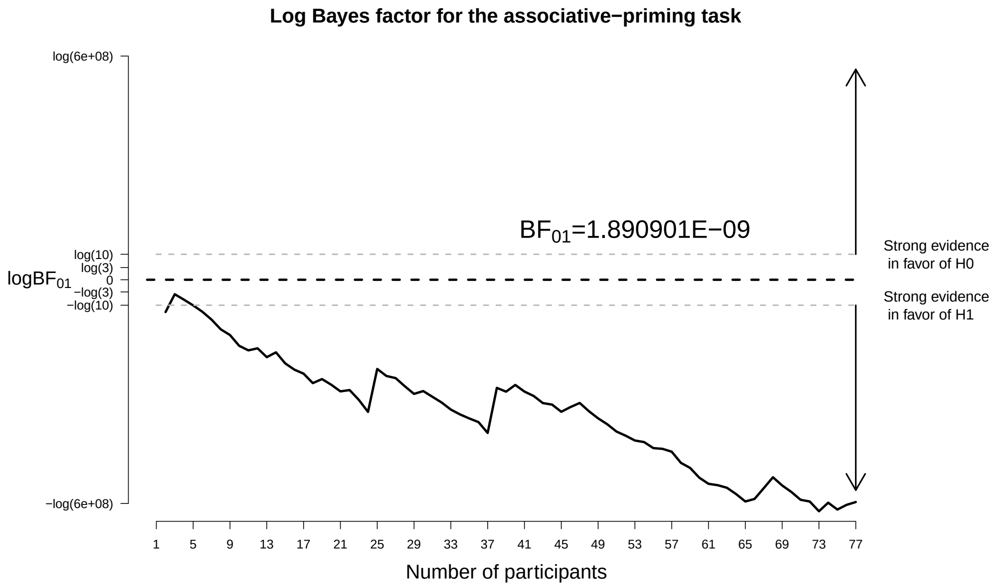


Figure 10.4 *Log Bayes factor for the comparison of mean RT for related vs. unrelated prime-word pairs.*

Here we investigate the extent to which the variability of the prior distribution of effect size influences the Bayes factor. We replaced the Cauchy prior on effect size with a zero centered normal prior and varied the standard deviation (sd) from 0 to 2, creating progressively more spread out —uninformative— priors. As we are concerned with one-sided tests, we used a normal prior that is defined for positive numbers only (i.e., $\delta \sim \text{Normal}(0, \text{sd})^{+}$).

Figure 10.5 shows changes in the log Bayes factor as a function of the standard deviation of the normal prior on effect size. The black triangle corresponds to the Bayes factor computed with the standard normal prior —the so-called unit information prior— on effect size (i.e., $\delta \sim \text{Normal}(0, 1)$). As before, log Bayes factors smaller than zero indicate evidence for the alternative hypothesis, whereas log Bayes factors higher than zero indicate evidence for the null hypothesis. Two aspects of the results are noteworthy. First, as the standard deviation of the normal prior increases (i.e., prior becomes progressively wider), the Bayes factor increasingly favors the null hypothesis. As mentioned above, this result reflects a typical aspect of Bayesian hypothesis testing. Second, the log Bayes factor is never smaller than zero. This result indicates that the Bayes factor never favors the alternative hypothesis over the null hypothesis regardless of the variability of the prior distribution. Even under the prior setting that maximally supports the alternative hypothesis (i.e., standard deviation very close to zero), the log Bayes factor is only around 0, indicating perfectly ambiguous evidence. This finding is not surprising; mean recall was highest in the no eye movement condition and lowest in the horizontal eye movement condition, a result that contradicts the order-restriction specified for the one-sided t test.

The results of the robustness analyses indicated that the Bayes factor, as expected, varied as a function of the standard deviation of the prior distribution of the effect size. Although the strength of the support for the null hypothesis varied as a function of the prior setting, the Bayes factor always favored the null hypothesis over the alternative hypothesis regardless of the variability of the prior.

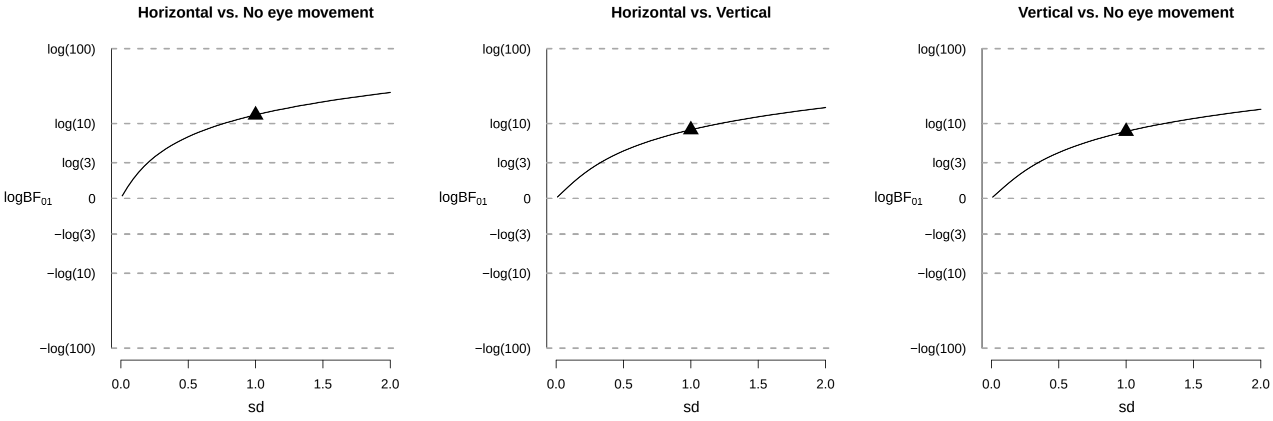


Figure 10.5 *Log Bayes factors (log BF_{01}) as a function of the standard deviation (sd) of the zero-centered normal prior on effect size.* Equal variances are assumed. The black triangle corresponds to the Bayes factor computed with a standard normal prior (i.e., unit-information prior) on effect size.

10.6 Discussion

Adversarial collaboration has been repeatedly advocated as a constructive method of scientific conflict resolution (Hofstee, 1984; Kahneman, 2003; Latham et al., 1988; Mellers et al., 2001). We believe that adversarial collaborations —especially when coupled with preregistration— may also remedy a number of factors that contributed to the crisis of confidence in psychological science and increase the transparency of scientific communication (see also Koole & Lakens, 2012; Wagenmakers et al., 2011). The present paper therefore introduced the notion of preregistered adversarial collaboration, a novel variant of scientific conflict resolution. The proposed approach combines the features of adversarial collaboration and purely confirmatory research (Wagenmakers et al., 2012).

We illustrated the use of preregistered adversarial collaboration with a joint proponent-skeptic investigation on the effect of horizontal eye movements on episodic memory performance. The rules of the collaboration were as follows. First, the adversaries reached consensus on an optimal research design. Specifically, the adversaries agreed to manipulate the type of eye movement between subjects: Participants were requested to perform either horizontal, or vertical, or no eye movements prior to the recall of the study list. Second, the two parties formulated their expectations and agreed to submit the findings to an academic journal whether or not those expectations are supported by the data. Third, the adversaries appointed an impartial referee whose task was to oversee the collaboration. Lastly, but importantly, the two parties set up a publicly available adversarial collaboration agreement that described the proposed design and all foreseeable aspects of the data analysis. The adversarial collaboration agreement was registered at the OSF before a single participant was tested. The adversarial collaboration agreement presented here may serve as a blueprint for future work.

As expected by the skeptics, the Bayes factor indicated strong evidence in favor of the null hypothesis for the horizontal eye movement vs. no eye movement as well as for the horizontal eye movement vs. vertical eye movement comparisons. As expected by both parties, the Bayes factor indicated substantial evidence in favor of the null hypothesis for the vertical eye movement vs.

no eye movement comparison. Lastly, the results of the associative-priming task supported the adversaries' expectation and indicated extreme evidence for the presence of an associative-priming effect. In what follows, the skeptics and the proponents will present their own perspectives on the results of the experiment and the process of the joint research effort.

Discussion by Skeptics

Reflection on the Results

The results clearly supported our expectations: Horizontal eye movements did not improve free recall performance in the present experiment. Our joint study thus failed to replicate the beneficial effect of bilateral eye movements on episodic memory. Despite our best efforts to carry out a prototypical experiment, the present study —and our two pilot studies— contradicts the seemingly well-established finding on the association between horizontal eye movements and memory retrieval.

Our failure to replicate may, of course, simply be due to chance; even if the effect under scrutiny truly exists, a certain number of replication attempts are necessarily doomed to be unsuccessful (e.g., Francis, 201s). Note, however, that our two pilot studies also yielded null results. We propose therefore that the conflicting findings may reflect mechanisms that are related to (1) statistical problems in the literature; (2) prevailing research practices in psychology; and (3) methodological shortcomings of the prototypical research design.

On the statistical side, we believe that the effect of horizontal eye movements on episodic memory may be overestimated as a result of the statistical problems associated with p value-based null hypothesis testing. A well-known problem of frequentist hypothesis testing is that p values overstate evidence against the null hypothesis (Berger & Delampady, 1987; Edwards et al., 1963; V. E. Johnson, 2013; Sellke et al., 2001). Wetzels et al. (2011) showed that 70% of the p values from t tests in experimental psychology that fall between .01 and .05 correspond to Bayes factors that indicate that the data are no more than three times more likely under the alternative hypothesis than under the null hypothesis. This suggests that a number of "significant" findings in the eye movement literature (e.g., Brunyé et al., 2009; Lyle et al., 2008; Samara et al., 2011) may in fact reflect negligible effects that are "not worth more than a bare mention" (Jeffreys, 1961). The present paper therefore advocates the use of Bayesian hypothesis testing with default Bayes factors.

Although it is likely that the eye movement literature is biased by the statistical peculiarities of p values, the results of the present experiment cannot be explained purely in terms of differences in statistical framework. The Bayesian conclusions were corroborated with the results of p value-based hypothesis tests. In fact, participants in the horizontal eye movement condition recalled on average the fewest words, a result that contradicts most —if not all— reported findings in the eye movement literature.

We therefore argue that the conflicting results may partly reflect bias and the use of questionable research practices, both of which can distort the literature. That is, the beneficial effect of horizontal eye movements on free recall may seem more established than it actually is, due to publication bias and the file-drawer problem (Rosenthal, 1979; Greenwald, 1975). Error mechanisms during the interpretation of the data, such as hindsight bias and positive confirmation bias, may likewise contribute to the unbalanced literature by fueling the use of questionable research practices (QRP). QRPs may include optional stopping (i.e., collecting data until the p value reaches a desired significance criterion), selectively reporting results from experimental conditions and dependent variables that produce significant effects, hypothesizing after the results are known (HARKing), and the use of post-hoc exclusion criteria, such as arbitrary handedness cut-off scores.[5]

[5] The following investigations all used different criteria for classifying participants as strongly right-handed:

The present paper therefore emphasizes the importance of preregistration and the strict separation of confirmatory and exploratory research (see also de Groot, 1961a).

Lastly, on the methodological side, we argue that limitations of the prototypical research design may contribute to the conflicting findings. In the present study, as in most eye movement studies, the experimenter was not blind to participants' eye movement condition. The expectations of the experimenter may have unintentionally influenced the outcome of the study by, say, selectively increasing participants' motivation in a given eye movement condition (Rosenthal, 1976). In the current study, the data were collected by the skeptics. Despite our best efforts, our expectations might have been subtly communicated to the participants and have contributed to the null finding in the present experiment and in our two pilot studies. The possibility of the experimenter effect as an explanation for our results warrants further investigation. Note however that if the experimenter's expectation can indeed eliminate or even reverse the effect of bilateral eye movement on free recall, the phenomenon is more fragile than suggested by the literature, a possibility that may explain the present failure to replicate.

Reflection on the Process

Preregistered adversarial collaboration is a labor-intensive undertaking that requires more planning and anticipation than carrying out standard research. Prior to data collection, the adversaries are required to reach consensus on an experimental design and have to anticipate and document —as far as possible— all foreseeable aspects of the data collection and the data analysis. We believe, however, that the advantages of the proposed approach outweigh the disadvantages, as the initial effort involved in setting up the joint research pays off in numerous ways. By critically evaluating and attempting to anticipate all aspects of the research effort, the two parties capitalize on expert knowledge and maximize the probability that the proposed experiment resolves the disagreement. Moreover, the public disclosure of the the experimental procedures and statistical analyses secures the purely confirmatory nature of the research and increases the transparency of the investigation.

Note that preregistration of the proposed experiment does not mean that all aspects of the research effort are carved in stone. If both parties agree, the adversarial collaboration agreement may be amended to account for unexpected events during data collection. For instance, as documented in the present adversarial collaboration, we agreed to modify the stopping rule and our strategy for participant recruitment during data collection (see amendment to the adversarial collaboration agreement on the OSF and footnote 5). Similarly, preregistration of the data analysis does not mean that investigators cannot follow up interesting patterns in the data or —as demonstrated here— investigate the robustness of the conclusions. We believe that exploratory research plays an essential role in science; it generates new testable hypotheses and facilitates scientific progress. We also believe, however, that researchers should explicitly acknowledge which results are based on explorations and which results are based on strictly confirmatory analyses.

In sum, setting up preregistered joint research requires more effort on behalf of the investigators than carrying out standard research. We believe, however, that the additional work is a small price to pay for the possibility of constructive conflict resolution and a great increase in transparency. We hope that preregistered adversarial collaboration —or some other variant of confirmatory joint research— will in the near future become the rule rather than the exception for settling scientific disputes in psychology. In light of the rather heated debates in our discipline, there is certainly room for improvement.

Brunyé et al. (2009) used EHI $>$ median, Christman et al. (2004, Experiment 1) used EHI $\geq$ median, Christman et al. (2004, Experiment 2) used EHI ≥ 75, and Lyle and Osborn (2011) used EHI ≥ 80.

Discussion by Proponents

Reflection on the Results

We were surprised by these results. In a previous study, we found a beneficial effect of horizontal eye movements on recall of emotional words but not neutral words (Samara et al., 2011). However, the null effect for neutral words may have been due to the small sample size ($N = 14$) and/or the relative long period between the horizontal eye movements and subsequent recall test due to an intermittent baseline EEG recording; in a subsequent study, using a much larger sample and no intermittent EEG recording, we did replicate the effect (Nieuwenhuis et al., 2013, Experiment 1). In additional experiments we found a similar beneficial effect on word recall of alternating (vs. simultaneous) left-right tactile but not auditory stimulation, a pattern of results predicted by the alternating hemispheric activation hypothesis (Christman et al., 2003; Propper & Christman, 2008). These and other studies (Propper & Christman, 2008) used procedures and stimulus material that were similar to those used in the current study. In addition, the current study only included consistently right-handed individuals as the effect of horizontal eye movements on memory is present in strong left- and right-handers but not in mixed-handers (Lyle et al., 2008, 2012). It is thus surprising that in the current study, previously reported positive effects of horizontal eye movements on memory performance were not replicated.

So how can we account for the current non-replication? As the skeptics suggest, the non-replication might be a false negative. Or it may be due to experimenter bias (Rosenthal & Rubin, 1978). To rule out this latter possibility, experimenters in future studies will have to be blind to the condition to which a participant is assigned. Here, we consider in more detail another explanation offered by the skeptics: the possibility that researchers selectively report positive studies or analyses, or use any of several questionable strategies (e.g., optional stopping; try different contrasts) for producing a significant effect of horizontal eye movements. To investigate this possibility we conducted a p-curve analysis (Simonsohn, Nelson, & Simmons, 2014). That is, we plotted the distribution of statistically significant p values ($< .05$) reported in studies on the beneficial effects of horizontal eye movements on memory and examined the form of the distribution. As Simonsohn and colleagues argue, "only right-skewed p-curves, those with more low (e.g., .01s) than high (e.g., .04s) significant p values, are diagnostic of evidential value. P-curves that are not right-skewed suggest that the set of findings lacks evidential value, and p-curves that are left-skewed suggest the presence of intense p-hacking" (i.e. obtaining statistically significant results using QRPs).

For this analysis, we selected all studies that examined the effects of 30 seconds of horizontal eye movements (relative to a control condition) on explicit memory in consistently-handed healthy individuals. The steps involved in the selection of p values that meet these selection criteria are documented in the recommended p-curve disclosure table (cf. Simonsohn et al., 2014) available as supplemental material at `http://dora.erbe-matzke.com/publications.html`. Figure 10.6 shows the results of the p-curve analysis based on these p values. As can be seen in this figure, the p-curve is significantly right-skewed, $\chi^2(36) = 102.33$, $p < .0001$, indicating that these studies do contain evidential value. This means that we can rule out p-hacking as the sole explanation for the reported effects of horizontal eye movements. As Simonsohn and colleagues show, with a sample size of ~ 20 p values, it is virtually impossible for p-curve analysis to indicate that the sample contains evidential value when in fact the studies were intensely p-hacked. Nevertheless, it is worth noting that there is an uptick in the p-curve at .05 (test for left skew: $\chi^2(36) = 28.23$, $p = .82$). A p-curve is markedly right-skewed when an effect is real but only mildly left-skewed when a finding is p-hacked. So Simonsohn and colleagues acknowledge that if a set of findings combines true effects with nonexistent ones, the p-curve will usually not detect the latter. Thus,

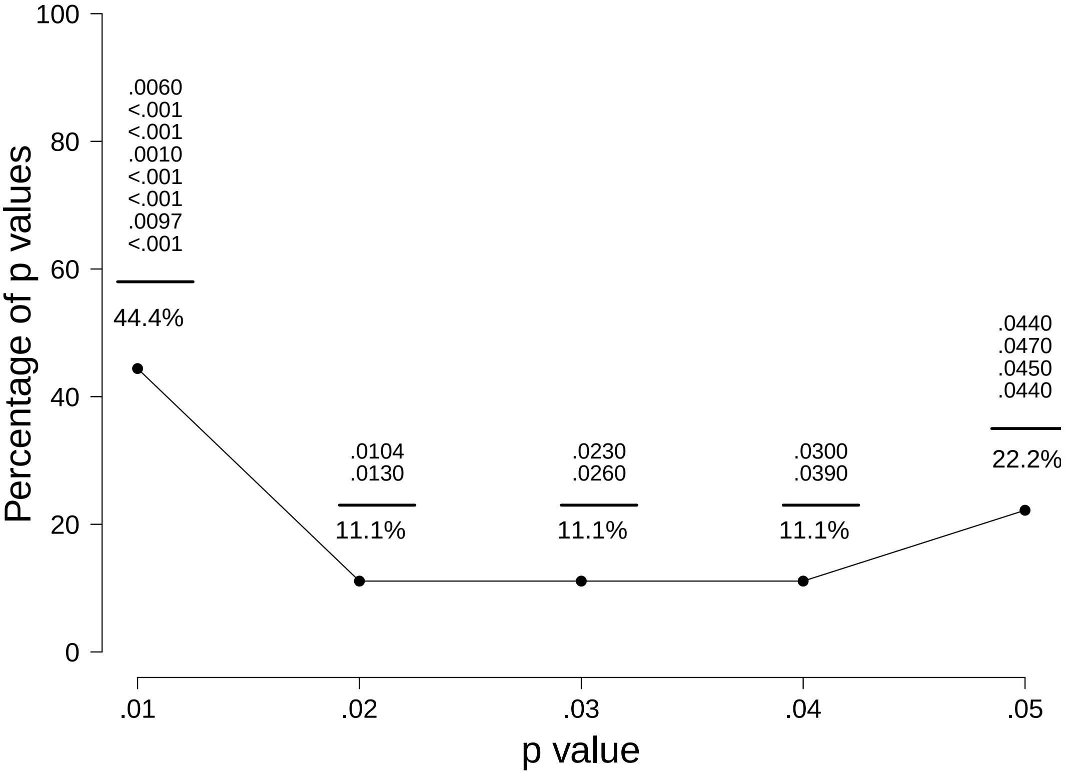


Figure 10.6 *P-curve: The distribution of statistically significant p values in the eye movement literature.* The p-curve shows the percentage of significant p values on the intervals $p < .01$, $.01 \leq p < .02$, $.02 \leq p < .03$, $.03 \leq p < .04$, $.04 \leq p < .05$. The exact p values in a given interval are printed above the corresponding percentage.

the p-curve analysis suggests that the effect of horizontal eye movements on explicit memory is a true effect, but leaves open the possibility that some of the significant findings were p-hacked.

The analysis yielded two other noteworthy findings. First, of the 18 p values that were selected for the p-curve analysis, 11 were $< .025$, and 7 of these 11 more significant p values were published by one group (i.e., Parker, Dagnall, and colleagues). Indeed, altogether only 5 different research groups have contributed to the literature examined here. It is thus important that more laboratories will replicate the effect. Second, in the current study, effects of horizontal eye movements on recall were examined. Therefore, we asked whether there was a difference in p values between studies using recall and studies using recognition tests, as it is possible that horizontal eye movements affect one type of memory more strongly than the other. This was not the case: of the 11 p values $< .025$, 5 reflected recall tests and 6 reflected recognition tests. Of the 7 significant p values $> .025$, 4 were based on recall tests, 3 on recognition tests.

Considering the empirical results and the p-curve analysis reported here, did the present adversarial collaboration resolve the disagreement between the skeptics and the proponents? No; the skeptics are probably no less skeptical, and we, the proponents, are not convinced by a single failure to replicate, especially given the results of the p-curve analysis. However, we have become more cautious about the conclusions that can be drawn from the studies reported so far, and will follow the further development of this field of research with a critical eye. It is important to note that although several authors have speculated about a link between this memory literature and a more clinical literature suggesting that eye movements reduce the vividness and distress associated with emotional autobiographical memories, we do not believe that the current results should lead researchers to call into question those clinical findings. A recent meta-analysis has found significant evidence that eye movements affect the processing of distressing memories in eye-movement desensitization and reprocessing (EMDR) therapy (moderate effect size) and in non-therapy contexts (large effect size; C. W. Lee & Cuijpers, 2013).

Reflection on the Process

Although our adversarial collaboration has not resolved the debate, it has generated new testable ideas and has brought the two parties slightly closer by demonstrating that the beneficial effect of bilateral eye movements on episodic memory is not unequivocal. We recommend that other researchers in this field use similar strict methods in future studies, and emphasize the importance of reporting non-replications.

Discussion by Referee

An impartial referee has been involved in the adversarial collaboration throughout the course of the process. The referee was asked to settle any dispute between parties that might arise with regard to issues not specified in the contract. That did not happen. The parties agreed on the "Adversarial Collaboration Agreement" contract without the need for a referee. The referee received weekly updates during data collection and observed that the parties were able to solve issues not specified in the contract, e.g., the required number of participants or outlier/exclusion criteria, on their own. Finally, and most importantly, the parties agreed upon the outcome of the adversarial collaboration. The results that emerged from this adversarial collaboration show that horizontal eye movements did not improve free recall. Game over and done with? It seems not to be the case. The results are clearly in support of the skeptics' expectations. However, while accepting the negative findings and acknowledging the benefits of preregistered adversarial collaboration, the proponents are not convinced by a single failure to replicate, especially given the results of the p-curve analysis. In retrospect, then, we have to conclude that the adversarial collaboration agreement was not watertight. It should have specified the conditions under which the parties would have been prepared to give up their point of view. If a single failure to replicate, based upon a strict agreement concerning the particulars of the experiment and associated data analysis, is not sufficient, the obvious danger is to encounter a situation well described by an unknown quote "Theories are like old soldiers, they never die but slowly fade away".

Chapter 11

Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests

This chapter has been published as:
Ruud Wetzels, Dora Matzke, Michael D. Lee, Jeffrey N. Rouder, Geoffrey J. Iverson, and Eric-Jan Wagenmakers (2011).
Statistical evidence in experimental psychology: An empirical comparison using 855 t tests.
Perspectives on Psychological Science, 6, 291-298.[1]

Abstract

Statistical inference in psychology has traditionally relied heavily on p value significance testing. This approach to drawing conclusions from data, however, has been widely criticized, and two types of remedies have been advocated. The first proposal is to supplement p values with complementary measures of evidence such as effect sizes. The second is to replace inference with Bayesian measures of evidence such as the Bayes factor. The authors provide a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. The comparison yields two main results. First, although p values and default Bayes factors almost always agree about what hypothesis is better supported by the data, the measures often disagree about the strength of this support; for 70% of the data sets for which the p value falls between 0.01 and 0.05, the default Bayes factor indicates that the evidence is only anecdotal. Second, effect sizes can provide additional evidence to p values and default Bayes factors. The authors conclude that the Bayesian approach is comparatively prudent, preventing researchers from overestimating the evidence in favor of an effect.

11.1 Introduction

Experimental psychologists use statistical procedures to convince themselves and their peers that the effect of interest is real, reliable, replicable, and hence worthy of academic attention. A representative example comes from Mussweiler (2006), who studied whether particular actions can

[1] The final publication is available at `http://pps.sagepub.com/content/6/3/291.short`.

activate a corresponding stereotype. To test this hypothesis empirically, Mussweiler unobtrusively induced half the participants, the experimental group, to move in a portly manner that is stereotypic for the overweight. The other half, the control group, made no such movements. Next, all participants were given an ambiguous description of a target person and then used a 9-point scale (1 = *not at all*, 9 = *very*) to rate this person on dimensions that correspond to the overweight stereotype (e.g., "unhealthy", "sluggish", "insecure"). To assess whether performing the stereotypic motion affected the rating of the ambiguous target person, Mussweiler computed a t statistic ($t(18) = 2.1$), and found that this value corresponded to a low p value ($p < 0.05$).[2] Following conventional protocol, Mussweiler concluded that the low p value should be taken to provide "initial support for the hypothesis that engaging in stereotypic movements activates the corresponding stereotype" (Mussweiler, 2006, p. 18).

The use of t tests and corresponding p values in this way constitutes a common and widely accepted practice in the psychological literature. It is, however, not the only possible or reasonable approach to measuring evidence and making statistical and scientific inferences. Indeed, the use of t tests and p values has been widely criticized (e.g., J. Cohen, 1994; Cumming, 2008; Dixon, 2003; Howard, Maxwell, & Fleming, 2000; M. D. Lee & Wagenmakers, 2005; G. R. Loftus, 1996; Nickerson, 2000; Wagenmakers, 2007). There are at least two different criticisms, coming from different perspectives and resulting in different remedies. First, many have argued that null hypothesis tests should be supplemented with other statistical measures, such as confidence intervals and effect sizes. Within psychology, this approach to remediation has sometimes been institutionalized, being required by journal editors or recommended by the American Psychological Association (e.g., American Psychological Association, 2010; J. Cohen, 1988; Erdfelder, 2010; Wilkinson & the Task Force on Statistical Inference, 1999).

A second, more fundamental criticism that comes from Bayesian statistics is that there are basic conceptual and practical problems with p values. Although Bayesian criticism of psychological statistical practice dates back at least to Edwards et al. (1963), it has become especially prominent and increasingly influential in the last decade (e.g., Dienes, 2008; Gallistel, 2009; Kruschke, 2010c, 2010a; M. D. Lee, 2008; Myung et al., 2000; Rouder et al., 2009). One standard Bayesian measure for quantifying the amount of evidence from the data in support of an experimental effect is the Bayes factor (Gönen, Johnson, Lu, & Westfall, 2005; Rouder et al., 2009; Wetzels et al., 2009). The measure takes the form of an odds ratio: It is the probability of the data under one hypothesis relative to that under another (Dienes, 2011; Kass & Raftery, 1995; M. D. Lee & Wagenmakers, 2005).

With this background, it seems that psychological statistical practice currently stands at a three-way fork in the road. Staying on the current path means continuing to rely on p values. A modest change is to place greater focus on the additional inferential information provided by effect sizes and confidence intervals. A radical change is struck by moving to Bayesian approaches, such as Bayes factors. The path that psychological science chooses seems likely to matter. It is not just that there are philosophical differences between the three choices. It is also clear that the three measures of evidence can be mutually inconsistent (e.g., Berger & Sellke, 1987; Rouder et al., 2009; Wagenmakers, 2007; Wagenmakers & Grünwald, 2006; Wagenmakers et al., 2010).

In this article, we assess the practical consequences of choosing among inference by p values, by effect sizes, and by Bayes factors. By practical consequences, we mean the extent to which conclusions of extant studies change according to the inference measure that is used. To assess these practical consequences, we reanalyzed 855 t tests reported in articles from the 2007 issues

[2] The findings suggest that Mussweiler (2006) conducted a one-sided t test. In the remainder of this article, we conduct two-sided t tests.

of *Psychonomic Bulletin & Review* (PBR) and *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP:LMC). For each t test, we compute the p value, the effect size, and the Bayes factor and study the extent to which they provide information that is redundant, complementary, or inconsistent. On the basis of these analyses, we suggest the best direction for measuring statistical evidence from psychological experiments.

11.2 Three Measures of Evidence

In this section, we describe how to calculate and interpret the p value, the effect size, and the Bayes factor. For concreteness, we use Mussweiler's (2006) study on the effect of action on stereotypes. The mean score of the control group, M_c, was 5.8 on a weight-stereotype scale ($s_c = 0.69, n_c = 10$), and the mean score of the experimental group, M_e, was 6.4 ($s_e = 0.66, n_e = 10$).

The p Value

The interpretation of p values is not straightforward, and their use in hypothesis testing is heavily debated (J. Cohen, 1994; Cortina & Dunlap, 1997; Cumming, 2008; Dixon, 2003; Frick, 1996; Gigerenzer, 1993, 1998; Hagen, 1997; Killeen, 2005, 2006; Kruschke, 2010c, 2010a; M. D. Lee & Wagenmakers, 2005; G. R. Loftus, 1996; Nickerson, 2000; Schmidt, 1996; Wagenmakers & Grünwald, 2006; Wainer, 1999). The p value is the probability of obtaining a test statistic (in this case, the t statistic) at least as extreme as the one that was observed in the experiment, given that the null hypothesis is true and the sample is generated according to a specific intended procedure, such as fixed sample size. Fisher (1935) interpreted these p values as evidence against the null hypothesis. The smaller the p value, the more evidence there was against the null hypothesis. Fisher viewed these values as self-explanatory measures of evidence that did not need further guidance. In practice, however, most researchers (and reviewers) adopt a 0.05 cutoff: p values less than 0.05 constitute evidence for an effect, and those greater than 0.05 do not. More fine-grained categories are possible, and Wasserman (2004, p. 157) proposes the gradations shown in Table 11.1. Note that Table 11.1 lists various categories of evidence against the null hypothesis. A basic limitation of null hypothesis significance testing is that it does not allow a researcher to gather evidence in favor of the null (Dennis, Lee, & Kinnell, 2008; Gallistel, 2009; Rouder et al., 2009; Wetzels et al., 2009).

Table 11.1 Evidence Categories for p Values (adapted from Wasserman, p. 157, 2004).

p value			Interpretation
	$<$	0.001	Decisive evidence against H_0
0.001	–	0.01	Substantive evidence against H_0
0.01	–	0.05	Positive evidence against H_0
	$>$	0.05	No evidence against H_0

For the data from Mussweiler (2006), we compute a p value based on the t test. The t test is designed to test whether a difference between two means is significant. First, we calculate the t statistic:

$$t = \frac{M_e - M_c}{\sqrt{s^2_{pooled}\left(\frac{1}{n_e} + \frac{1}{n_c}\right)}} = \frac{6.42 - 5.79}{\sqrt{0.46\left(\frac{1}{10} + \frac{1}{10}\right)}} = 2.09,$$

where M_c and M_e are the means of both groups, n_c and n_e are the sample sizes, and s^2_{pooled} estimates the common population variance:

$$s^2_{pooled} = \frac{(n_e - 1)s^2_e + (n_c - 1)s^2_c}{n_e + n_c - 2}.$$

Next, the t statistic with $n_e + n_c - 2 = 18$ degrees of freedom results in a p value slightly larger than 0.05 (≈ 0.051). For our concrete example, Table 11.1 leads to the conclusion that the p value is on the cusp between "no evidence against H_0" and "positive evidence against H_0".

The Effect Size

Effect sizes quantify the magnitude of an effect and serves as a measure of how much the results deviate from the null hypothesis (J. Cohen, 1988; Thompson, 2002; Richard, Bond, & Stokes-Zoota, 2003; Rosenthal, 1990; Rosenthal & Rubin, 1982). For the data from Mussweiler (2006), the effect size d is calculated as follows:

$$d = \frac{M_e - M_c}{s_{pooled}} = \frac{6.42 - 5.79}{0.68} = 0.93.$$

Note that in contrast to the p value, the effect size is independent of sample size; increasing the sample size does not increase effect size but instead allows it to be estimated more accurately.

Effect sizes are often interpreted in terms of the categories introduced by J. Cohen (1988), as listed in Table 11.2, ranging from "small" to "very large". For our concrete example, $d = 0.93$, and we conclude that this effect is large to very large. Interestingly, the p value was on the cusp between the categories "no evidence against H_0" and "positive evidence against H_0" whereas the effect size indicates the effect to be strong.

Table 11.2 Evidence Categories for Effect Sizes as Proposed by J. Cohen (1988).

Effect Size	Interpretation
< 0.2	Small effect size
0.2 – 0.5	Small to medium effect size
0.5 – 0.8	Medium to large effect size
> 0.8	Large to very large effect size

The Bayes Factor

In Bayesian statistics, uncertainty (or degree of belief) is quantified by probability distributions over parameters. This makes the Bayesian approach fundamentally different from the classical "frequentist" approach, which relies on sampling distributions of data (Berger & Delampady, 1987; Berger & Wolpert, 1988; Jaynes, 2003; Lindley, 1972).

Within the Bayesian framework, one may quantify the evidence for one hypothesis relative to another. The Bayes factor is the most commonly used (although certainly not the only possible) Bayesian measure for doing so (Jeffreys, 1961; Kass & Raftery, 1995). The Bayes factor is the probability of the data under one hypothesis relative to the other. When a hypothesis is a simple point, such as the null, then the probability of the data under this hypothesis is simply the likelihood evaluated at that point. When a hypothesis consists of a range of points, such as all positive effect sizes, then the probability of the data under this hypothesis is the weighted average of the likelihood across that range. This averaging automatically controls for the complexity of different models, as has been emphasized in Bayesian literature in psychology (e.g., Pitt et al., 2002; Rouder et al., 2009).

We take as the null that a parameter α is restricted to 0 (i.e., $H_0 : \alpha = 0$), and take as the alternative that α is not zero (i.e., $H_A : \alpha \neq 0$). In this case, the Bayes factor given data D is simply the ratio

$$BF_{A0} = \frac{p\,(D \mid H_A)}{p\,(D \mid H_0)} = \frac{\int p\,(D \mid H_A, \alpha)\, p\,(\alpha \mid H_A)\ \mathrm{d}\alpha}{p\,(D \mid H_0)},$$

where the integral in the denominator takes the average evidence over all values of α, weighted by the prior probability of those values $p\,(\alpha \mid H_A)$ under the alternative hypothesis.

An alternative —but formally equivalent— conceptualization of the Bayes factor is as a measure of the change from prior model odds to posterior model odds, brought about by the observed data. This change is often interpreted as the *weight of evidence* (Good, 1983; Good, 1985). Before seeing the data D, the two hypotheses H_0 and H_A are assigned prior probabilities $p(H_0)$ and $p(H_A)$. The ratio of the two prior probabilities defines the *prior odds*. When the data D are observed, the prior odds are updated to *posterior odds*, which is defined as the ratio of the posterior probabilities $p(H_0 \mid D)$ and $p(H_A \mid D)$:

$$\frac{p(H_A \mid D)}{p(H_0 \mid D)} = \frac{p(D \mid H_A)}{p(D \mid H_0)} \times \frac{p(H_A)}{p(H_0)}. \tag{11.1}$$

Equation 11.1 shows that the change from prior odds to posterior odds is quantified by $p(D \mid H_A)/p(D \mid H_0)$, the Bayes factor BF_{A0}.

Under either conceptualization, the Bayes factor has an appealing and direct interpretation as an odds ratio. For example, $BF_{A0} = 2$ implies that the data are twice as likely to have occurred under H_A than under H_0. Jeffreys (1961), proposed a set of verbal labels to categorize the Bayes factor according to its evidential impact. This set of labels, presented in Table 11.3, facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence (Kass & Raftery, 1995).

Table 11.3 Evidence Categories for the Bayes Factor BF_{A0} (Jeffreys, 1961).

Bayes factor			Interpretation
	>	100	Decisive evidence for H_A
30	–	100	Very strong evidence for H_A
10	–	30	Strong evidence for H_A
3	–	10	Substantial evidence for H_A
1	–	3	Anecdotal evidence for H_A
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Substantial evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very strong evidence for H_0
	<	1/100	Decisive evidence for H_0

Note. We replaced the label “worth no more than a bare mention” with “anecdotal”. Note that, in contrast to p values, the Bayes factor can quantify evidence in favor of the null hypothesis.

In general, calculating Bayes factors is more difficult than calculating p values and effect sizes. However, psychologists can now turn to easy-to-use Web pages to calculate the Bayes factor for many common experimental situations or use software such as WinBUGS (Lunn et al., 2000; Wetzels et al., 2009; Wetzels, Lee, & Wagenmakers, 2010).[3] In this article, we use the Bayes factor calculation described in Rouder et al. (2009). Rouder et al.'s development is suitable for one-sample and two-sample designs, and the only necessary input is the t value and sample size.

The Bayes factor that we report in this article is the result of a default Bayesian t test (for details see Rouder et al., 2009). The test is default because it applies regardless of the phenomenon under study: For every experiment, one uses the same prior on effect size for the alternative hypothesis, the Cauchy(0,1) distribution. This prior has statistical advantages that make it an appropriate default choice (for example, it has excellent theoretical properties in the limit, when $N \to \infty$ and $t \to \infty$; for details see Liang et al., 2008).

The default test is easy to use and avoids informed specification of prior distributions that other researchers may contest. Conversely, one may argue that the informed specification of priors is the appropriate way to take problem-specific prior knowledge into account. Bayesian statisticians are divided over the relative merits of default versus informed specifications of prior distributions (Press, Chib, Clyde, Woodworth, & Zaslavsky, 2003). In our opinion, the default test provides an excellent starting point of analysis, one that may later be supplemented with a detailed problem-specific analysis (see Dienes, 2011, 2008; Kruschke, 2010a, 2010b, 2011, for additional discussion of informed priors).

In our concrete example, the resulting Bayes factor for $t = 2.09$ and a sample size of 20 observations is $BF_{A0} = 1.56$. Accordingly, the data are 1.56 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. This Bayes factor falls into the category "anecdotal". In other words, this Bayes factor indicates that although the alternative hypothesis is slightly favored, we do not have sufficiently strong evidence from the data to reject or accept either hypothesis.

11.3 Comparing p Values, Effect Sizes and Bayes Factors

For our concrete example, the three measures of evidence are not in agreement. The p value was on the cusp between the categories "no evidence against H_0" and "positive evidence against H_0", the effect size indicates a large to very large effect size, and the Bayes factor indicates that the data support the null hypothesis almost as much as they support the alternative hypothesis. If this example is not an isolated one, and the measures differ in many psychological applications, then it is important to understand the nature of those differences.

To address this question, we studied all of the empirical results evaluated by a t test in the 2007 volumes of *PBR* and *JEP:LMC*. This sample was comprised of 855 t tests from 252 articles. These articles covered 2,394 journal pages and addressed many topics that are important in modern experimental psychology. Our sample suggests, on average, that an article published in *PBR* and *JEP:LMC* contains about 3.4 t tests, which amounts to one t test for every 2.8 pages. For simplicity we did not include t tests that result from multiple comparisons in ANOVA designs (for a Bayesian perspective on multiple comparisons see Scott & Berger, 2006). Even though our t tests are sampled from the field of experimental/cognitive psychology, we expect our findings to generalize to many other subfields of psychology, as long as the studies in these subfields use the same level

[3] A Web page for computing a Bayes factor online is `http://pcl.missouri.edu/bayesfactor`, and a Web page to download a tutorial and a flexible R/WinBUGS function to calculate the Bayes factor can be found at `www.ruudwetzels.com`.

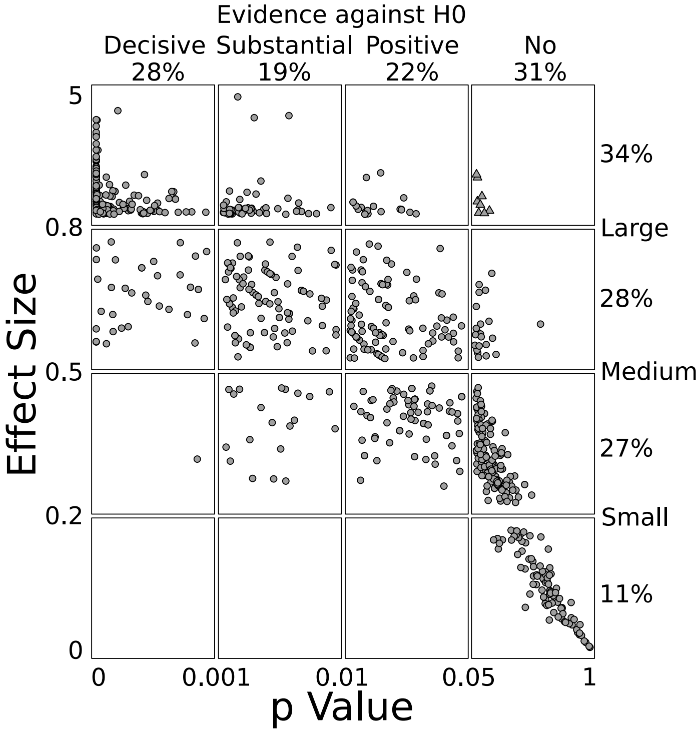


Figure 11.1 The relationship between effect size and p values. Points denote comparisons (855 in total). Points denoted by circles indicate relative consistency between the effect size and p value, whereas those denoted by triangles indicate gross inconsistency. The scale of the axes is based on the decision categories, as given in Table 11.1 and Table 11.2.

of statistical significance, approximately the same number of participants, and approximately the same number of trials per participant (Howard et al., 2000).

In the next sections we describe the empirical relation between the three measures of evidence, starting with the relation between effect sizes and p values.

Comparing Effect Sizes and p Values

The relationship between the obtained p values and effect sizes is shown as a scatter plot in Figure 11.1. Each point corresponds to one of the 855 comparisons. Different panels are introduced to distinguish the different evidence categories, as given in Table 11.1 and Table 11.2.

Figure 11.1 suggests that p values and effect sizes capture roughly the same information in the data. Large effect sizes tend to correspond to low p values, and small effect sizes tend to correspond to large p values. The two measures, however, are far from identical. For instance, a p value of 0.01 can correspond to effect sizes ranging from about 0.2 to 1, and an effect size near 0.5 can

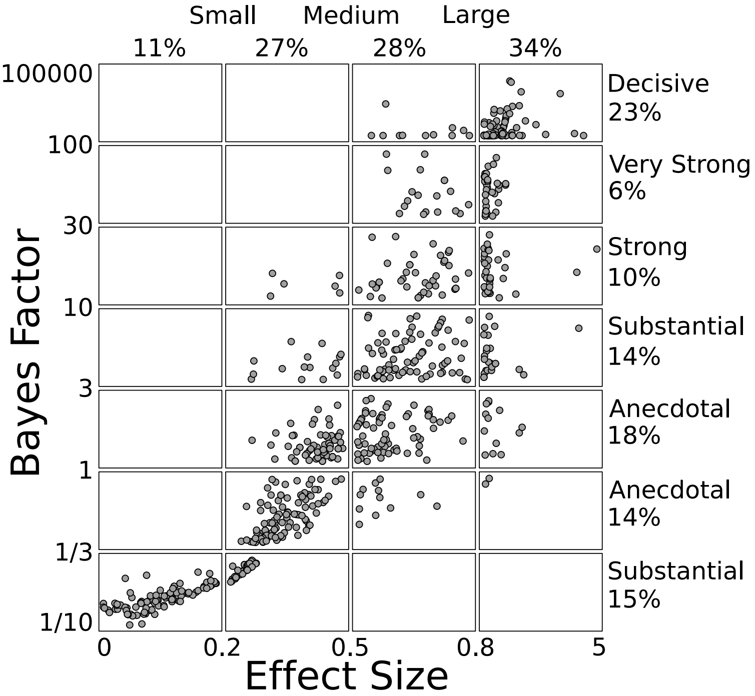


Figure 11.2 The relationship between Bayes factor and effect size. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 11.2 and Table 11.3.

correspond to p values ranging from about 0.001 to 0.05. The triangular points in the top-right panel of Figure 11.1 highlight gross inconsistencies. These 8 studies have a large effect size, above 0.8, but their p values do not indicate evidence against the null hypothesis. A closer examination revealed that these studies had p values very close to 0.05, and were comprised of small sample sizes.

Comparing Effect Sizes and Bayes Factors

The relationship between the obtained Bayes factors and effect sizes is shown in Figure 11.2. Much as with the comparison of p values with effect sizes, it seems clear that the default Bayes factor and effect size generally agree, though not exactly. No striking inconsistencies are apparent: No study with an effect size greater than 0.8 coincides with a Bayes factor below 1/3, nor does a study with very low effect size below 0.2 coincide with a Bayes factor above 3. The two measures, however, are not identical. They differ in the assessment of strength of evidence. Effect sizes above 0.8 range all the way from anecdotal to decisive evidence in terms of the Bayes factor. Also note that small to medium effect sizes (i.e., those between 0.2 and 0.5) can correspond to Bayes factor evidence in favor of either the alternative or the null hypothesis.

This last observation highlights that Bayes factors may quantify support for the null hypothesis. Figure 11.2 shows that about one-third of all studies produced evidence in favor of the null hypothesis. In about half of these studies favoring the null, the evidence is substantial. Because of the file-drawer problem (i.e., only significant effects tend to get published) this is an underestimate of the true amount of null findings and their Bayes factor support.

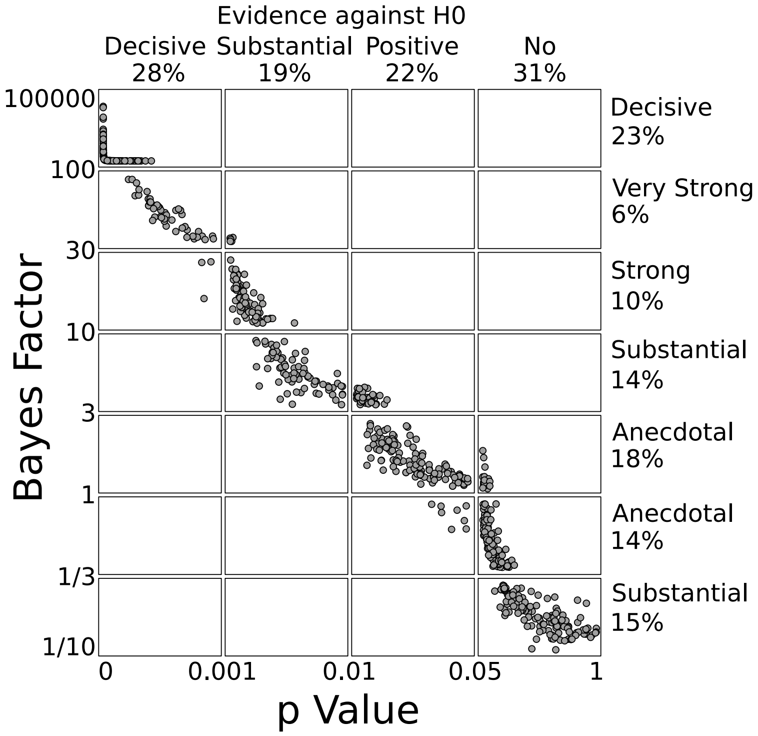


Figure 11.3 The relationship between Bayes factor and p value. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 11.1 and Table 11.3.

Comparing p Values and Bayes Factors

The relationship between the obtained Bayes factors and p values is shown in Figure 11.3, again using interpretative panels. It is clear that default Bayes factors and p values largely covary with each other. Low Bayes factors correspond to high p values and high Bayes factors correspond to low p values, a relationship that is much more exact than for our previous two comparisons. The main difference between default Bayes factors and p values is one of calibration; p values accord more evidence against the null than do Bayes factors. Consider the p values between 0.01 and 0.05, values that correspond to “positive evidence” and that usually pass the bar for publishing in academia. According to the default Bayes factor, 70% of these experimental effects convey evidence in favor of the alternative hypothesis that is only “anecdotal”. This difference in the assessment of the strength of evidence is dramatic and consequential.

11.4 Conclusions

We compared p values, effect sizes and default Bayes factors as measures of statistical evidence in empirical psychological research. Our comparison was based on a total of 855 different t statistics from all published articles in two major empirical journals in 2007. In virtually all studies, the three different measures of evidence are broadly consistent: Small p values correspond to large effect sizes and large Bayes factors in favor of the alternative hypothesis. Despite the fact that the measures of evidence reach the same conclusion about what hypothesis is best supported by the data, however, the measures differ with respect to the strength of that support. In particular, we noted that p values between 0.01 and 0.05 often correspond to what, in Bayesian terms, is only anecdotal evidence favor of the alternative hypothesis. The practical ramifications of this are considerable.

Practical Ramifications

Our results showed that when the p value falls in the interval from 0.01 to 0.05, there is a 70% chance that the default Bayes factor indicates the evidence for the alternative hypothesis to be only anecdotal or "worth no more than a bare mention"; this means that the data are no more than three times more likely under the alternative hypothesis than they are under the null hypothesis. Hence, for the studies under consideration here, it seems that a p value criterion more conservative than 0.05 is appropriate. Alternatively, researchers could avoid computing a p value altogether and instead compute the Bayes factor. Both methods help prevent researchers from overestimating the strength of their findings and help the field from incorporating ambiguous findings as if they were real and reliable (Ioannidis, 2005).

As a practical illustration, consider a series of recent experiments on precognition (Bem, 2011). In nine experiments with over 1,000 participants, Bem intended to show that precognition exists, that is, that people can foresee the future. And indeed, eight out of nine experiments yielded a significant result. However, most p values fell in the ambiguous range of 0.01 to 0.05, and, across all nine experiments, a Bayes factor analysis indicates about as much evidence for the alternative hypothesis as against it (Kruschke, 2011; Wagenmakers et al., 2011). We believe that this situation typifies part of what could be improved in psychological research today. It is simply too easy to obtain a p value below 0.05 and subsequently publish the result.

When researchers publish ambiguous results as if they were real and reliable, this damages the field as a whole: Time, effort, and money will be invested to replicate the phenomenon, and, when replication fails, the burden of proof is almost always on the part of the researcher who, after all, failed to replicate a phenomenon that was demonstrated to be present (with a p value in between 0.01 and 0.05).

Thus, our empirical comparison shows that the academic criterion of 0.05 is too liberal. Note this problem would not be solved by opting for a stricter significance level, such as 0.01. It is well known that the p value decreases as the sample size n increases. Hence, if psychologists switch to a significance level of 0.01 but inevitably increase their sample sizes to compensate for the stricter statistical threshold, then the phenomenon of anecdotal evidence will start to plague p values even when these p values are lower than 0.01. Therefore, we make a case for Bayesian statistics in the next section.

A Case for Bayesian Statistics

We have compared the conclusions from the different measures of evidence. It is easy to make a case for Bayesian statistical inference in general, based on arguments already well documented in statistics and psychology (e.g., Dienes, 2008; Jaynes, 2003; Kruschke, 2010c, 2010a; M. D. Lee & Wagenmakers, 2005; Lindley, 1972; Wagenmakers, 2007). We briefly mention three arguments here.

First, unlike null hypothesis testing, Bayesian inference does not violate basic principles of rational statistical decision making such as the stopping rule principle or the likelihood principle (Berger & Delampady, 1987; Berger & Wolpert, 1988). This means that the results of Bayesian inference do not depend on the intention with which the data were collected. As stated by Edwards et al. (1963, p. 193), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience".

Second, Bayesian inference takes model complexity into account in a rational way. Specifically, the Bayes factor has the attraction of not assigning a special status to the null hypothesis and so makes it theoretically possible to measure evidence in favor of the null (e.g., Dennis et al., 2008; Gallistel, 2009; Kass & Raftery, 1995; Rouder et al., 2009).

Third, we believe that Bayesian inference provides the kind of answers that researchers care about. In our experience, researchers are usually not that interested in the probability of encountering data at least as extreme as those that were observed, given that the null hypothesis is true and the sample was generated according to a specific intended procedure. Instead, most researchers want to know what they have learned from the data about the relative plausibility of the hypotheses under consideration. This is exactly what is quantified by the Bayes factor.

These advantages notwithstanding, the Bayes factor is not a measure of the mere size of an effect. Hence, the measure of effect size confers additional information, particularly when small numbers of participants or trials are involved. So, especially for these sorts of studies, there is an argument for reporting both a Bayes factor and an effect size. We note that, from a Bayesian perspective, the effect size can naturally be conceived as a (summary statistic of) the posterior distribution of a parameter representing the effect, under an uninformative prior distribution. In this sense, a standard Bayesian combination of parameter estimation and model selection could encompass all of the useful measures of evidence we observed (for an example of how Bayes factor estimation can be incorporated in a Bayesian estimation framework, see, for instance, Kruschke, 2011).

Our final thought is that reasons for adopting a Bayesian approach now are amplified by the promise of using an extended Bayesian approach in the future. In particular, we think the hierarchical Bayesian approach, which is standard in statistics (e.g., Gelman & Hill, 2007), and is becoming more common in psychology (e.g., Kruschke, 2010c, 2010b; M. D. Lee, 2011; Rouder & Lu, 2005) could fundamentally change how psychologists identify effects. Hierarchical Bayesian analysis can be a valuable tool both for meta-analyses and for the analysis of a single study. In the meta-analytical context, multiple studies can be integrated, so that what is inferred about the existence of effects and their magnitude is informed, in a coherent and quantitative way, by a domain of experiments. In the context of a single experiment, a hierarchical analysis can be used to take variability across participants or items into account.

In sum, our empirical comparison of 855 t tests shows that three often-used measures of evidence —p values, effect sizes, and Bayes factors— almost always agree about what hypothesis is better supported by the data. The measures often disagree about the strength of this support: for those data sets with p values in between 0.01 and 0.05, about 70% are associated with a Bayes factor that indicates the evidence to be only anecdotal or “worth no more than a bare mention” (Jeffreys, 1961). This analysis suggests that many results that have been published in the literature are not established as strongly as one would like.

Chapter 12

Hidden Multiplicity in Multiway ANOVA: Prevalence, Consequences, and Remedies

A modified version of this chapter has been submitted for publication as:
Angélique O. J. Cramer, Don van Ravenzwaaij, Dora Matzke, Helen Steingroever, Ruud Wetzels,
Raoul P. P. P. Grasman, Lourens Waldorp, and Eric-Jan Wagenmakers (2013).
Hidden multiplicity in multiway ANOVA: Prevalence, consequences, and remedies.

Abstract

Many empirical researchers do not realize that the common multiway analysis of variance (ANOVA) harbors a multiple comparison problem. In the case of two factors, three separate null hypotheses are subject to test (i.e., two main effects and one interaction). Consequently, the probability of a Type I error is 14% rather than 5%. He we describe the multiple comparison problem and demonstrate that researchers seldom correct for it. We then illustrate the use of the sequential Bonferroni correction —one of several correction procedures— and show that its application alters at least one of the substantive conclusions in 45 out of 60 articles considered. An alternative method to combat the multiplicity problem in multiway ANOVA is preregistration of the hypotheses.

12.1 Introduction

The factorial or multiway analysis of variance (ANOVA) is one of the most popular statistical procedures in psychology. Whenever an experiment features two or more factors, researchers usually apply a multiway ANOVA to gauge the evidence for the effect of each of the separate factors as well as their interactions. For instance, consider a response time experiment with a 2×3 balanced design (i.e., a design with equal number of participants in each condition); factor A is speed-stress (high or low) and factor B is the age of the participants (14-20 years, 50-60 years, and 75-85 years). The standard multiway ANOVA tests whether factor A is significant, whether factor B is significant, and whether the interaction term $A \times B$ is significant. In the same vein, the standard multiway ANOVA is also frequently used in non-experimental settings (e.g., to assess the potential influence of gender and age on major depression).

Despite its popularity, few researchers realize that the multiway ANOVA harbors a multiple comparisons problem, particularly when detailed hypotheses have not been specified a priori (to be discussed in more detail later). Consider, for example, the 2×3 scenario introduced above. Without a-priori hypotheses (i.e., when the researcher's attitude can best be described by "let us see what we can find"; de Groot, 1969), three independent tests are carried out. Hence, given the null hypothesis (H_0) is true and $\alpha = 0.05$, the probability of at least one significant result equals $1 - (1 - 0.05)^3 = 0.14$. This is called a Type I error or familywise error rate. The problem of Type I error is not trivial: add a third, balanced factor to the 2×3 scenario (e.g., a $2\times3\times3$ design), and the probability of finding at least one significant result when H_0 is true increases to $1 - (1 - 0.05)^7 = 0.30$. The situation becomes even more troublesome for designs with unequal numbers of participants per condition: in such unbalanced designs, the three tests in our hypothetical 2×3 experiment are no longer independent and this further increases the probability of a Type I error (Rao & Toutenburg, 1999). Thus, in the absence of strong a priori expectations about the tests that are relevant, the α-inflation is dramatic and should be cause for great concern.

The goal of the present article is to highlight the problem of multiple comparison inherent in multiway ANOVA. To this end, we first conduct a literature review and demonstrate that the problem is widely ignored: recent articles published in leading psychology journals contain virtually no procedures to correct for the multiple comparison problem. Next, we outline one possible remedy, the sequential Bonferroni procedure (Hartley, 1955; Hochberg, 1988; Holm, 1979; McHugh, 1958; Shaffer, 1986; Wright, 1992). Finally, we demonstrate that the sequential Bonferroni procedure alters at least one of the substantive conclusions in 45 of 60 randomly chosen articles. In order to prevent the loss of power that is inherent to all correction procedures, we recommend the pre-registration of the hypotheses of interest.

12.2 Type I Errors and the Oneway ANOVA

A Type I error occurs when the null hypothesis (H_0) is falsely rejected in favor of the alternative hypothesis (H_1). With a single test, such as the oneway ANOVA, the probability of a Type I error can be controlled by setting the significance level α. For example, when $\alpha = 0.05$, the probability of a Type I error is 5%. It is well-known, however, that the multiple comparison problem arises even in the oneway ANOVA whenever the independent variable has more than two levels and post-hoc tests are employed to investigate which condition means differ significantly from one another. Consider, for instance, a researcher who uses a oneway ANOVA and obtains a significant effect for *Ethnicity* on the total score of a depression questionnaire. Assume that *Ethnicity* has three levels (e.g., Caucasian, African-American, and Asian); hence, the researcher will usually perform multiple post-hoc tests to investigate which ethnic groups differ significantly from one another —here the three post-hoc tests are Caucasian vs. African-American, Caucasian vs. Asian, and African-American vs. Asian. Note that when the three test statistics are independent —as for balanced designs— the overall Type I error equals $1 - (1 - 0.05)^3 = 0.14$. That is, the probability that at least one post-hoc test leads to a false rejection of H_0 has increased almost threefold. Fortunately, for the oneway ANOVA, the multiple comparison problem has been thoroughly studied. Software programs such as SPSS (IBM Corp., 2012) explicitly address the multiple comparison problems by offering a host of correction methods including Tukey's HSD test, Hochberg's GT2, and the Scheffé method (Hochberg, 1974; Scheffé, 1953; Tukey, 1973).

12.3 The Explorative Multiway ANOVA: A Family of Hypotheses

Now consider a design that is only slightly more complicated. Suppose a researcher wants to examine the effect of *Ethnicity* (E; three levels) as well as *Gender* (G; two levels) on the total score on a depression questionnaire. Furthermore, suppose that the researcher in question has no firm a priori hypotheses about how E and G influence the depression total score; that is, the researcher is predominantly interested in finding out whether *any* kind of relationship exists between E, G and depression, a classic example of the *guess* phase of the empirical cycle in which hypotheses are formed rather than tested (de Groot, 1969).

In the above example, the multiway ANOVA without strictly formulated a priori hypotheses is an *explorative* one: Using statistical software, such as SPSS, the researcher obtains the results for all three hypotheses involved (i.e., main effect of E, main effect of G, and the $E \times G$ interaction) by means of a single mouse click. As such, in an explorative setting, all hypotheses implied by the design are considered and tested jointly, rendering this collection of hypotheses a *family*, where "... the term 'family' refers to the collection of hypotheses [...] that is being considered for joint testing" (Lehmann & Romano, 2005). We therefore argue that a multiple comparison problem lurks in the explorative use of the multiway ANOVA.

To see this, consider the results of a fictitious explorative multiway ANOVA reported in Table 12.1. When interpreting the ANOVA table, most researchers would conclude that both main effects as well as the interaction are significant because all p values are smaller than $\alpha = 0.05$. This conclusion is intuitive and directly in line with the numbers reported in Table 12.1. Nevertheless, this conclusion is incorrect. The researcher does not have firm a priori hypotheses and tests all three hypotheses simultaneously and is therefore engaged in an explorative "fishing expedition". The tests in the multiway ANOVA for balanced designs are independent (Toutenburg, 2002) and thus the multiple comparison problem, when unaddressed, results in a 14% Type I error probability. Note that multiway ANOVAs in the psychological literature often consist of three or four factors and this compounds the problem. In the case of an explorative multiway ANOVA with three factors, we are dealing with seven tests (i.e., three main effects, three first-order interactions, and one second-order interaction), resulting in a 30% Type I error probability; with four factors, the probability of incorrectly rejecting one or more null hypotheses increases to 54%. It is therefore incorrect to evaluate the p values from a multiway ANOVA table with $\alpha = 0.05$.

Note that the above sketched scenario is different from the situation where the researcher uses a multiway ANOVA for *confirmatory* purposes; that is, when the researcher tests one or more a priori postulated hypotheses (i.e., hypothesis testing in the *predict* phase of the empirical cycle; de Groot, 1969). For instance, consider a design with two factors and one pre-defined hypothesis: the family is no longer defined as encompassing all three hypotheses implied by the design, but as all to-be-tested hypotheses, in this case a single hypothesis, rendering it unnecessary to adjust the α level.

The realization that explorative multiway ANOVAs inherently harbor a multiple comparison problem may come as a surprise to many empiricists, even to those who regularly use multiway ANOVAs. In standard statistical textbooks, the multiple comparison problem is almost exclusively discussed in the context of oneway ANOVAs. In addition, statistical software packages, such as SPSS, do not present the possible correction procedures for the multiway case, inviting researchers to evaluate the resulting p values with $\alpha = 0.05$.

We are by no means the first to identify the multiplicity problem in multiway ANOVAs (see, e.g., Didelez, Pigeot, & Walter, 2006; Fletcher, Daw, & Young, 1989; Kromrey & Dickinson, 1995;

Table 12.1 Example ANOVA Table for the Three Tests Associated with a 2 × 3 Design with Gender (G) and Ethnicity (E) as Independent Factors.

Effect	Factor	df_1	df_2	F	p value
Main	G	1	30	5	0.0329*
Main	E	2	30	4	0.0288*
Interaction	$G \times E$	2	30	4.5	0.0195*

Note. $*$,significant at $\alpha = 0.05$

Olejnik, Li, Supattathum, & Huberty, 1997; Ryan, 1959; R. A. Smith, Levine, Lachlan, & Fediuk, 2002). Earlier work, however, does not feature in mainstream statistical textbooks and is written in a technical style that is inaccessible to scholars without sophisticated statistical knowledge. As a result, empirical work has largely ignored the multiplicity problem. Unfortunately, as we will demonstrate shortly, the ramifications can be profound.

One may argue that the problem sketched above is less serious than it appears. Perhaps the majority of researchers test only a single pre-specified hypothesis, thereby circumventing the multiple comparison problem. Or perhaps, whenever researchers conduct multiple tests, they use some sort of procedure to adjust the α level for the individual tests. This is unfortunately not the case.

With respect to the former, it is quite common to perform what Gigerenzer (2004) has termed the "null ritual" where researchers specify H_0 in purely statistical terms (e.g., equality of means) without providing an alternative hypothesis in substantive terms (e.g., women are more depressed than men). Additionally, Kerr (1998) notes that researchers are often lurked into presenting a post-hoc hypothesis (e.g., Caucasian people are more depressed than African-American people: main effect of *Ethnicity* on depression) as if it were an a priori hypothesis (i.e., Hypothesizing After the Results are Known: HARKing). Hence, hindsight bias and confirmation bias make it difficult for researchers to ignore unexpected "significant" (i.e., individual test with $p < 0.05$) effects.

With respect to the latter, in the next section, we investigate whether researchers correct for multiple comparisons when they use multiway ANOVAs. The short answer is that, almost without exception, researchers interpret the results of the individual tests in isolation, without any correction for multiple comparisons.

12.4 Practice: Multiway Corrections in Six Journals

We selected six journals that rank among the most widely read and cited journals in experimental, social, and clinical psychology. Specifically, we investigated all 2010 publications of the following journals:

1. *Journal of Experimental Psychology: General* (volume 139; issues 1-4; 40 papers).
2. *Psychological Science* (volume 21; issues 1-12; 285 papers).
3. *Journal of Abnormal Psychology* (volume 119; issues 1-4; 88 papers).
4. *Journal of Consulting and Clinical Psychology* (volume 78; issues 1-6; 92 papers).
5. *Journal of Experimental Social Psychology* (volume 46; issues 1-6; 178 papers).

Table 12.2 Percentage of Articles Overall and in the Six Selected Journals that Used a Multiway ANOVA, and the Percentage of These Articles that Used Some Sort of Correction Procedure.

Journal	% with mANOVA	% with mANOVA & correction
Journal of Experimental Psychology: General	84.61	0
Psychological Science	43.16	0
Journal of Abnormal Psychology	31.82	0
Journal of Consulting and Clinical Psychology	16.30	0
Journal of Experimental Social Psychology	65.17	2.59
Journal of Personality and Social Psychology	54.41	1.35
Overall	47.62	1.03

Note. Overall, all papers from the six journals; mANOVA, multiway ANOVA.

6. *Journal of Personality and Social Psychology* (volumes 98-99; issues 1-6; 136 papers).

For each article, we assessed whether the papers featured one of more multiway ANOVAs and whether the authors had reported some sort of correction procedure (e.g., an omnibus test; see below) to remedy the multiple comparison problem. The results are summarized in Table 12.2.

Two results stand out. First, almost half of the articles under investigation here used a multiway ANOVA, underscoring the popularity of this testing procedure. Second, only around 1% of the articles used a correction procedure. In all four cases where a correction procedure was used, this was an omnibus F test, where one pools the sums of squares and degrees of freedom for all main effects and interactions into a single F statistic. The individual F tests should only be conducted if both this omnibus H0 is rejected as well as all other combinations of null hypotheses (Fletcher et al., 1989; Wright, 1992). For example, for a 2×2 ANOVA, one should first test the omnibus hypothesis with all three hypotheses included (two main effects and the interaction). If significant, then one proceeds to test the three combinations of two null hypotheses (i.e., main effects A and B, main effect A and interaction, main effect B and interaction). Finally, if significant, only then can one safely continue and test the individual hypotheses. When this closed test procedure is followed, one is safeguarded against capitalization on chance both for unbalanced and balanced designs (Shaffer, 1995).

In sum, our literature review confirms that the multiway ANOVA is a highly popular statistical method in psychological research, but that its use is almost never accompanied by a correction for multiple comparisons. Note that this state of affairs is different for fMRI and genetics research where the problem is more evident and it is common practice to correct for multiplicity (e.g., Poldrack et al., 2008).

12.5 Possible Remedy: Sequential Bonferroni Correction

As noted earlier, statisticians have long been aware of the multiple comparison problem in multiway ANOVAs. However, our literature review demonstrated that this awareness has not resonated in the arena of empirical research in psychology.

In the few cases where a correction procedure was used, this involved an omnibus F test, a test that cannot control the familywise Type I error under partial null conditions (Kromrey & Dickinson, 1995). For example, suppose that in a threeway ANOVA, a main effect is present

Table 12.3 The Sequential Bonferroni Procedure for the Hypothetical Example of Table 1.

Effect	p value	α_{adj}	H_0
$G \times E$	0.0195	0.0167	not rejected
E	0.0288	0.0250	not rejected
G	0.0329	0.0500	not rejected

Note. The sequential Bonferroni procedure entails: (1) sorting p values in ascending order; (2) computing adjusted α level per test (α_{adj}); (3) sequentially evaluating each p value with adjusted α level (i.e., reject or not reject H_0); and (4) stopping whenever H_0 is not rejected (and do not reject all remaining untested H_0).

for one factor but not for the remaining two factors; then the overall F test is likely to yield a significant F value because, indeed, the omnibus H_0 is false. However, the omnibus test does not remedy the multiple comparison problem involving the remaining two factors.

A more general correction is known as the sequential Bonferroni procedure (also known as the Bonferroni-Holm correction). The sequential Bonferroni correction was first introduced by Hartley (1955) and subsequently (independently) re-invented and/or modified by others (Hochberg, 1988; Holm, 1979; McHugh, 1958; Shaffer, 1986; Rom, 1990; Wright, 1992). How does the procedure work? Let us revisit our hypothetical example in which a researcher conducts a twoway ANOVA with G and E as independent factors (uncorrected results are listed in Table 12.1). The results of the sequential Bonferroni correction procedure for this example are presented in Table 12.3. First, one sorts all significant p values in ascending order, that is, with the smallest p value first. Next, one computes an adjusted α level, α_{adj}. For the smallest p value, α_{adj} equals α divided by the number of tests. In the present example, we conduct three tests so α_{adj} for the smallest p value equals $0.05/3 = 0.01667$. For the second p value, α_{adj} equals α divided by the number of tests minus 1: $\alpha_{adj} = 0.05/2 = 0.025$. For the final p value, α_{adj} equals α divided by the total number of tests minus 2: $\alpha_{adj} = 0.05/1 = 0.05$. Next, one evaluates each p value with the adjusted α level, sequentially, with the smallest p value evaluated first. Importantly, if the H_0 associated with a given p value is not rejected (i.e., $p > \alpha_{adj}$), all testing ends and all remaining tests are also considered non-significant.

In our example, we evaluate $p = 0.0195$ with $\alpha_{adj} = 0.01667$: $p > \alpha_{adj}$ and therefore we conclude that the $G \times E$ interaction is not significant. We therefore stop testing and conclude that the remaining main effects are not significant either. Thus, with the sequential Bonferroni correction procedure, we conclude that none of the effects are significant; without the correction procedure, we had concluded that all of the effects were significant.

We note that other correction procedures are available, for instance those that focus on the *false discovery rate* (Benjamini & Hochberg, 1995); these other procedures might result in a different conclusion. The false discovery rate procedure, for example, which we will later discuss in more detail, is less conservative than the sequential Bonferroni correction and would have resulted in more effects being judged significant.

Thus, the sequential Bonferroni correction procedure allows one to control for the familywise error by evaluating each H_0 –from the one associated with the smallest to the one associated with the largest p value– against an α level that is adjusted in order to control for the inflated probability of a Type I error. In this way, the probability of incorrectly rejecting one or more null hypotheses will be no larger than 5% (for a proof, see Hartley, 1955). Note that for a relatively small number of tests k, the sequential Bonferroni correction is notably less conservative than the standard Bonferroni correction where one divides α by k for all null hypotheses. However, sequential Bonferroni is a conservative procedure in that it never rejects the remaining null hypotheses whenever a given H_0 is not rejected, regardless of how many null hypotheses remain. That is, it does not matter whether we deal with five or 50 null hypotheses, one single H_0 that is not rejected means that the remaining null hypotheses cannot be rejected either. As such, it has been argued that procedures such as (sequential) Bonferroni, while adequately reducing the probability of a Type I error, reduce power and hence inflate the probability of a Type II error (i.e., not rejecting H_0 when H_1 is true; e.g., Benjamini & Yekutieli, 2001; Nakagawa, 2004).

An alternative might be to forego control of the familywise error and instead control the false discovery rate, which is the expected proportion of erroneous rejections of H_0 among all rejections of H_0 (e.g., Benjamini, Drai, Elmer, Kafkafi, & Golani, 2001). With the false discovery rate method, the probability of a Type II error is smaller than with the sequential Bonferroni correction but this comes at the expense of a higher probability of a Type I error.

12.6 Consequences: Sequential Bonferroni Applied to 60 Published Articles

In our hypothetical example, none of the null hypotheses were rejected after the sequential Bonferroni correction (see Table 12.3), whereas, without any correction, all null hypotheses were rejected (see Table 12.1). One may argue that this example is extreme and contrived, and that such dramatic changes in conclusions will not regularly occur in the empirical literature. We addressed this claim quantitatively as follows. For each of the six journals listed in Table 12.2, we randomly chose 10 articles that reported one or more multiway ANOVAs. For these 60 papers, we re-evaluated the results (see `www.aojcramer.com` for R code (R Core Team, 2012) to perform the sequential Bonferroni procedure) after applying the sequential Bonferroni correction. The results paint a grim picture: in 75% (45/60) of the cases, one or more p values were no longer significant after correcting for multiple comparisons. That is, in the majority of cases, one or more conclusions are not substantiated by the corrected outcomes of the statistical analyses.

12.7 Conclusion

Our literature review showed that, across a total of 819 articles from six leading psychology journals, hardly any researchers have corrected for the multiple comparison problem that is an inherent property of multiway ANOVA. A reanalysis of a subset of 60 papers showed that the results of foregoing such correction procedures are worrying: Many conclusions reported in the literature may no longer hold after applying a correction procedure. The good news is that the sequential Bonferroni procedure (Hartley, 1955) is a simple, easy-to-use correction method that controls the α level, that is, the probability of falsely rejecting true null hypotheses.

One disadvantage of the sequential Bonferroni procedure is conceptual: The significance of a particular factor depends on the significance of other, unrelated factors. For instance, the main effect for G reported in Table 12.1 and Table 12.3 is associated with $p = 0.0329$. If the effects for the other two factors (i.e., $E \times G$ and E) had been more compelling (e.g., $p < 0.01$), the final and third test for G would have been conducted with $\alpha = 0.05$ level, and the result would have been labeled significant. This dependence on the results of unrelated tests may strike one as odd. However, such oddities are a general characteristic of p values (e.g., Wagenmakers, 2007). Note that the regular Bonferroni correction does not have this conceptual drawback, but it is inferior in terms of power.

We do not wish to suggest that the sequential Bonferroni procedure is the only or even the best procedure to correct for multiple comparisons in the multiway ANOVA. As noted before, several other procedures exist. These alternative procedures differ in terms of the balance between safeguarding against Type I and Type II errors. On the one hand, it is crucial to control the probability of incorrectly rejecting the H_0 (i.e., the Type I error). On the other hand, it is also important to minimize the Type II error, that is, to maximize power (Button et al., 2013).

So what is a researcher to do? Using the sequential Bonferroni correction, one is safeguarded against Type I errors at the expense of failing to detect some effects that are true. Using the false discovery rate procedure, one obtains more power, but relinquishes strict control over the Type I error rate. We encourage researchers to report the results from multiple correction methods: this allows readers to assess the robustness of the statistical evidence. Of course, the royal road to obtaining sufficient power is not to choose lenient correction methods; instead, one is best advised to increase sample size.

In sum, we have shown that multiway ANOVA harbors a multiplicity problem that has been ignored in empirical practice. The problem can be addressed in a straightforward fashion by various correction procedures, such as the sequential Bonferroni correction. Another fruitful avenue to remedy the problem is the *pre-registration* of the hypotheses and the corresponding statistical analyses (e.g., Chambers, 2013; Chambers et al., 2013; de Groot, 1969; Goldacre, 2009; Wagenmakers et al., 2012; Wolfe, 2013). Pre-registration forces researchers to consider beforehand the exact hypotheses of interest. In doing so, as we have argued earlier, one engages in confirmative hypothesis testing (i.e., the confirmative multiway ANOVA). "Fishing expeditions", however, in which one has no a priori hypotheses, come at a rather high price: one will have to use some sort of correction procedure to adjust the α level when engaging in an explorative multiway ANOVA.

The view on differential uses of the multiway ANOVA (i.e., explorative vs. confirmative) hinges on the specific definition of what constitutes a family of hypotheses, and we acknowledge that other definitions of such a family exist. However, in our view, the intentions of the researcher (explorative hypothesis *formation* vs. confirmative hypothesis *testing*) play a crucial part in determining the size of the family of hypotheses. It is vital to recognize the multiplicity inherent in the explorative multiway ANOVA and correct the current unfortunate state of affairs[1]; the alternative is to accept that our findings are much less compelling than advertised.

[1] Fortunately, some prominent psychologists such as Dorothy Bishop, are acutely aware of the multiple comparison problem in multiway ANOVA and urge their readers to rethink their analysis strategies: `http://deevybee.blogspot.co.uk/2013/06/interpreting-unexpected-significant.html`.

Chapter 13

Summary and Future Directions

In what follows, I will summarize the results and the main conclusions presented in this dissertation, accompanied by suggestions about avenues for future development.

13.1 The Analysis of Response Time Distributions

Cognitive Interpretation of the Ex-Gaussian and Shifted-Wald Parameters

In Chapter 2, I investigated the cognitive interpretation of parameters of the ex-Gaussian and shifted Wald distributions. A growing number of researchers use descriptive distributions such as the ex-Gaussian and the shifted Wald to summarize response time (RT) data for speeded two-choice tasks. Some of these researchers also assume that the parameters of these distributions uniquely correspond to specific cognitive processes. We studied the validity of this cognitive interpretation by examining the extent to which the ex-Gaussian and shifted Wald parameters could be associated with the kind of psychological processes that are hypothesized by the Ratliff diffusion model (Ratcliff, 1978), a successful model whose parameters have a well-established cognitive interpretation (e.g.,Voss et al., 2004). In a simulation study, we fit the ex-Gaussian and shifted Wald distributions to data generated from the diffusion model by systematically varying its parameters across a wide range of plausible values. In an empirical study, we fit the two descriptive distributions to published data that featured manipulations of task difficulty (i.e., corresponding to changes in drift rate v), response caution (i.e., boundary separation a), and a priori bias (i.e., starting point z; Wagenmakers, Ratcliff, et al., 2008). The results were clear-cut: In the context of a two-choice task, the ex-Gaussian and shifted Wald parameters cannot be associated uniquely with the parameters of the diffusion model. We concluded that researchers should resist temptation to interpret changes in the ex-Gaussian and shifted Wald parameters in terms of cognitive processes. A possible reason for this unfortunate result may be that the descriptive distributions do not take response accuracy into account. Without any knowledge of response accuracy, it is very difficult to distinguish between the effects of task difficulty (or subject ability) and the effects of response caution.

Bayesian Estimation of Stop-Signal Reaction Time Distributions

In Chapter 3, I introduced a Bayesian parametric approach for the estimation of stopping latencies in the stop-signal paradigm. The stop-signal paradigm is frequently used to study response inhibition. In this paradigm, participants perform a two-choice RT task where, on some of the trials, the primary task is interrupted by a stop signal that prompts participants to withhold their response. The dependent variable of interest is the latency of the unobservable stop response (stop-signal RT or SSRT). Based on the horse race model (Logan & Cowan, 1984), several methods have been developed to estimate SSRTs. Unfortunately, none of these approaches allow for the accurate estimation of the entire distribution of SSRTs. Here we presented a Bayesian parametric approach (BPA) that addresses this limitation. Our method is based on the assumptions of the horse race model and rests on the concept of censored distributions. The BPA treats response inhibition as a censoring mechanism, where the distribution of RTs on the primary task (go RTs) is censored by the distribution of SSRTs. The method assumes that go RTs and SSRTs are ex-Gaussian distributed and uses Markov chain Monte Carlo sampling (MCMC; Gamerman & Lopes, 2006; Gilks et al., 1996) to obtain posterior distributions for the model parameters. The BPA can be applied to individual as well as hierarchical data structures. We presented the results of a number of parameter recovery and robustness studies and applied the new approach to published data from a stop-signal experiment. The WinBUGS (Lunn et al., 2012) and WBDev (Wetzels, Lee, & Wagenmakers, 2010) codes that implement the BPA are available online.

Releasing the BEESTS

In Chapter 4, I presented BEESTS, an efficient and user-friendly software implementation of the BPA introduced in Chapter 3. BEESTS comes with an easy-to-use graphical user interface and provides users with summary statistics of the posterior distribution of the parameters as well various diagnostic tools to assess the quality of the parameter estimates. The software is open source and runs on Windows and OS X operating systems. BEESTS relies on Python for parameter estimation (Patil et al., 2010; Wiecki et al., 2013) and on R (R Core Team, 2012) for the post-processing of the output. For computational speed, the likelihood functions are coded in Cython (Behnel et al., 2011). We illustrated the use of the individual and the hierarchical BEESTS analysis with a published stop-signal data set.

Future Directions

The first part of the dissertation focused on modeling RTs —observed and unobserved— in two-choice tasks with descriptive RT models, such as the ex-Gaussian and the shifted Wald distributions. First, I showed that the parameters of these descriptive distributions should not be interpreted in terms of the cognitive processes assumed by the diffusion model. However, the parameters of the ex-Gaussian and shifted-Wald distributions need not be considered in isolation. Unlike the individual parameters, certain —possibly nonlinear— combinations of the ex-Gaussian or shifted Wald parameters might map uniquely onto parameters of the diffusion model. This possibility awaits further investigation.

Second, I showed that —despite its lack of theoretical underpinning in terms of specific cognitive processes— the ex-Gaussian distribution may be successfully used to describe and model the distribution of stopping latencies in the stop-signal paradigm. However, if the processes underlying response inhibition are of interest, cognitive models, such as the interactive race model (Boucher et al., 2007), the Linear Approach to Threshold with Ergodic Rate (Carpenter, 1981; Carpenter & Williams, 1995; Hanes & Carpenter, 1999), or a modified version of the linear ballistic accumulator model (Brown & Heathcote, 2008) are the appropriate choice. For other alternatives, see Logan et al. (2014). Nevertheless, the ex-Gaussian based BPA is certainly not a redundant tool in the growing arsenal of techniques targeted at estimating stopping latencies. To the contrary: the BPA may be used to aid model development and evaluate the predictions of competing process models beyond the level of mean SSRT.

Third, I introduced BEESTS, a user-friendly software implementation of the ex-Gaussian based BPA. In order to assess the absolute goodness-of-fit of the BPA, BEESTS relies on posterior predictive model checks. To formalize the model checks, BEESTS computes posterior predictive p values using the median of the observed signal-respond RTs and the median of the signal-respond RTs predicted by the joint posterior of the model parameters. As I repeatedly stressed throughout the dissertation, this approach is not ideal; adequate analysis of RT data should not only focus on the median, but should consider the shape of the entire RT distribution. Accordingly, the posterior predictive model checks in BEESTS should preferably compare the entire distribution of observed signal-respond RTs to the distribution of signal-respond RTs predicted by the model. Unfortunately, this is easier said than done. The assessment of goodness-of-fit using the entire distribution of signal-respond RTs does not only involve the formal comparison of nonparametric distributions; it also involves the comparison of a single observed signal-respond RT distribution to multiple —often thousands of— predicted signal-respond RT distributions. Note also that BEESTS only allows user to assess the *absolute* goodness-of-fit of the model. The assessment of the *relative* goodness-of-fit of the BPA involves the specification of an alternative model and the application of formal Bayesian model selection. The improvement of the posterior predictive checks and the implementation of formal Bayesian model selection methods require further development.

13.2 Multinomial Processing Tree Models

Crossed-Random Effects Multinomial Processing Tree Models

In Chapter 5, I focused on a Bayesian approach that accounts for parameter heterogeneity as a result of differences between participants as well as items in multinomial processing tree (MPT) models. MPT models are theoretically motivated stochastic models for the analysis of categorical data. Traditionally, statistical analysis for MPT models is carried out on aggregated data, assuming homogeneity in participants and items (Hu & Batchelder, 1994). However, in many applications it is reasonable to assume that the model parameters differ both between participants and items. We should then treat both participant and items effects as random and base statistical inference on unaggregated data. Here we introduced a hierarchical crossed-random effects extension of the pair-clustering model (Batchelder & Riefer, 1980), one of the most extensively studied MPT models for the analysis of free recall data. Our approach assumed that participant and item effects combine additively on the probit scale and postulated multivariate normal distributions for the random effects. We provided a WinBUGS implementation of the crossed-random effects pair-clustering model and an application to novel experimental data that featured the manipulation of word frequency.

Model Comparison for Multinomial Processing Tree Models

In Chapter 6, I discussed various procedures for model comparison in the context of MPT models. A careful model comparison procedure involves both qualitative and quantitative elements. Important qualitative elements include, for example, plausibility, consistency with known behavioral phenomena, and coherence of the underlying assumptions. The single most important quantitative element of model comparison relates to the tradeoff between parsimony and goodness-of-fit (Pitt & Myung, 2002). The topic of quantitative model comparison has received —and continues to receive— considerable attention in the field of statistics. Here we focused on two popular information criteria, the AIC ("an information criterion", Akaike, 1973) and the BIC ("Bayesian information criterion", G. Schwarz, 1978), on the Fisher information approximation of the minimum description length principle (MDL; Grünwald, 2007), and on Bayes factors as obtained from importance sampling (Hammersley & Handscomb, 1964). We first provided a general description of the procedures and then applied them to three competing MPT models of memory interference (Wagenaar & Boer, 1987). The R codes (R Core Team, 2012) that implement the MDL and Bayes factor calculations are available online.

Future Directions

The second part of the dissertation focused on parameter estimation and model selection in MPT models. First, I introduced a hierarchical crossed-random effects extension to MPT models that assumes the additivity of participant and item effects. Although I focused exclusively on the pair-clustering model, the crossed-random effects approach may be extended to many other MPT models. The issue of model identification, however, must be carefully considered. The present approach deals only with models that are identified for each participant after collapsing across items and for each item after collapsing across the participants. In paradigms where items are restricted to certain category systems, model identification remains an issue that requires further development.

Second, I reviewed a number of procedures for model comparison in MPT models, with special emphasis on Bayes factors obtained from importance sampling. The chapter exclusively focused on MPT models that assume homogeneity in participants and items. For (crossed-) random effects hierarchical MPT models, however, the computation of Bayes factors using importance sampling is computationally infeasible. The development of more sophisticated model selection methods that are appropriate for hierarchical models is presently an active area of research; preliminary results indicate that reversible jump MCMC (Green, 1995) is a promising tool for the computation of Bayes factors in hierarchical MPT models.

13.3 Correlations, Partial Correlations, and Mediation

Power to Reject the Hypothesis of Perfect Correlation

In Chapter 7, I examined the power to reject the hypothesis of perfect correlation in the context of higher-order structural equation models (SEM). In higher-order factor models, general intelligence (g) is often found to correlate perfectly with lower-order common factors, suggesting that g and some well-defined cognitive ability, such as working memory, may be identical. However, the results of studies that addressed the equivalence of g and lower-order factors are inconsistent. We suggested that this inconsistency may partly be attributable to the lack of statistical power to detect the distinctiveness of the two factors. We therefore investigated the power to reject the hypothesis that g and a lower-order factor are perfectly correlated using artificial datasets, based on realistic parameter values and on the results of selected publications. The results of the power analyses indicated that power was substantially influenced by the effect size and the number and the reliability of the indicators. The examination of published studies revealed that most case studies that reported a perfect correlation between g and a lower-order factor were severely underpowered, with power coefficients rarely exceeding 0.30. We concluded by emphasizing the importance of considering power in the context of identifying g with lower-order factors. The R code for the power calculation is available online.

Bayesian Correction for Attenuated Correlations

In Chapter 8, I discussed a Bayesian method for correcting the correlation coefficient for the uncertainty of the observations. The Pearson product-moment correlation coefficient can be severely underestimated when the observations are subject to measurement noise. Various approaches exist to correct the estimation of the correlation in the presence of measurement error, but none are routinely applied in psychological research. Here we outlined a Bayesian correction method for the attenuation of correlations proposed by Behseta et al. (2009) that is conceptually straightforward and easy to apply. We illustrated the Bayesian correction with two empirical data sets; in each data set, we first estimated posterior distributions for the uncorrected and corrected correlation coefficient and then computed Bayes factors to quantify the evidence that the data provided for the presence of an association. We demonstrated that correcting for measurement error can substantially increase the correlation between noisy observations. The WinBUGS and R codes that implement the Bayesian correction method and the Bayes factor calculations are available online.

A Default Bayesian Mediation Test

In Chapter 9, I described a default Bayesian hypothesis test for mediation. In order to quantify the relationship between multiple variables, researchers often carry out a mediation analysis. In such an analysis, a mediator (e.g., knowledge of healthy diet) transmits the effect from an independent variable (e.g., classroom instruction on healthy diet) to a dependent variable (e.g., consumption of fruits and vegetables). Almost all mediation analyses in psychology use frequentist estimation and hypothesis testing techniques. A recent exception is Yuan and MacKinnon (2009), who outlined a Bayesian parameter estimation procedure for mediation analysis. Here we completed the Bayesian alternative to frequentist mediation analysis by specifying a default Bayesian hypothesis test based on the Jeffreys-Zellner-Siow approach (Rouder et al., 2009). We further extend the default test by allowing the computation of directional or one-sided Bayes factors, using MCMC techniques implemented in JAGS (Plummer, 2009). All Bayesian tests are implemented in the `R` package `BayesMed`.

Future Directions

The third part of the dissertation focused on estimating and testing observed and unobserved (partial) correlations. First, I showed that the majority of studies that use SEM to evaluate the hypothesis of perfect correlation between g and a lower-order factor are underpowered. In contrast to previous chapters, here I relied on classical p value-based hypothesis testing. Second, I moved back to the domain of Bayesian inference, and described a method for correcting observed correlations for the uncertainty of the observations. Moreover, I illustrated a straightforward Bayesian procedure for testing the presence of a correlation using Bayes factors obtained with the Savage-Dickey density ratio method (Dickey & Lientz, 1970). The Bayesian correction method can be viewed as a simple Bayesian structural equation model with two latent variables, each with a single indicator. Third, I stayed within the Bayesian framework but moved away from latent variables, and described a default Bayesian hypothesis test for mediation. The mediation analysis relies on default Bayes factors for correlations and partial correlations (Wetzels & Wagenmakers, 2012).

Possible extensions for the techniques presented above are straightforward. Hypothesis test for assessing (perfect) correlations in SEMs may be implemented in a Bayesian setting. The simple Bayesian attenuation correction may be extended to (higher-order) SEMs featuring multiple latent factors and indicators. The Bayesian mediation test may be adapted to handle latent variables. These extensions all rely on Bayesian parameter estimation and model selection in SEMs.

Bayesian parameter estimation can be easily implemented in standard statistical software, such as WinBUGS. Also, recent versions of Mplus (i.e, popular software for fitting and testing SEMs; Muthén & Asparouhov, 2012) support Bayesian parameter estimation and posterior predictive assessment of goodness-of-fit. Formal Bayesian model selection methods are also available for SEMs. The computation of Bayes factors, however, relies on sophisticated sampling methods, such as path sampling (S.-Y. Lee, 2007; Song & Lee, 2012) and reversible jump MCMC (Lopes & West, 2004), and is not yet implemented in standard statistical software. Hence it is all but impossible for most research psychologists to take advantage of the Bayesian developments. A notable exception is the work of van de Schoot, Hoijtink, and Deković (2010) that uses Mplus output to compute Bayes factors for inequality-constrained hypotheses. The development and implementation of Bayesian model selection in SEMs is an active and exciting area of research.

13.4 Improving Research Practice

A Preregistered Adversarial Collaboration

In Chapter 10, I introduced a novel variant of proponent-skeptic collaboration that focused on the association between horizontal eye movements and episodic memory. A growing body of research suggests that horizontal saccadic eye movements facilitate the retrieval of episodic memories in free recall and recognition memory tasks. Nevertheless, a minority of studies have failed to replicate this effect. Here we attempted to resolve the inconsistent results by introducing a novel variant of proponent-skeptic joint research. The proposed approach combined the features of adversarial collaboration (Kahneman, 2003) and purely confirmatory preregistered research (Wagenmakers et al., 2012). Prior to data collection, the adversaries reached consensus on an optimal research design, formulated their expectations, and agreed to submit the findings to an academic journal regardless of the outcome. To increase transparency and to secure the purely confirmatory nature of the investigation, the two parties set up a publicly available adversarial collaboration agreement that detailed the proposed design and all foreseeable aspects of the data analysis. As anticipated by the skeptics, a series of Bayesian hypothesis tests indicated that horizontal eye movements did not improve free recall performance. The skeptics suggested that the non-replication may partly reflect the use of suboptimal and questionable research practices in earlier eye movement studies. The proponents countered this suggestion and used a p-curve analysis to argue that the effect of horizontal eye movements on explicit memory does not merely reflect selective reporting. The preregistered adversarial collaboration agreement and the data are available on the Open Science Framework.

Bayes Factors, p Values, and Effect Sizes

In Chapter 11, I presented a comparison of the statistical evidence provided by p values, effect sizes, and default Bayes factors. Statistical inference in psychology has traditionally relied heavily on p value significance testing. This approach to drawing conclusions from data, however, has been widely criticized, and two types of remedies have been advocated. The first proposal is to supplement p values with complementary measures of evidence such as effect sizes. The second is to replace inference with Bayesian measures of evidence such as the Bayes factor. Here we provided a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. The comparison yielded two main results. First, although p values and default Bayes factors almost always agreed about what hypothesis is better supported by the data, the measures often disagreed about the strength of this support; 70% of the p values that fell between .01 and .05 correspond to Bayes factors that indicate that the data are no more than three times more likely under the alternative hypothesis than under the null hypothesis. Second, effect sizes can provide additional evidence to p values and default Bayes factors. We concluded that the Bayesian approach is comparatively prudent, preventing researchers from overestimating the evidence in favor of an effect.

Sequential Bonferroni Correction for Multiple Comparisons

In the twelfth and final chapter, I focused on the sequential Bonferroni correction in multiway analysis of variance (ANOVA). Many empirical researchers do not realize that the common multiway ANOVA harbors a multiple comparison problem. In the case of two factors, three separate null hypotheses are subject to test (i.e., two main effects and one interaction). Consequently, the probability of a Type I error is 14% rather than 5%. Here we described the multiple comparison problem and demonstrated that researchers seldom correct for it. We then illustrated the use of the sequential Bonferroni (Hartley, 1955) correction —one of several correction procedures— and showed that its application alters at least one of the substantive conclusions in 45 out of 60 articles considered. We argued that preregistration of the hypotheses provides an alternative method to combat the multiplicity problem in multiway ANOVA.

Future Directions

The fourth and final part of the dissertation focused on suboptimal research practices in psychology. First, I focused on questionable research practices and the replication crisis in psychology, and advocated the use of preregistered adversarial collaborations for scientific conflict resolution. I described a proponent-skeptic collaboration on the beneficial effects of horizontal eye movement on memory performance and illustrated how the Bayes factor can be used to quantify evidence *in favor of* the null hypothesis. Second, I showed that although p values and Bayes factors almost always agree about which hypothesis is better supported by the data, p values often overestimate evidence against the null hypothesis. Third, I revisited the frequentist approach, and described a hidden multiplicity problem in multiway ANOVAs and showed that the application of sequential Bonferroni correction often alters conclusions drawn from ANOVA designs.

My main goal was to highlight the advantages of adopting the Bayesian approach in original as well as replication research. Bayesian inference —as opposed to frequentist inference— does not depend on the intention with which the data were collected, it can be used to quantify evidence in favor of the null hypothesis, and enables researchers to assess what they would like to know in the first place when they engage in hypothesis testing, that is, the probability of the data under one hypothesis relative to the other. Various user-friendly default Bayesian procedures are now available for t tests (Rouder et al., 2009; Wetzels et al., 2009), ANOVAs (Masson, 2011; Wetzels et al., 2012), correlations and partial correlations, (Wetzels & Wagenmakers, 2012), mediation (Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, submitted) and regression analyses (Liang et al., 2008; Rouder & Morey, 2012). Despite considerable progress over the past decade, the user-friendly Bayesian implementation of many popular techniques, such as structural equation models and contingency tables, awaits further development. Similarly, the further development and the implementation of Bayesian correction methods for multiple comparison (e.g., Berry & Hochberg, 1999; Marchini, Howie, Myers, McVean, & Donnelly, 2007; Scott & Berger, 2006, 2010) are exciting research areas that will hopefully receive due attention from the statistical community in the near future.

Part V

Appendices

Appendix A

Appendix to Chapter 2: "Psychological Interpretation of the Ex–Gaussian and Shifted Wald Parameters: A Diffusion Model Analysis"

A.1 The Distribution of the Diffusion Model Parameter Values

Figure A.1 presents histograms of the best–fitting diffusion model parameter values and the corresponding z/a and s_z/a ratios found in 23 applications of the diffusion model. The histograms are based on the the parameter values reported for each experimental condition of the 23 articles. The exact parameter values, including references, are available as supplemental materials at `http://dora.erbe-matzke.com/publications.html`.

A.2 Results for the Diffusion Model Trial–to–Trial Variability Parameters

This appendix shows how the ex–Gaussian and shifted Wald parameters change as a function of the manipulation of the trial–to–trial variability in drift rate η, the trial–to–trial variability in starting point s_z, and the trial–to–trial variability in nondecision time s_t parameters of the diffusion model. Table A.1 gives a summary of the associations between the ex–Gaussian and shifted Wald parameter and the diffusion model variability parameters. Figure A.2 and Figure A.3 then show the detailed changes in the ex–Gaussian and shifted Wald parameters as a function of changes in the diffusion model variability parameters.

Ex–Gaussian Parameters

With respect to trial–to–trial variability in drift rate η, Figure A.2a shows that both μ and σ decrease as η increases. In contrast, τ increases for low values of η and decreases for high values of η. However, the changes in the three ex–Gaussian parameters are all extremely small. Turning to trial–to–trial variability in starting point s_z, Figure A.2b shows that both σ and τ increase as s_z increases, but in contrast, μ decreases with increasing s_z. However, the changes in the three

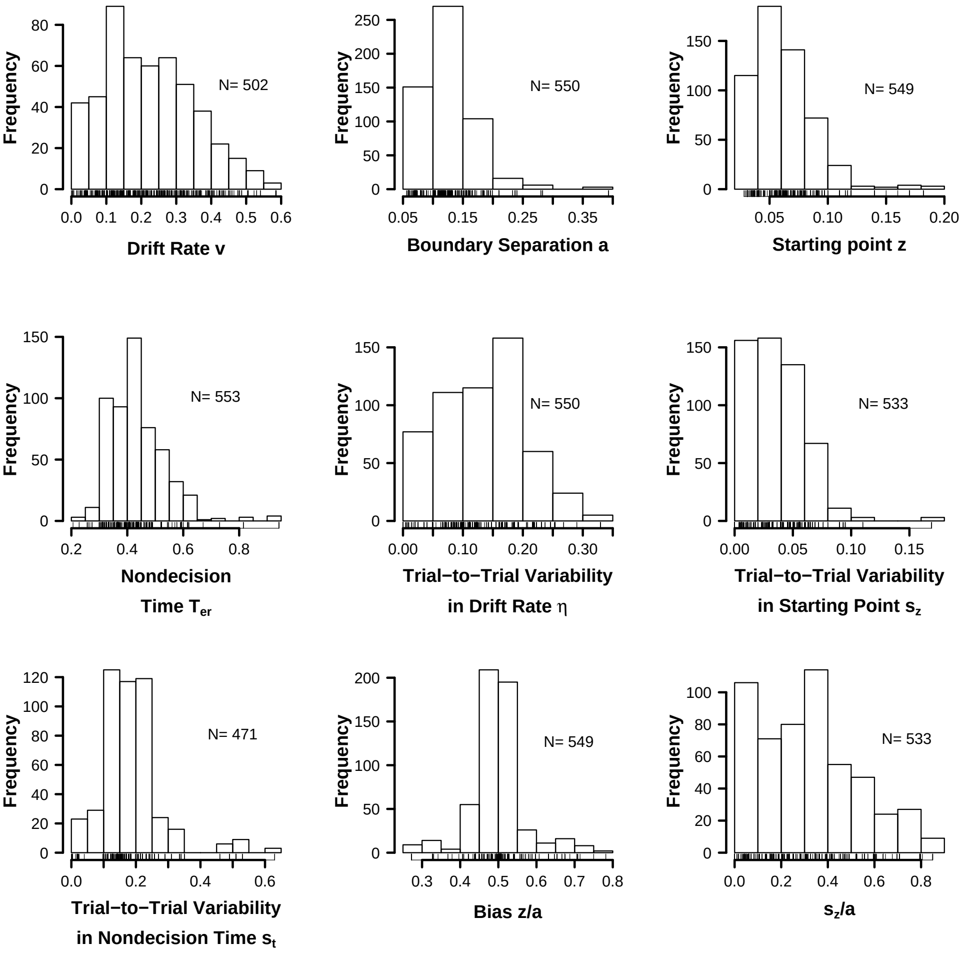


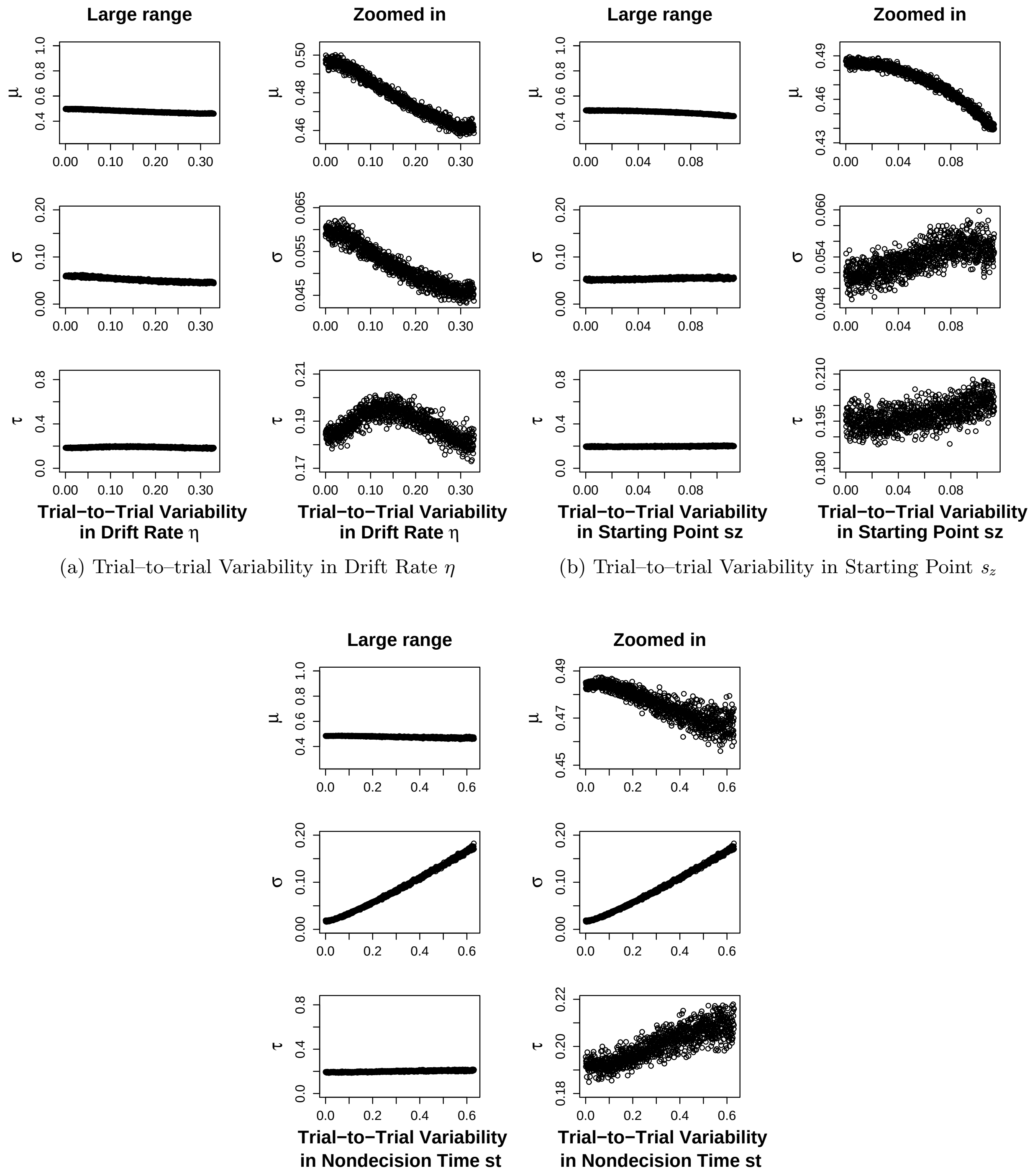
Figure A.1 *Histograms of the diffusion model parameter values.*

ex–Gaussian parameters are all negligible. With respect to trial–to–trial variability in nondecision time s_t, Figure A.2c shows that both σ and τ parameters as s_t increases, whereas μ decreases with increasing s_t. Note that σ is the only parameter that is substantially influenced by s_t. In fact, σ changes substantially more as function of s_t than as a function of any other diffusion model parameter.

To summarize, these results further support the conclusion that the two most important parameters of the ex–Gaussian distribution, μ and τ, do not correspond uniquely to parameters of the diffusion model. Neither of these ex–Gaussian parameters is influenced substantially by any of the variability parameters of the diffusion model. In contrast, σ seems to be uniquely associated with s_t, the parameter for trial–to–trial variability in nondecision time.

Shifted Wald Parameters

With respect to trial–to–trial variability in drift rate η, Figure A.3a shows that both α and γ decrease as η increases, but in contrast, θ increases with increasing η. However, the changes in the three shifted Wald parameters are all extremely small. Turning to trial–to–trial variability in



(a) Trial–to–trial Variability in Drift Rate η

(b) Trial–to–trial Variability in Starting Point s_z

(c) Trial–to–trial Variability in Nondecision Time s_t

Figure A.2 *Changes in the ex–Gaussian parameters μ, σ, and τ as a function of systematic changes in the diffusion model parameters trial–to–trial variability in drift rate η (panel a), trial–to–trial variability in starting point s_z (panel b), and trial–to–trial variability in nondecision time s_t (panel c).* The left-hand figures in each panel plot the results on scales ranging from the minimum to the maximum values of the ex–Gaussian parameters found across all simulations. The right-hand figures in each panel plot the same results on scales ranging from the minimum to the maximum values of the ex–Gaussian parameters found for the manipulation of the given diffusion model parameter.

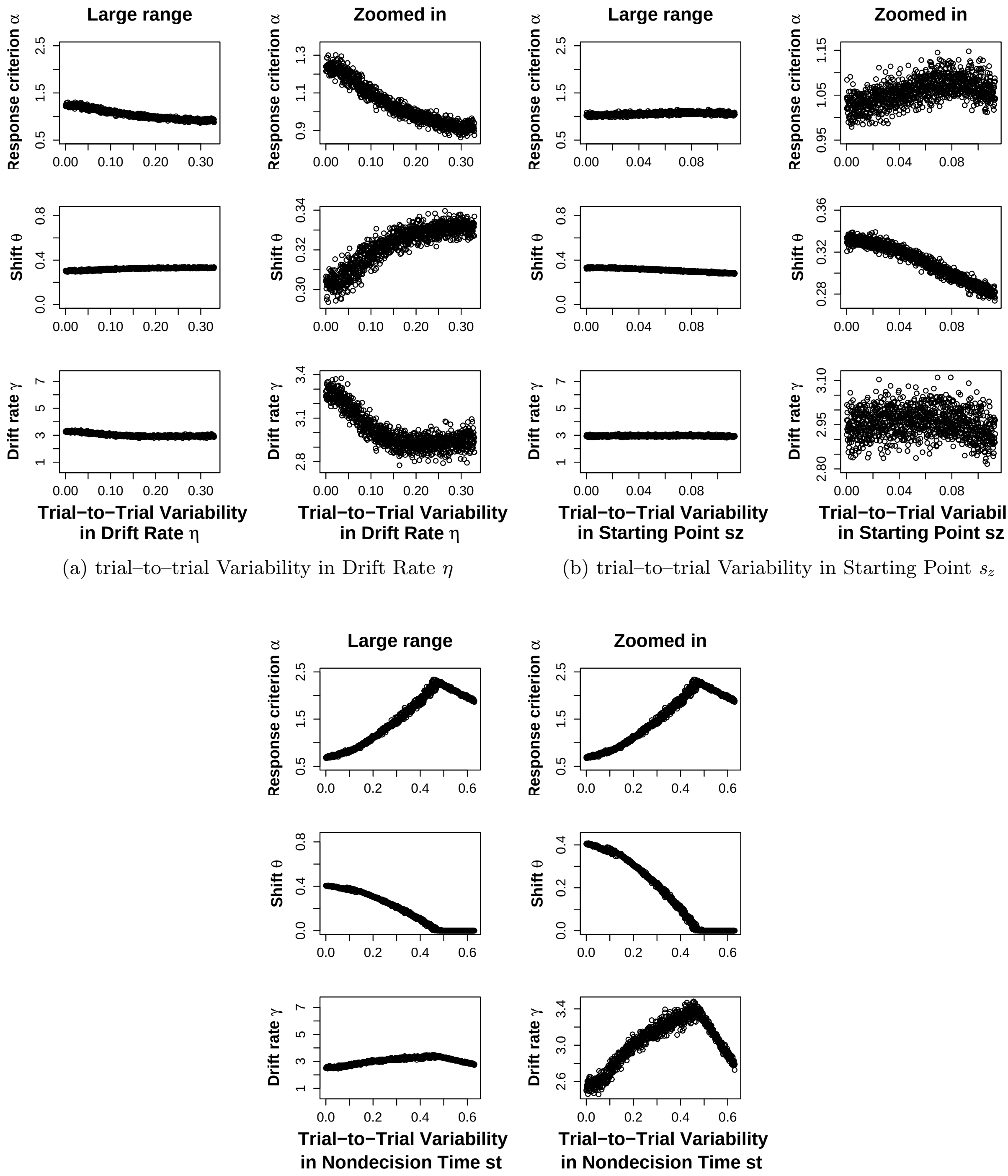
Table A.1 The Associations Between the Parameters of the Ex–Gaussian and Shifted Wald Distributions and the Variability Parameters of the Diffusion Model.

		Diffusion model parameters		
		η	s_z	s_t
	μ	−	−	−
Ex–Gaussian	σ	−	+	++
	τ	+/−	+	+
	α	−	+	++/− −
Shifted Wald	θ	+	−	− −
	γ	−	×	+/−

Note. ++, substantial positive association; +, weak positive association; − −, substantial negative association; −, weak negative association; ×, no association; η, variability in drift rate; s_z, variability in starting point; s_t, variability in nondecision time.

starting point s_z, Figure A.3b shows that γ is unaffected by changes in s_z, whereas α increases and θ decreases with increasing s_z. However, the changes in both α and θ are extremely small. With respect to trial–to–trial variability in nondecision time s_t, Figure A.3c shows that both α and γ increase for low and intermediate values of s_t and then decrease for high values. In contrast, θ decreases for low and intermediate values of s_t and equals 0 for high values. Although α changes more than either θ or γ, the change in θ is also substantial. Note that α changes just as much as a function of s_t than as a function of boundary separation a.

To summarize, these results further support the conclusion that the shifted Wald parameters do not correspond uniquely to parameters of the diffusion model. The γ parameter is not influenced substantially by any of the variability parameters of the diffusion model. In contrast, both α and θ are substantially influenced by s_t, the trial–to–trial variability in nondecision time. In addition to the influence of the key diffusion model parameters, changes in α and θ can therefore also reflect the influence of s_t.



(a) trial–to–trial Variability in Drift Rate η

(b) trial–to–trial Variability in Starting Point s_z

(c) trial–to–trial Variability in Nondecision Time s_t

Figure A.3 *Changes in the shifted Wald parameters α, θ, and γ as a function of systematic changes in the diffusion model parameters trial–to–trial variability in drift rate η (panel a), trial–to–trial variability in starting point s_z (panel b), and trial–to–trial variability in nondecision time s_t (panel c).* The left-hand figures in each panel plot the results on scales ranging from the minimum to the maximum values of the shifted Wald parameters found across all simulations. The right-hand figures in each panel plot the same results on scales ranging from the minimum to the maximum values of the shifted Wald parameters found for the manipualtion of the given diffusion model parameter.

Appendix B

Appendix to Chapter 3: "Bayesian Parametric Estimation of Stop-Signal Reaction Time Distributions"

B.1 WinBUGS Script

Individual Bayesian Parametric Approach (BPA) Model

The WinBUGS script for the individual BPA is as follows:

```
model
{

   # Priors for parameters
   mu_go ~ dunif(1,1000)
   sigma_go ~ dunif(1,300)
   tau_go ~ dunif(1,300)

   mu_stop ~ dunif(1,600)
   sigma_stop ~ dunif(1,250)
   tau_stop ~ dunif(1,250)

   # Go RTs come from an ex-Gaussian distribution
   for (g in 1:n.gort){
      go_rt[g] ~ ExGaussian(mu_go, sigma_go, tau_go)
   }

   # Signal respond trials; signal-respond RTs (srrt) at each SSD come from
   # a censored ex-Gaussian distribution
   # (see first part of Equation 14.)
   for (d in 1:end_SR){
      for (r in 1:n.srrt[d]){
         srrt[d,r] ~ CensoredExGaussian_SR(mu_go, sigma_go, tau_go,
                           mu_stop, sigma_stop, tau_stop, ssd_SR[d])
      }
   }

   # Signal inhibit trials; Succesful inhibitions come from a censored ex-Gaussian
   # distribution (see second part of Equation 14.)
   for (h in 1:end_I){
      for (i in 1:n.inhibitions[h]){
         zeros[h,i] <- 0
         zeros[h,i] ~ dpois(phi[h,i])
         phi[h,i] <- - intg[h]
      }
      # Compute integral in Equation 14 using Simpson’s rule of numerical integration
      # The first and the second arguments define the limits of integration,
      # and the third argument defines the number of subintervals used for
      # computing the integral.
      intg[h] <- CensoredExGaussian_I(1, 6000, 2000, mu_go, sigma_go, tau_go,
                                     mu_stop, sigma_stop, tau_stop, ssd_I[h])
   }
}
```

The ExGaussian and CensoredExGaussian_SR distributions and the CensoredExGaussian_I function are implemented with the WinBUGS Development Interface (WBDev; Lunn, 2003). For a WBDev tutorial for social scientists, see Wetzels, Lee, and Wagenmakers (2010). The WinBUGS and WBDev scripts are available in the supplemental materials at http://dora.erbe-matzke.com/publications.html. For computational reasons, the indefinite integral in Equation 3.14 is replaced by a definite integral (i.e., CensoredExGaussian_I) with limits of integration well beyond the range of stop-signal reaction times that may be encountered in the stop-signal paradigm.

Hierarchical Bayesian Parametric Approach (BPA) Model

The WinBUGS script for the hierarchical BPA is as follows:

```
model
{
   # Priors for the group-level parameters
   # The I(0,) construct denotes distributional censoring, with a lower bound of 0,
   # and an upper bound of infinity
   mu_mu_go ~ dnorm(500,0.0001)I(0,)
   lambda_mu_go <- 1/pow(sigma_mu_go,2)
   sigma_mu_go ~ dunif(0,300)

   mu_sigma_go ~ dnorm(100,0.001)I(0,)
   lambda_sigma_go <- 1/pow(sigma_sigma_go,2)
   sigma_sigma_go ~ dunif(0,200)

   mu_tau_go ~ dnorm(80,0.001)I(0,)
   lambda_tau_go <- 1/pow(sigma_tau_go,2)
   sigma_tau_go ~ dunif(0,200)

   mu_mu_stop ~ dnorm(200,0.0001)I(0,)
   lambda_mu_stop <- 1/pow(sigma_mu_stop,2)
   sigma_mu_stop ~ dunif(0,200)

   mu_sigma_stop ~ dnorm(40,0.001)I(0,)
   lambda_sigma_stop <- 1/pow(sigma_sigma_stop,2)
   sigma_sigma_stop ~ dunif(0,100)

   mu_tau_stop ~ dnorm(30,0.001)I(0,)
   lambda_tau_stop <- 1/pow(sigma_tau_stop ,2)
   sigma_tau_stop ~ dunif(0,100)

   # C has to be large enough so that all phi[s,k,n] are positive
   # C <- 10000
   # Participant loop
   for (j in 1:n.subjects){

      # Go RTs come from an ex-Gaussian distribution
      for (g in 1:n.gort){
         go_rt[g,j] ~ ExGaussian(mu_go[j], sigma_go[j], tau_go[j])
      }

      # Signal respond trials; signal-respond RTs (srrt) at each SSD come from
      # a censored ex-Gaussian distribution
      # (see first part of Equation 14.)
      for (d in 1:end_SR[j]){
         for (r in 1:n.srrt[d,j]){
            srrt[d,r,j] ~ CensoredExGaussian_SR(mu_go[j], sigma_go[j], tau_go[j],
                             mu_stop[j], sigma_stop[j], tau_stop[j], ssd_SR[d,j])
         }
      }

      # Signal inhibit trials; Succesful inhibitions come from a censored ex-Gaussian
      # distribution (see second part of Equation 14.)
      # The following code implements the zeros trick (see WinBUGS manual)
      # Because phi[s,k,n] is a Poisson mean, it should always be positive.
      # As a result, we may need to add constant C to ensure that all phi[s,k,n]
      # are positive
      for (h in 1:end_I[j]){
         for (i in 1:n.inhibitions[h,j]){
            zeros[h,i,j] <- 0
            zeros[h,i,j] ~ dpois(phi[h,i,j])
            phi[h,i,j] <- - intg[h,j] #+C
         }

         # Compute integral in Equation 14 using Simpson's rule of numerical
         # integration. The first and the second arguments define the limits
         # of integration, and the third argument defines the number of
         # subintervals used for computing the integral.
         intg[h,j] <- CensoredExGaussian_I(1, 3000, 1000, mu_go[j], sigma_go[j],
                                           tau_go[j], mu_stop[j], sigma_stop[j],
                                           tau_stop[j], ssd_I[h,j])
      }

      # Individual parameters come from truncated normal distributions
      # The third argument specifies the truncation point
      mu_go[j] ~ TruncatedNormal(mu_mu_go, lambda_mu_go,0)
      sigma_go[j] ~ TruncatedNormal(mu_sigma_go,lambda_sigma_go,1)
      tau_go[j] ~ TruncatedNormal(mu_tau_go,lambda_tau_go,1)

      mu_stop[j]~ TruncatedNormal(mu_mu_stop, lambda_mu_stop,0)
      sigma_stop[j] ~ TruncatedNormal(mu_sigma_stop,lambda_sigma_stop,1)
      tau_stop[j] ~ TruncatedNormal(mu_tau_stop,lambda_tau_stop,1)
   }
}
```

The `TruncatedNormal` distribution is implemented with WBDev and is available in the supplemental materials.

Appendix C

Appendix to Chapter 4: "Release the BEESTS: Bayesian Estimation of Ex-Gaussian Stop-Signal Reaction Time Distributions"

C.1 Prior Distribution of the Model Parameters

This appendix presents the prior distributions of the model parameters in the BEESTS implementation of the Bayesian parametric approach (BPA). The name of each parameter as shown in the BEESTS output is in brackets.

Individual BPA

The priors for the go and stop parameters are uniform distributions, spanning a plausible but wide range of values. BEESTS relies on slightly more diffuse priors than the WinBUGS implementation of the BPA (see Matzke et al., 2013):

$$\begin{aligned}
\mu_{go} \text{ (mu\_go) } &\sim \text{Uniform}(0.001, 1000)\\
\sigma_{go} \text{ (sigma\_go) } &\sim \text{Uniform}(1, 500)\\
\tau_{go} \text{ (tau\_go) } &\sim \text{Uniform}(1, 500)\\
\mu_{stop} \text{ (mu\_stop) } &\sim \text{Uniform}(0.001, 600)\\
\sigma_{stop} \text{ (sigma\_stop) } &\sim \text{Uniform}(1, 350)\\
\tau_{stop} \text{ (tau\_stop) } &\sim \text{Uniform}(1, 350).
\end{aligned} \tag{C.1}$$

Hierarchical BPA

Individual Parameters

The hierarchical BPA assumes that the μ_{go}, σ_{go}, τ_{go}, μ_{stop}, σ_{stop}, and τ_{stop} parameters of each participant $j = 1, ..., J$ come from truncated normal group-level distributions. The group-level distributions are themselves characterized by a group mean (μ) and a group standard deviation (σ) parameter. The WinBUGS implementation relies on normal group-level distributions that are truncated only at the lower end, whereas BEESTS uses normal distributions that are truncated at the lower *and* the upper ends:

$$\begin{aligned}
\mu_{go_j} \text{ (mu\_go.subj) } &\sim \text{Normal}(\mu_{\mu_{go}}, \sigma_{\mu_{go}})[0.001, 1000] \\
\sigma_{go_j} \text{ (sigma\_go.subj) } &\sim \text{Normal}(\mu_{\sigma_{go}}, \sigma_{\sigma_{go}})[1, 500] \\
\tau_{go_j} \text{ (tau\_go.subj) } &\sim \text{Normal}(\mu_{\tau_{go}}, \sigma_{\tau_{go}})[1, 500] \\
\mu_{stop_j} \text{ (mu\_stop.subj) } &\sim \text{Normal}(\mu_{\mu_{stop}}, \sigma_{\mu_{stop}})[0.001, 600] \\
\sigma_{stop_j} \text{ (sigma\_stop.subj) } &\sim \text{Normal}(\mu_{\sigma_{stop}}, \sigma_{\sigma_{stop}})[1, 350] \\
\tau_{stop_j} \text{ (tau\_stop.subj) } &\sim \text{Normal}(\mu_{\tau_{stop}}, \sigma_{\tau_{stop}})[1, 350].
\end{aligned} \tag{C.2}$$

Group-Level Parameters

The priors for the group mean and group standard deviations are uniform distributions. Note that the WinBUGS implementation uses censored normal priors for the group-level means and relies on slightly less diffuse priors for the group-level standard deviations than BEESTS:

$$\begin{aligned}
\mu_{\mu_{go}} \text{ (mu\_go) } &\sim \text{Uniform}(0.001, 1000) \\
\sigma_{\mu_{go}} \text{ (mu\_go\_var) } &\sim \text{Uniform}(0.01, 300) \\
\mu_{\sigma_{go}} \text{ (sigma\_go) } &\sim \text{Uniform}(1, 500) \\
\sigma_{\sigma_{go}} \text{ (sigma\_go\_var) } &\sim \text{Uniform}(0.01, 200) \\
\mu_{\tau_{go}} \text{ (tau\_go) } &\sim \text{Uniform}(1, 500) \\
\sigma_{\tau_{go}} \text{ (tau\_go\_var) } &\sim \text{Uniform}(0.01, 200) \\
\mu_{\mu_{stop}} \text{ (mu\_stop) } &\sim \text{Uniform}(0.001, 600) \\
\sigma_{\mu_{stop}} \text{ (mu\_stop\_var) } &\sim \text{Uniform}(0.01, 300) \\
\mu_{\sigma_{stop}} \text{ (sigma\_stop) } &\sim \text{Uniform}(1, 350) \\
\sigma_{\sigma_{stop}} \text{ (sigma\_stop\_var) } &\sim \text{Uniform}(0.01, 200) \\
\mu_{\tau_{stop}} \text{ (tau\_stop) } &\sim \text{Uniform}(1, 350) \\
\sigma_{\tau_{stop}} \text{ (tau\_stop\_var) } &\sim \text{Uniform}(0.01, 200).
\end{aligned} \tag{C.3}$$

Appendix D

Appendix to Chapter 7: "The Issue of Power in the Identification of 'g' with Lower-Order Factors"

D.1 R Code for Power Calculations

This appendix presents the R code that can be used to calculate power for various sample sizes. The code takes as inputs the goodness-of-fit statistic of M_A (TA), the chosen Type I error probability (alpha), df_{diff} (df), the sample size (N) used to obtain the non-centrality parameter λ, and the minimum (minN) and maximum (maxN) sample sizes of interest. The output provided by the code consists of the power coefficients corresponding to sample sizes ranging from the minimum and the maximum sample size of interest.

```
#Goodness-of-fit statistic of M_A (i.e., non-centrality parameter lambda)
TA = 7.23
#Type I error probability
alpha = 0.05*2
#Degrees of freedom (i.e., df.diff)
df = 1
#Sample size used to calculate the non-centrality parameter lambda
N = 200
#Critical value
C = qchisq(alpha, df=df, ncp=0, lower.tail=F)
#Minimum sample size of interest
minN = 100
#Maximum sample size of interest
maxN = 2000
power = matrix(0,maxN-minN+1,2)

#Nnew is new sample size of interest
for (Nnew in minN:maxN){
   #lambda.new is the value of the non-centrality parameter
   #corresponding to the new sample size
   lambda.new = (TA/N)*Nnew
   #calculate power
   power[Nnew-minN+1,1] = pchisq(C, df=df, ncp=lambda.new, lower.tail=F)
   power[Nnew-minN+1,2] = Nnew
}
#power plot
plot(power[,2],power[,1],type='l',xlab="Sample size", ylab="Power")
#power for original N
print(power[power[,2]==N,1])
```

Appendix E

Appendix to Chapter 8: “Accounting for Measurement Error and the Attenuation of Correlation: A Bayesian Approach”

E.1 WinBUGS Script

The WinBUGS script that implements the Bayesian correction for the attenuation of the correlation is as follows (see Chapter 8 and Behseta et al., 2009 for details):

```
# Bayesian correction for the attenuation of correlation as a results of uncertainty in measurement
model {
  # Data
  for (i in 1:N){
    # eta[i,1] = theta[i]; eta[i,2] = beta[i];
    # mu[1] = mu_theta; mu[2] = mu_beta;
    # ISigma_cov = inverse of Sigma_cov matrix
    eta[i,1:2] ~ dmnorm(mu[],ISigma_cov[,])

    # observed[i,1] = theta_hat[i]; observed[i,2] = beta_hat[i];
    # Isigma_epsilon[i,1] = Inverse of sigma_epsilon^2_theta[i];
    # Isigma_epsilon[i,2] = Inverse of sigma_epsilon^2_beta[i]
    for (j in 1:2){
      observed[i,j] ~ dnorm(eta[i,j],Isigma_epsilon[i,j])
    }
  }

  # Priors
  mu[1] ~ dnorm(0,.001)
  mu[2] ~ dnorm(0,.001)

  # sigma[1] = sigma_theta; sigma[2] = sigma_beta
  sigma[1] ~ dunif(0,mysigma_1)
  sigma[2] ~ dunif(0,mysigma_2)

  rho ~ dunif(-1,1)

  # Reparameterization
  Sigma_cov[1,1] <- pow(sigma[1],2)
  Sigma_cov[1,2] <- rho*sigma[1]*sigma[2]
  Sigma_cov[2,1] <- rho*sigma[1]*sigma[2]
  Sigma_cov[2,2] <- pow(sigma[2],2)
  ISigma_cov[1:2,1:2] <- inverse(Sigma_cov[1:2,1:2])
}
```

The R script that calls the WinBUGS script using the R2WinBUGS (Sturtz, Ligges, & Gelman, 2005) package is available in the supplemental materials at http://dora.erbe-matzke.com/publications.html. The R script allows users to adjust the range of the uniform prior distribution of sigma[1] and sigma[2] by specifying the value of my_sigma1 and my_sigma2.

Appendix F

Appendix to Chapter 9: "A Default Bayesian Hypothesis Test for Mediation"

F.1 JAGS Code

JAGS Code for Correlation

```
####### Cauchy-prior on alpha #######
model
{
 for (i in 1:n)
  {
    mu[i] <- intercept + alpha*x[i]
    y[i]   ~ dnorm(mu[i],phi)
  }

# uninformative prior on intercept,
# Jeffreys' prior on precision phi
  intercept ~ dnorm(0,.0001)
  phi   ~ dgamma(.0001,.0001)


# inverse-gamma prior on g:
  g       <- 1/invg
  a.gamma <- 1/2
  b.gamma <- n/2
  invg     ~ dgamma(a.gamma,b.gamma)


# g-prior on beta:
  vari <- (g/phi) * invSigma
  prec <- 1/vari
  alpha    ~ dnorm(0, prec)
}

# Explanation---------------------------------
    # Prior on g:
    # We know that g ~ inverse_gamma(1/2, n/2), with 1/2 the shape
    # parameter and n/2 the scale parameter.
    # It follows that 1/g ~ gamma(1/2, 2/n).
    # However, BUGS/JAGS uses the *rate parameterization* 1/theta instead of the
    # scale parametrization theta. Hence we obtain, in de BUGS/JAGS rate notation:
    # 1/g ~ dgamma(1/2, n/2)
    #----------------------------------------
```

JAGS Code for Partial Correlation

```
####### Cauchy-prior on beta and tau' #######

# theta contains beta and tau'

model
{
  for (i in 1:n)
   {
     mu[i] <- intercept + theta[1]*x[i,1] + theta[2]*x[i,2]
     y[i]   ~ dnorm(mu[i],phi)
   }

# uninformative prior on intercept,
# Jeffreys' prior on precision phi
  intercept ~ dnorm(0,.0001)
  phi   ~ dgamma(.0001,.0001)


# inverse-gamma prior on g:
  g       <- 1/invg
  a.gamma <- 1/2
  b.gamma <- n/2
  invg     ~ dgamma(a.gamma,b.gamma)


# calculation of the inverse matrix of V
  inverse.V <- inverse(V)
# calculation of the elements of prior precision matrix
  for(i in 1:2)
  {
    for (j in 1:2)
    {
      prior.T[i,j] <- inverse.V[i,j] * phi/g
    }
  }
# multivariate prior for the theta vector
  theta[1:2] ~ dmnorm( mu.theta, prior.T )
  for(i in 1:2) { mu.theta[i] <- 0 }

}
```

F.2 Testing the Correctness of Our JAGS Implementation

To assess the correctness of our JAGS implementation, we compared the analytical results for the two-sided Bayes factor against the Savage-Dickey density ratio results based on the MCMC samples from JAGS. The distribution that fit the posterior samples best[1] is the non-standardized

[1] We compared the fit of four distributions: a non-standardized t-distribution, a normal distribution, a non-parametric distribution estimated with the spline interpolation function `splinefun` in `R`, and a non-parametric

t-distribution with the following density:

$$p(x|\nu, \mu, \sigma) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu\sigma)}} \left(1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}, \tag{F.1}$$

with ν degrees of freedom, location parameter μ, and scale parameter σ. With the samples of the parameter of interest, we can estimate ν, μ, and σ and thus the exact shape of the distribution and the exact height of the distribution at the point of interest.

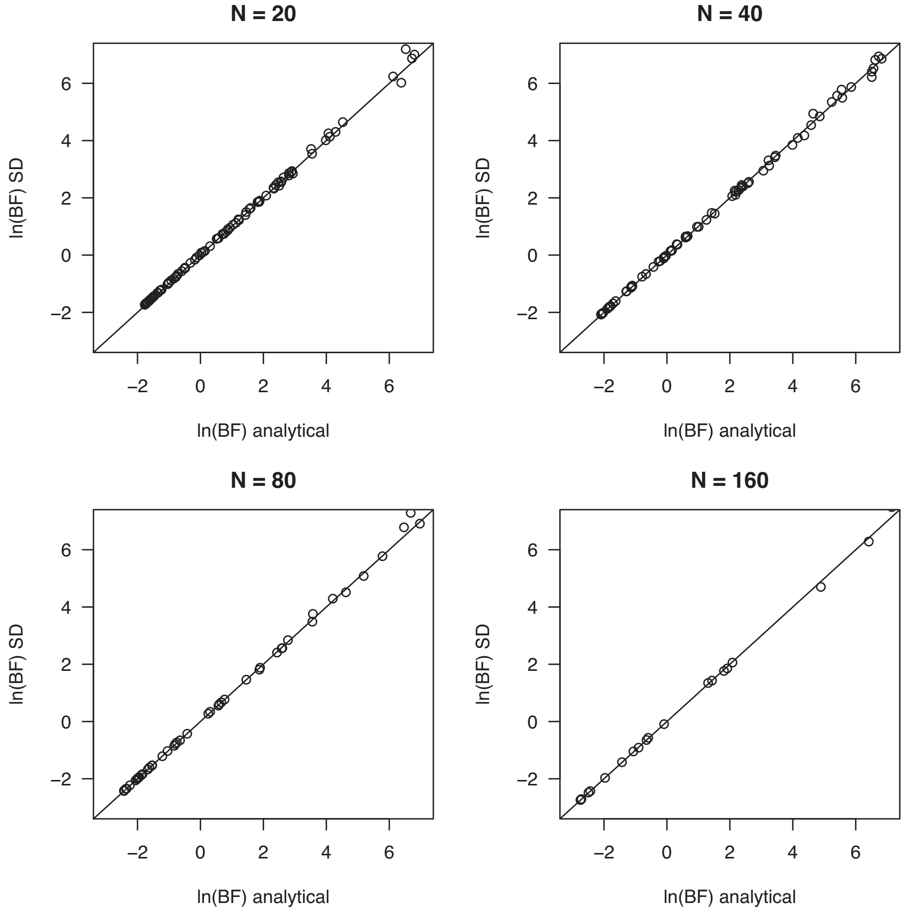


Figure F.1 Natural logarithm of the Bayes factors for correlation obtained with analytical calculations (x axis) or obtained with the SD method based on a non-standardized t-distribution (y axis) for different sample sizes (N). The graphs show fewer points as the samples grow larger, because in these situations there are more extreme Bayes factors that fall outside the axis limits. We restricted the graphs, since it is most important that the lower Bayes factors lie on the diagonal: it is not important whether a Bayes factor is 2000 or 3000, since it is overwhelming evidence in any case.

We checked the fit of this distribution and the performance of the SD method in a small simulation study. We considered the following sample sizes: N = 20, 40, 80, or 160. We simu-

distribution estimated with the R function logspline that also uses splines to estimate the log density. All four distributions fitted reasonably well: the Bayes factors of the analytical test and the SD method are similar with all different posterior distributions. All four distributions are therefore included in the R package BayesMed and can be used when applying the SD method.

lated correlational data by drawing N values for the X from a standard normal distribution, and conditional on X we simulated values for Y according to the following equation:

$$Y_i = \beta_0 + \tau X_i + \epsilon, \tag{F.2}$$

where the subscript i denotes subject i and τ represents the relation between X and Y. For each of the four sample sizes, we generated 100 datasets, each in which τ was drawn from a standard uniform distribution.

Next, we tested the correlation in each dataset with both the analytical Bayesian correlation test and the SD method with the non-standardized t-distribution and compared the results. The results are shown in Figure F.1. The figure shows that the proposed SD method performs well: the Bayes factors of the analytical test and the SD method are similar for all sample sizes and correlations.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*, 567-589. 148

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs. *Psychological Bulletin*, *131*, 30–60. 148

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akadémiai Kiadó. 4, 132, 238, 293

Akaike, H. (1974a). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. 132

Akaike, H. (1974b). On the likelihood of a time series model. *The Statistician*, *27*, 217–235. 133

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author. 216

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060. 126, 141

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 514–544. 12

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43. 138

Ardia, D., Baştürk, N., Hoogerheide, L., & van Dijk, H. K. (2012). A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, *56*, 3398–3414. 137

Armstrong, A. M., & Dienes, Z. (2013). Subliminal understanding of negation: Unconscious control by subliminal processing of word pairs. *Consciousness & Cognition*, *22*, 1022–1040. 178, 183

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144–151. 94

Asrress, K. N., & Carpenter, R. H. S. (2001). Saccadic countermanding: A comparison of central and peripheral stop signals. *Vision Research*, *41*, 2645–2651. 69, 70

Azzelini, A. (1996). *Statistical inference based on the likelihood.* London, UK: Chapman & Hall. 150

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, UK: Cambridge University Press. 96

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. 197

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, *128*, 32–55. 13, 14

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry. *Current Directions in Psychological Science*, *20*, 160–166. 34

Band, G. P. H., van der Molen, M. W., & Logan, G. D. (2003). Horse-race model simulations of the stop-signal procedure. *Acta Psychologica*, *112*, 105–142. 38, 39, 40, 43, 47, 49, 70

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). A new lease of life for Thomson's bonds model for intelligence. *Psychological Review*, *116*, 567–579. 148

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534. 70, 206

Basilevsky, A. (1983). *Applied matrix algebra in the statistical sciences.* Amsterdam, The Netherlands: Elsevier Science Publishing. 147

Batchelder, W. H. (1975). Individual differences and the all-or-none vs incremental learning controversy. *Journal of Mathematical Psychology*, *12*, 53–74. 95

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344. 96

Batchelder, W. H. (2009). Cognitive psychometrics: Using multinomial processing tree models as measurement tools. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model based measurement* (pp. 71–93). Washington, DC: American Psychological Association. 96

Batchelder, W. H., & Crowther, C. S. (1997). Multinomial processing tree models of factorial categorization. *Journal of Mathematical Psychology*, *41*, 45–55. 96

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review*, *87*, 375–397. 3, 97, 126, 132, 237, 292

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*, 129–149. 97, 98, 99, 115

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564. 121

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86. 3, 93, 95, 97, 132, 292

Batchelder, W. H., & Riefer, D. M. (2007). Using multinomial processing tree models to measure cognitive deficits in clinical populations. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling of processes and symptoms* (pp. 19–50). Washington, DC: American Psychological Association. 95, 97, 121, 132

Bateman, I., Kahneman, D., Munro, A., Starmer, C., & Sugden, R. (2005). Testing competing models of loss aversion: An adversarial collaboration. *Journal of Public Economics*, *89*, 1561–1580. 196

Bayarri, M. J., & Berger, J. O. (1998). Quantifying surprise in the data and model verification. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6: Proceedings of the Sixth Valencia International Meeting, June 6-10, 1998* (pp. 53–82). Oxford, UK: Oxford University Press. 83

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., & Smith, K. (2011). Cython: The best of both worlds. *Computing in Science & Engineering*, *13*, 31–39. 80, 236

Behseta, S., Berdyyeva, T., Olson, C. R., & Kass, R. E. (2009). Bayesian correction for attenuation of correlation in multi-trial spike count data. *Journal of Neurophysiology*, *101*, 2186–2193. 4, 159, 160, 161, 163, 165, 166, 173, 174, 239, 259, 293

Belin, T. R., & Rubin, D. B. (1995). The analysis of repeated-measures data on schizophrenic reaction times using mixture models. *Statistics in Medicine*, *14*, 747–768. 34

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. 224

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., & Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behavioural Brain Research*, *125*, 279–284. 233

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Methodological*, 289–300. 232

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*, 1165–1188. 233

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (2nd ed., Vol. 1, pp. 378–386). Hoboken, NJ: Wiley. 136, 181, 189

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352. 180, 209, 218, 225

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*, 542–554. 199

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122. 68, 70, 135

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112–139. 216

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics. 180, 218, 225

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York, NY: Wiley. 135

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*, 215–227. 241

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*, 309–328. 148

Bissett, P. G., & Logan, G. D. (2011). Balancing cognitive demands: Control adjustments in the stop-signal paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 392–404. 39, 53, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 84, 85, 86, 87, 88

Blough, D. S. (1988). Quantitative relations between visual search speed and target-distractor similarity. *Perception & Psychophysics*, *43*, 57–71. 14

Blough, D. S. (1989). Contrast as seen in visual search reaction times. *Journal of the Experimental Analysis of Behavior*, *52*, 199–211. 14

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review*, *114*, 376–397. 69, 236

Bröder, A., Herwig, A., Teipel, S., & Fast, K. (2008). Different storage and retrieval deficits in normal aging and mild cognitive impairment: A multinomial modeling analysis. *Psychology and Aging*, *23*, 353–365. 97

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455. 26, 103

Brown, S. D., & Heathcote, A. (2005). Practice increases the efficiency of evidence accumulation in perceptual choice. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 289–298. 126

Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178. 126, 141, 236

Brunyé, T. T., Mahoney, C. R., Augustyn, J. S., & Taylor, H. A. (2009). Horizontal saccadic eye movements enhance the retrieval of landmark shape and location information. *Brain and Cognition*, *70*, 279–288. 198, 209, 210

Buonaccorsi, J. P. (2010). *Measurement error: Models, methods, and applications.* Boca Raton, FL: Chapman & Hall/CRC. 161

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception & Psychophysics*, *32*, 117–133. 15

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer Verlag. 132, 133

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling.* Thousand Oaks, CA: Sage. 125

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 1-12. 234

Cadsby, C. B., Croson, R., Marks, M., & Maynes, E. (2008). Step return versus net reward in the voluntary provision of a threshold public good: An adversarial collaboration. *Public Choice*, *135*, 277–289. 196

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. 157

Carpenter, R. H. S. (1981). Oculomotor procrastination. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 237–246). Hillsdale, NJ: Erlbaum. 69, 236

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*, 59–62. 15, 30, 69, 236

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609-610. 234

Chambers, C. D., Munafo, M., & et al. (2013). *Trust in science would be improved by study pre-registration.* Retrieved from `http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration` 197, 234

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods*, *10*, 206-226. 159, 161

Chechile, R. A. (1973). *The relative storage and retrieval losses in short–term memory as a function of the similarity and amount of information processing in the interpolated task* (Unpublished doctoral dissertation). University of Pittsburgh. 132

Chechile, R. A., & Meyer, D. L. (1976). A Bayesian procedure for separately estimating storage and retrieval components of forgetting. *Journal of Mathematical Psychology*, *13*, 269–295. 132

Chen, M.-H. (2005). Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, *59*, 16–29. 141

Chen, M.-H., Shao, Q.-M., & Ibrahim, J. G. (2002). *Monte Carlo methods in Bayesian computation.* New York, NY: Springer. 137

Cheng, C.-L., & Van Ness, J. W. (1999). *Statistical regression with measurement error.* London, UK: Arnold. 161

Christman, S. D., Garvey, K. J., Propper, R. E., & Phaneuf, K. A. (2003). Bilateral eye movements enhance the retrieval of episodic memories. *Neuropsychology*, *17*, 221–229. 198, 211

Christman, S. D., Propper, R. E., & Dion, A. (2004). Increased interhemispheric interaction is associated with decreased false memories in a verbal converging semantic associates paradigm. *Brain and Cognition*, *56*, 313–319. 198, 210

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. 94, 96

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates. 149, 216, 218

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. 149

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. 216, 217

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332–361. 126

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300-315. 161

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*, 1503–1514. 148

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by *g*. *Intelligence*, *32*, 277–296. 148, 156

Colonius, H. (1990). A note on the stop-signal paradigm, or how to observe the unobservable. *Psychological Review*, *97*, 309–312. 40, 53, 68

Colonius, H., Özyurt, J., & Arndt, P. A. (2001). Countermanding saccades with auditory stop signals: Testing the race model. *Vision Research*, *41*, 1951–1968. 69, 70

Congdon, P. (2006). *Bayesian statistical modelling* (2nd ed.). Chichester, UK: John Wiley & Sons Ltd. 161

Consonni, G., Forster, J. J., & La Rocca, L. (2013). The whetstone and the alum block: Balanced objective Bayesian comparison of nested models for discrete data. *Statistical Science*, *28*, 398–423. 182

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163–183. 148

Corneil, B. D., & Elsley, J. K. (2005). Countermanding eye-head gaze shifts in humans: Marching orders are delivered to the head first. *Journal of Neurophysiology*, *94*, 883–895. 70

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–172. 217

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300. 216, 217

Curran, T., & Hintzman, D. L. (1995). Violations of the independence assumption in process dissociation. *Journal of Experimental Psycholog: Learning Memory, and Cognition*, 531–547. 94

D'Agostino, R. B., & Stephens, M. A. (1986). *Goodness-of-fit techniques.* New York, NY: Marcel Dekker Inc. 127

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. 96

De Boeck, P., & Partchev, I. (2012). IRTTrees: Tree–based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. 96

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721. 95

Deese, J. (1960). Frequency of usage and number of words in free recall: The role of association. *Psychological Reports*, 337–344. 111

de Groot, A. D. (1961a). *De betekenis van significantie bij verschillende typen onderzoek.* 'S-Gravenhage, The Netherlands: Uitgeverij Mouton. 197, 210

de Groot, A. D. (1961b). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen.* 'S-Gravenhage, The Netherlands: Uitgeverij Mouton. 197

de Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences.* Den Haag, The Netherlands: Mouton. 228, 229, 234

de Groot, A. M. B. (1980). *Mondelinge woordassociatienormen.* Lisse, The Netherlands: Swets & Zeitlinger. 201

de Groot, A. M. B. (1984). Primed lexical decision: Combined effects of the proportion of related prime-target pairs and the stimulus-onset asynchrony of prime and target. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *36*, 253–280. 199, 201

de Groot, A. M. B. (1987). The priming of word associations: A levels-of-processing approach. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *39*, 721–756. 199, 201

de Jong, R., Coles, M. G., Logan, G. D., & Gratton, G. (1990). In search of the point of no return: The control of response processes. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 164–182. 36, 40, 68

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1136–1146. 111

Demetriou, A., Kui, Z. X., Spanoudis, G., Christou, C., Kyriakides, L., & Platsidou, M. (2005). The architecture, dynamics, and development of mental processing: Greek, Chinese, or Universal? *Intelligence*, *33*, 109–141. 148

Dempster, F. N. (1991). Inhibitory processes: A neglected dimension of intelligence. *Intelligence*, *15*, 157–173. 148

Dennis, S. J., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*, 361–376. 217, 225

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42*, 204–223. 68, 70, 141

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226. 141, 165, 188, 239

Didelez, V., Pigeot, I., & Walter, P. (2006). Modifications of the Bonferroni-Holm procedure for a multi-way ANOVA. *Statistical Papers*, *47*, 181–209. 229

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* New York, NY: Palgrave Macmillan. 180, 216, 220, 225

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274-290. 178, 180, 183, 191, 216, 220

Dixon, P. (2003). The *p*-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*, 189–202. 216, 217

Dolan, C. V. (2000). A model-based approach to Spearman's hypothesis. *Multivariate Behavioral Research*, *35*, 21–50. 155

Dolan, C. V., & Hamaker, E. L. (2001). Investigating black-white differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC, and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in psychological research* (Vol. VI, pp. 31–60). Huntington, NY: Nova Science Publishers, Inc. 155

Dolan, C. V., van der Maas, H. L. J., & Molenaar, P. C. M. (2002). A framework for ML estimation of parameters of (mixtures of) common reaction time distributions given optional truncation or censoring. *Behavior Research Methods*, *34*, 304–323. 44, 67

Dominicus, A., Skrondal, A., Gjessing, H. K., Pedersen, N. L., & Palmgren, J. (2006). Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior Genetics*, *36*, 331–340. 152

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*, e29081. 196

Duncan, C. P. (1974). Retrieval of low-frequency words from mixed lists. *Bulletin of the Psychonomic Society*, *4*, 137–138. 111

Duncan, J., Burgess, P., & Emslie, H. (1995). Fluid intelligence after frontal lobe lesions. *Neuropsychologia*, *33*, 261–268. 148

Dunham, M., McIntosh, D., & Gridley, B. E. (2002). An independent confirmatory factor analysis of the Differential Ability Scales. *Journal of Psychoeducational Assessment*, *20*, 152-163. 148, 156

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2009). A diffusion model decomposition of the practice effect. *Psychonomic Bulletin & Review*, *16*, 1026–1036. 126

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242. 2, 77, 174, 180, 199, 200, 209, 216, 225

Elandt-Johnson, R. C., & Johnson, N. L. (1980). *Survival models and data analysis.* New York, NY: Wiley. 41

Elliot, D. L., Goldberg, L., Kuehl, K. S., Moe, E. L., Breger, R. K. R., & Pickering, M. A. (2007). The PHLAME (Promoting healthy lifestyles: Alternative models' effects) firefighter study: Outcomes of two models of behavior change. *Journal of Occupational and Environmental Medicine*, *49*, 204–213. 189

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence*, *20*, 169–189. 148

Emerson, P. L. (1970). Simple reaction time with Markovian evolution of Gaussian discriminal processes. *Psychometrika*, *35*, 99–109. 15

Epstein, J. N., Conners, C. K., Hervey, A. S., Tonev, S. T., Arnold, L. E., Abikoff, H. B., … others (2006). Assessing medication effects in the MTA study using neuropsychological outcomes. *Journal of Child Psychology and Psychiatry*, *47*, 446–456. 14

Erdfelder, E. (2010). A note on statistical analysis. *Experimental Psychology*, *57*, 1–4. 216

Erdfelder, E., Auer, T. S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie*, *217*, 108–124. 93, 132

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140. 94

Farrell, S., & Ludwig, C. J. H. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*, 1209–1217. 24, 48, 78, 94, 163

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York, NY: Springer-Verlag. 96

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, UK: Oliver and Boyd. 217

Fletcher, H. J., Daw, H., & Young, J. (1989). Controlling multiple F test errors with an overall F test. *The Journal of Applied Behavioral Science*, *25*, 101–108. 229, 231

Francis, G. (201s). Publication bias in "Red, rank, and romance in women viewing men" by Elliot et al. (2010). *Journal of Experimental Psychology: General*, *142*, 292–296. 209

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390. 217

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models.* New York, NY: Springer. 144

Fuller, W. A. (1987). *Measurement error models.* New York, NY: John Wiley & Sons, Inc. 161

Gallistel, C. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453. 216, 217, 225

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC. 2, 35, 45, 68, 70, 77, 95, 137, 163, 165, 236, 292

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* Boca Raton, FL: Chapman & Hall. 94, 95

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press. 24, 35, 46, 48, 49, 60, 70, 78, 83, 94, 95, 102, 103, 108, 118, 144, 163, 167, 174, 225

Gelman, A., Meng, X., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807. 60, 83, 118, 167

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472. 45, 78, 103, 139

Gholson, B., & Hohle, R. H. (1968a). Choice reaction time to hues printed in conflicting hue names and nonsense words. *Journal of the Experimental Psychology*, *76*, 413–418. 14

Gholson, B., & Hohle, R. H. (1968b). Verbal reaction times to hues and hue names and forms and form names. *Perception & Psychophysics*, *3*, 191–196. 14

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum. 217

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200. 217

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606. 230

Gignac, G. E. (2007). Working memory and fluid intelligence are both identical to g?! Reanalyses and critical evaluation. *Psychology Science*, *42*, 187–207. 157

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* Boca Raton, FL: Chapman & Hall/CRC. 2, 35, 45, 77, 95, 161, 163, 165, 236, 292

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton, FL: CRC Press. 77, 94, 95, 135, 174

Gilovich, T., Medvec, V. H., & Kahneman, D. (1998). Varieties of regret: A debate and partial resolution. *Psychological Review*, *105*, 602–605. 196

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*, 21–32. 166, 167

Goldacre, B. (2009). *Bad science.* London, UK: Fourth Estate. 197, 234

Golz, D., & Erdfelder, E. (2004). Effekte von L–Dopa auf die Speicherung und den Abruf verbaler Informationen bei Schlaganfallpatienten [Effects of L–Dopa on storage and retrieval of verbal information in stroke patients]. *Zeitschrift für Neuropsychologie*, *15*, 275–286. 97

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*, 389–413. 30

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician*, *59*, 252–257. 216

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* Minneapolis, MN: University of Minnesota Press. 219

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York, NY: Elsevier. 135, 219

Gordon, B., & Carson, K. (1990). The basis for choice reaction time slowing in Alzheimer's disease. *Brain & Cognition*, *13*, 148–166. 13, 14

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732. 188, 238

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20. 197, 209

Gregg, V. H. (1976). Word frequency, recognition and recall. In J. Brown (Ed.), *Recall and recognition* (pp. 183–216). London, UK: Wiley. 111

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, *44*, 133–152. 127, 134

Grünwald, P. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press. 4, 126, 127, 134, 135, 238, 293

Grünwald, P., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT Press. 134

Guo, X., Li, F., Yang, Z., & Dienes, Z. (2013). Bidirectional transfer between metaphorical related domains in implicit learning of form-meaning connections. *PLoS ONE*, *8*, e68100. 178, 183

Gustafson, P. (2004). *Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments.* Boca Raton, FL: Chapman & Hall/CRC. 161

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, *8*, 179–203. 148, 154, 155, 156, 157

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24. 217

Hall, J. F. (1954). Learning as a function of word-frequency. *The American Journal of Psychology*, *67*, 138–140. 111

Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *66*, 1039–1046. 170

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods.* London, UK: Methuen. 4, 137, 138, 238, 293

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing, Inc. 150

Hanes, D. P., & Carpenter, R. H. S. (1999). Countermanding saccades in humans. *Vision Research*, *39*, 2777–2791. 69, 70, 236

Hanes, D. P., Patterson, W. F., & Schall, J. D. (1998). Role of frontal eye fields in countermanding saccades: Visual, movement, and fixation activity. *Journal of Neurophysiology*, *79*, 817–834. 69

Hanes, D. P., & Schall, J. D. (1995). Countermanding saccades in macaque. *Visual Neuroscience*, *12*, 929–937. 69

Hartley, H. O. (1955). Some recent developments in analysis of variance. *Communications on Pure and Applied Mathematics*, *8*, 47–72. 5, 228, 232, 233, 241, 294

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S–PLUS package. *Behavior Research Methods, Instruments & Computers*, *36*, 678–694. 15, 18, 30, 69, 77

Heathcote, A., Brown, S., & Cousineau, D. (2004). QMPE: Estimating Lognormal, Wald, and Weibull RT distributions with a parameter-dependent lower bound. *Behavior Research Methods*, *36*, 277–290. 69, 77

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. 95, 125

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: Different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology*, *66*, 125–136. 126

Heathcote, A., Popiel, S. J., & Mewhort, D. J. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, *109*, 340–347. 10, 12, 13, 34, 35, 40, 42, 43, 68, 74, 75, 292

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society Series A*, *134*, 486–501. 155

Hintze, J. L., & Nelson, R. D. (1998). Violin plots: A box plot-density trace synergism. *American Statistician*, *52*, 181–184. 83, 119, 169

Hintzman, D. L. (1980). Simpson's paradox and the analysis of memory retrieval. *Psychological Review*, *87*, 398–410. 94

Hintzman, D. L. (1993). On variability, Simpson's paradox, and the relation between recognition and recall: Reply to Tulving and Flexser. *Psychological Review*, *100*, 143–148. 94

Hochberg, Y. (1974). Some generalizations of the T-method in simultaneous inference. *Journal of Multivariate Analysis*, *4*, 224–234. 228

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800–802. 228, 232

Hockley, W. E. (1982). Retrieval processes in continuous recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 497–512. 13, 35, 40, 42, 68, 75

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 598–615. 13, 35, 40, 42, 68, 75

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417. 133

Hofstee, W. K. B. (1984). Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychologica*, *56*, 93–109. 196, 197, 208

Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*, 382-386. 10, 11, 12, 14

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses.* New York, NY: Springer. 181, 186

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70. 228, 232

Howard, G., Maxwell, S., & Fleming, K. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*, 315–332. 216, 221

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. 3, 94, 95, 237, 292

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods*, *31*, 220–234. 94

Hunter, J. E. (2001). The desperate need for replications. *Journal of Consumer Research*, *28*, 149–158. 196

IBM Corp. (2012). IBM SPSS statistics for Windows [Computer software manual]. Armonk, NY. (Version 21.0) 228

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701. 197, 224

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181. 181

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge, UK: Cambridge University Press. 218, 225

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, *80*, 64–72. 127, 136

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press. 127, 135, 136, 137, 140, 143, 165, 167, 172, 181, 182, 199, 200, 206, 209, 218, 219, 226

Jensen, A. R. (1998). *The g factor: The science of mental ability.* New York, NY: Praeger. 148, 169

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences*, *3*, 423–438. 155

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532. 197

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*, 19313–19317. 209

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*, 393–416. 148, 154, 155, 156, 157

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7: A guide to the program and applications.* Mooresville, IN: Scientific Software, Inc. 155

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8.50: Users reference guide.* Chicago, IL: Scientific Software International. 153

Kahneman, D. (2003). Experiences of collaborative research. *American Psychologist*, *58*, 723. 4, 196, 197, 208, 240, 294

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, *86*, 199–225. 148

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*, 66–71. 148

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, *68*, 373–389. 95, 96

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. 68, 70, 114, 135, 137, 165, 199, 216, 218, 219, 225

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, *90*, 928–934. 183

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service. 157

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217. 230

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*, 627–633. 201

Kieffaber, P. D., Kappenman, E. S., Bodkins, M., Shekhar, A., O'Donnell, B. F., & Hetrick, W. P. (2006). Switch and maintenance of task set in schizophrenia. *Schizophrenia Research*, *84*, 345–358. 10, 13, 14

Killeen, P. R. (2005). An alternative to null–hypothesis significance tests. *Psychological Science*, *16*, 345–353. 217

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, *13*, 549–562. 217

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*, 7–31. 94, 95

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98. 94, 96, 99, 102, 103, 104, 114, 118, 119, 121, 122, 132, 144, 174

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493. 68, 70, 188

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614. 196, 197, 208

Kornylo, K., Dill, N., Saenz, M., & Krauzlis, R. J. (2003). Canceling of pursuit and saccadic eye movements in humans and monkeys. *Journal of Neurophysiology*, *89*, 2984–2999. 70

Kraemer, H. C., & Thiemann, S. (1989). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage. 149

Kramer, A. F., Humphrey, D. G., Larish, J. F., Logan, G. D., & Strayer, D. (1994). Aging and inhibition: Beyond a unitary view of inhibitory processing in attention. *Psychology and Aging*, *9*, 491–512. 34, 66, 74

Krijnen, W. P. (2004). Positive loadings and factor correlations from positive covariance matrices. *Psychometrika*, *69*, 655–660. 149

Kromrey, J. D., & Dickinson, W. B. (1995). The use of an overall F test to control type I error rates in factorial analyses of variance: Limitations and better strategies. *The Journal of Applied Behavioral Science*, *31*, 51–64. 229, 231

Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658–676. 216, 217, 220, 225

Kruschke, J. K. (2010b). *Doing Bayesian data analysis: A tutorial introduction with R and BUGS.* Burlington, MA: Academic Press. 45, 95, 165, 180, 220, 225

Kruschke, J. K. (2010c). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300. 216, 217, 225

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299-312. 220, 224, 225

Kyllonen, P. C. (1993). Aptitude testing inspired by information processing: A test of the four-sources model. *The Journal of General Psychology*, *120*, 375–405. 148

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 389–433. 148

Lappin, J. S., & Eriksen, C. W. (1966). Use of a delayed signal to stop a visual reaction-time response. *Journal of Experimental Psychology*, *72*, 805–811. 34, 36, 73

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez–Latham dispute regarding participation in goal setting. *Journal of Applied Psychology*, *73*, 753–772. 196, 208

Lee, C. W., & Cuijpers, P. (2013). A meta-analysis of the contribution of eye movements in processing emotional memories. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*, 231.239. 198, 213

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15. 25, 46, 100, 163, 216

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7. 46, 48, 78, 94, 95, 163, 225

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision-making. *Judgment and Decision Making*, *6*, 832–842. 94

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668. 216, 217, 225

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian modeling for cognitive science: A practical course.* Cambridge, UK: Cambridge University Press. 49, 70, 77, 94, 95, 132, 136, 165, 174, 180, 181

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*, 605–621. 95

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach.* New York, NY: John Wiley & Sons. 174, 240

Lehmann, E. L., & Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, *33*, 1138-1154. 229

Leth-Steensen, C., King Elbaz, Z., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica*, *104*, 167–190. 10, 14, 34

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice.* Thousand Oaks, CA: Sage. 125

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655. 136, 181

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423. 68, 137, 178, 183, 200, 220, 241

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192. 189

Lindley, D. V. (1972). *Bayesian statistics: A review.* Philadelphia, PA: Society for Industrial and Applied Mathematics. 218, 225

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: Methodological*, *34*, 1–41. 46

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375. 70, 206

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31. 129

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171. 216, 217

Logan, G. D. (1981). Attention, automaticity, and the ability to stop a speeded choice respons. In J. B. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 205–222). Hillsdale, NJ: Erlbaurn. 34, 36, 37, 66, 74, 75

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527. 125

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883–914. 125

Logan, G. D. (1994). On the ability to inhibit thought and action: A users' guide to the stop signal paradigm. In D. Dagenbach & T. H. Carr (Eds.), *Inhibitory processes in attention, memory and language* (pp. 189–240). San Diego, CA: Academic Press. 37, 39, 40, 74

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*, 376–400. 125

Logan, G. D., & Burkell, J. (1986). Dependence and independence in responding to double stimulation: A comparison of stop, change, and dual-task paradigms. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 549–563. 36, 37, 38

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, *91*, 295–327. 1, 3, 33, 34, 36, 37, 38, 39, 60, 66, 73, 74, 75, 236, 292

Logan, G. D., Schachar, R. J., & Tannock, R. (1997). Impulsivity and inhibitory control. *Psychological Science*, *8*, 60–64. 39

Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review*, *121*, 66-95. 74, 236

Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, *14*, 41–68. 240

Lord, F. M., & Novick, M. R. (1986). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley. 96

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York, NY: Oxford University Press. 10, 12, 15, 40, 77

Lunn, D. (2003). WinBUGS development interface (WBDev). *ISBA Bulletin*, *10*, 10–11. 252

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis.* Boca Raton, FL: Chapman & Hall/CRC. 74, 94, 138, 161, 165, 236

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067. 94, 122, 165

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS–A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337. 26, 45, 94, 220

Lyle, K. B., Hanaver-Torrez, S. D., Hackländer, R. P., & Edlin, J. M. (2012). Consistency of handedness, regardless of direction, predicts baseline memory accuracy and potential for memory enhancement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 187-193. 198, 211

Lyle, K. B., Logan, J. M., & Roediger, H. L. (2008). Eye movements enhance memory for individuals who are strongly right-handed and harm it for individuals who are not. *Psychonomic Bulletin & Review*, *15*, 515–520. 198, 200, 209, 211

Lyle, K. B., & Osborn, A. E. (2011). Inconsistent handedness and saccade execution benefit face memory without affecting interhemishperic interction. *Memory*, *19*, 613–624. 198, 210

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge, UK: Cambridge University Press. 127, 136

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*, 593. 177, 178

MacKinnon, D. P., Lockwood, C. M., & Hoffman, J. (1998). A new method to test for mediation. In *Annual meeting of the society for prevention research.* Park City, UT. 179, 180

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*, 83–104. 180

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128. 189

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*, 41–62. 180

Madden, D. J., Gottlob, L. R., Denny, L. L., Turkington, T. G., Provenzale, J. M., Hawk, T. C., & Coleman, R. E. (1999). Aging and recognition memory: Changes in regional cerebral blood flow associated with components of reaction time distributions. *Journal of Cognitive Neuroscience*, *11*, 511–520. 13, 14

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, *39*, 906–913. 241

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null–hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. 68, 112, 241

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (in press). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika.* 132, 144, 174

Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, *142*, 1047-1073. 73, 74, 75, 76, 77, 78, 79, 80, 83, 84, 88, 89, 90, 255

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-gaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817. 34, 43, 74, 75, 77, 78, 163, 170

McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 309–360). New York, NY: Wiley. 11

McGill, W. J., & Gibbon, J. (1965). The general-gamma distribution and reaction times. *Journal of Mathematical Psychology*, *2*, 1–18. 11

McHugh, R. (1958). Significance level in factorial design. *The Journal of Experimental Education*, *26*, 257–260. 228, 232

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275. 196, 197, 208

Merritt, P. S., DeLosh, E. L., & McDaniel, M. A. (2006). Effects of word frequency on individual-item and serial order retention: Tests of the order-encoding view. *Memory & Cognition*, *34*, 1615–1627. 111

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419. 191

Morey, R. D., & Wagenmakers, E.-J. (2014). Simple relation between Bayesian order-restriced and point-null hypothesis tests. *Statistics and Probability Letters*, *92*, 121-124. 166, 188

Moshagen, M. (2010). MultiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54. 94

Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement*, *56*, 63–75. 162

Mussweiler, T. (2006). Doing is for thinking! *Psychological Science*, *17*, 17–21. 215, 216, 217, 218

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*, 313-335. 240

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204. 128

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100. 19, 44, 67

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*. 126, 143, 216

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, *50*, 167–179. 134, 135

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95. 136, 181

Naglieri, J. A., & Jensen, A. R. (1985). Comparison of black-white differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence*, *11*, 21–43. 155

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, *15*, 1044–1045. 233

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122. 95

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & Cognition*, *4*, 648–654. 199

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*, 226–254. 199

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. 197

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301. 216, 217

Nier, J. A., & Campbell, S. D. (2012). Two outsiders view on feminism and evolutionary psychology: An opportune time for adversarial collaboration. *Sex Roles*, 1–4. 196

Nieuwenhuis, S., Elzinga, B. M., Ras, P. H., Berends, F., Duijs, P., Samara, Z., & Slagter, H. A. (2013). Bilateral saccadic eye movements and tactile stimulation, but not auditory stimulation, enhance memory retrieval. *Brain and Cognition*, *81*, 52–56. 198, 211

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*, 84–93. 48, 94, 166

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. 197

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (in preparation). BayesMed: Default Bayesian hypothesis tests for correlation, partial correlation, and mediation [Computer software manual]. (R package version 0.1.0) 190

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (submitted). A default Bayesian hypothesis tests for mediation. *Manuscript submitted for publication.* 241

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society Series B*, *57*, 99–138. 135

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B. Bayesian inference* (2nd ed.). London, UK: Arnold. 68, 70, 133, 141, 180

Okada, R. (1971). Decision latencies in short-term recognition memory. *Journal of Experimental Psychology*, *90*, 27–32. 18

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh Inventory. *Neuropsychologia*, *9*, 97–113. 200

Olejnik, S., Li, J., Supattathum, S., & Huberty, C. J. (1997). Multiple testing and statistical power with modified Bonferroni procedures. *Journal of Educational and Behavioral Statistics*, *22*, 389–406. 229, 230

Oosterlaan, J., Logan, G. D., & Sergeant, J. A. (1998). Response inhibition in AD/HD, CD, comorbid AD/HD+CD, anxious, and control children: A meta-analysis of studies with the stop task. *Journal of Child Psychology and Psychiatry*, *39*, 411–425. 34, 66, 74

Osman, A., Kornblum, S., & Meyer, D. E. (1986). The point of no return in choice reaction time: Controlled and ballistic stages of response preparation. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 243-258. 39

Overstall, A. M., & Forster, J. J. (2010). Default Bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, *54*, 3269–3288. 182

Parker, A., Buckley, S., & Dagnall, N. (2009). Reduced misinformation effects following saccadic bilateral eye movements. *Brain and Cognition*, *69*, 89–97. 198

Parker, A., & Dagnall, N. (2007). Effects of bilateral eye movements on gist based false recognition in the DRM paradigm. *Brain and Cognition*, *63*, 221–225. 198

Parker, A., & Dagnall, N. (2010). Effects of handedness and saccadic bilateral eye movements on components of autobiographical recollection. *Brain and Cognition*, *73*, 93–101. 198

Parker, A., & Dagnall, N. (2012). Effects of saccadic bilateral eye movements on memory in children and adults: An exploratory study. *Brain and Cognition*, *78*, 238–247. 198

Parker, A., Relph, S., & Dagnall, N. (2008). Effects of bilateral eye movements on the retrieval of item, associative, and contextual information. *Neuropsychology*, *22*, 136. 198

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536. 197

Pashler, H., & Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. 196

Patil, A., Huard, D., & Fonnesbeck, C. (2010). PyMC: Bayesian stochastic modelling in Python. *Journal of Statistical Software*, *35*, 1–81. 80, 236

Penner-Wilger, M., Leth-Steensen, C., & LeFevre, J. A. (2002). Decomposing the problem-size effect: A comparison of response time distributions across cultures. *Memory & Cognition*, *30*, 1160–1167. 14

Pericchi, L. R., Liu, G., & Torres, D. (2008). Objective Bayes factors for informative hypotheses: "Completing" the informative hypothesis and "splitting" the Bayes factor. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 131–154). New York, NY: Springer-Verlag. 188

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425. 4, 126, 129, 134, 237, 292

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491. 134, 219

Plourde, C. E., & Besner, D. (1997). On the locus of the word frequency effect in visual word recognition. *Canadian Journal of Experimental Psychology*, *51*, 181–194. 12

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing.* Vienna, Austria. 122, 138, 165

Plummer, M. (2009). JAGS Version 1.0.3 manual [Computer software manual]. Retrieved from `http://www-ice.iarc.fr/~{}martyn/software/jags/jags_user_manual.pdf` 188, 239

Plummer, M. (2013). JAGS Version 3.1.0 [Computer software manual]. Retrieved from `http://mcmc-jags.sourceforge.net/` 165, 166

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage*, *40*, 409–414. 231

Possamaï, C. A. (1991). A responding hand effect in a simple-RT precuing experiment: Evidence for a late locus of facilitation. *Acta Psychologica*, *77*, 47–63. 13, 14

Postman, L. (1970). Effects of word frequency on acquisition and retention under conditions of free-recall learning. *The Quarterly Journal of Experimental Psychology*, *22*, 185–195. 111

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single-and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46. 121

Press, S., Chib, S., Clyde, M., Woodworth, G., & Zaslavsky, A. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications.* Hoboken, NJ: Wiley-Interscience. 220

Propper, R. E., & Christman, S. D. (2008). Interhemispheric interaction and saccadic horizontal eye movements. Implications for episodic memory, EMDR, and PTSD. *Journal of EMDR Practice and Research*, *2*, 269–281. 198, 211

Purdy, B. P., & Batchelder, W. H. (2009). A context-free language for binary multinomial processing tree models. *Journal of Mathematical Psychology*, *53*, 547–561. 94

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0) 80, 166, 190, 233, 236, 238

R Development Core Team. (2006). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` 153

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (p. 111-196). Cambridge, UK: Blackwells. 112, 133

Raftery, A. E. (1999). Bayes factors and BIC. *Sociological Methods & Research*, *27*, 411–417. 112

Rao, C. R., & Toutenburg, H. (1999). *Linear models: Least squares and alternatives* (2nd ed.). New York, NY: Springer-Verslag. 228

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. 1, 3, 10, 13, 16, 18, 29, 35, 40, 42, 43, 68, 75, 126, 141, 160, 166, 170, 174, 235, 292

Ratcliff, R. (1993). A method for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532. 13, 35, 40, 42, 68, 75

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*, 278–291. 16

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159–182. 16

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922. 10, 16, 17, 170

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*, 190–214. 13, 18, 35, 40, 42, 68, 75, 76

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127–140. 16

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. 170

Ratcliff, R., & Strayer, D. (2014). Modeling simple driving tasks with a one-boundary diffusion model. *Psychonomic Bulletin & Review*, *21*, 577–589. 161

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*, 278–289. 16

Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*, 323–341. 16

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65*, 523–535. 16

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory & Language*, *50*, 408–424. 16

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*, 127–157. 170

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 4: The Advanced Progressive Matrices*. San Antonio, TX: Harcourt Assessment. 170

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. 218

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General*, *126*, 288–311. 126

Ridderinkhof, K. R., Band, G. P. H., & Logan, G. D. (1999). A study of adaptive behavior: Effects of age and irrelevant information on the ability to inhibit one's actions. *Acta Psychologica*, *101*, 315–337. 34, 66, 74

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. 94, 132

Riefer, D. M., & Batchelder, W. H. (1991). Statistical inference for multinomial processing tree models. In J.-P. Doignon & J.-C. G. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 313–335). New York, NY: Springer–Verlag. 95, 97

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–200. 97, 114

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 445–471. 134

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society Series B*, *49*, 223–239. 134

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*, 40–47. 134

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, *47*, 1712–1717. 134, 135

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science*, *7*, 411–426. 197

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition*, *24*, 188–201. 14

Rohrer, D. (2002). The breadth of memory search. *Memory*, *10*, 291–301. 14

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition*, *22*, 511–524. 13, 14

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*, 663–665. 232

Rosenthal, R. (1976). *Experimenter effects in behavioral research.* New York, NY: Irvington. 210

Rosenthal, R. (1979). An introduction to the file drawer problem. *Psychological Bulletin*, *86*, 683–641. 197, 209

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*, 775–777. 218

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *1*, 377–386. 211

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, *74*, 166–169. 218

Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember-know paradigm. *Psychonomic Bulletin & Review*, *15*, 825–832. 13, 14

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*, 377-381. 77

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. 48, 94, 95, 96, 101, 102, 106, 121, 225

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process-dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389. 94, 96, 101, 106, 121, 132, 174

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223. 24, 46, 48, 77, 78, 163

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642. 95, 96, 101, 106, 121, 122, 123, 174

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903. 178, 241

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. 178, 180, 199

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. 2, 4, 68, 70, 178, 181, 199, 200, 216, 217, 219, 220, 225, 239, 241, 293

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606. 24, 46, 48, 95

Ryan, T. A. (1959). Multiple comparison in psychological research. *Psychological Bulletin*, *56*, 26-47. 230

Samara, Z., Elzinga, B. M., Slagter, H. A., & Nieuwenhuis, S. (2011). Do horizontal saccadic eye movements increase interhemispheric coherence? Investigation of a hypothesized neural mechanism underlying EMDR. *Frontiers in Psychiatry*, *2*, 1–7. 198, 209, 211

Saris, W. E., & Satorra, A. (1993). Power evaluation in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing strucural equation models.* (pp. 181–204). Newburry Park, CA: Sage. 149, 150

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, *50*, 83–90. 149, 150

Schachar, R., & Logan, G. D. (1990). Impulsivity and inhibitory control in normal development and childhood psychopathology. *Developmental Psychology*, *26*, 710–720. 34, 66, 74

Schachar, R., Mota, V. L., Logan, G. D., Tannock, R., & Klim, P. (2000). Confirmation of an inhibitory control deficit in attention-deficit/hyperactivity disorder. *Journal of Abnormal Child Psychology*, *28*, 227–235. 34, 66, 74

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, *40*, 87–110. 228

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*, 39–64. 144

Schlitz, M., Wiseman, R., Watt, C., & Radin, D. (2006). Of two minds: Sceptic-proponent collaboration within parapsychology. *British Journal of Psychology*, *97*, 313–322. 196

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129. 217

Schmiedek, F., Oberauer, K., Wilhelm, O., Suss, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414–429. 14, 17, 18, 170

Schmittmann, V. D., Dolan, C. V., Raijmakers, M. E., & Batchelder, W. H. (2010). Parameter identification in multinomial processing tree models. *Behavior Research Methods*, *42*, 836–846. 121

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*, 143–153. 31

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. 4, 133, 183, 238, 293

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, *33*, 457–469. 15, 18, 69, 77

Schwarz, W. (2002). On the convolution of inverse Gaussian and exponential random variables. *Communications in Statistics: Theory & Methods*, *31*, 2113–2121. 15

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*, 2144–2162. 220, 241

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587–2619. 241

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71. 180, 209

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, *81*, 826–831. 228, 232

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*, 561–584. 231

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., … Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*, *8*, e56515. 196

Shapiro, F. (1989). Eye movement desensitization: A new treatment for post-traumatic stress disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, *20*, 211–217. 198

Shenoy, P., Rao, R., & Yu, A. J. (2010). A rational decision-making framework for inhibitory control. In J. Lafferty, C. k. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 23, p. 2146-2154). Cambridge, MA: MIT Press. 70

Shenoy, P., & Yu, A. J. (2011). Rational decision-making in inhibitory control. *Frontiers in Human Neuroscience*, *5*, 48. 70

Sheu, C., & O'Curry, S. L. (1998). Simulation-based Bayesian inference using BUGS. *Behavior Research Methods*, *30*, 232–237. 95

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284. 24, 25, 48, 94

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166. 141

Silver, N. (2012). *The signal and the noise: The art and science of prediction.* London, UK: Allen Lane. 127

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 3). London, UK: Chapman & Hall. 40

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. 196, 197

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. 211

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575. 132, 133, 135

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. 94, 95

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. 94, 95, 132

Smith, P. L. (1995). Psychophysically principled models of visual simple reaction time. *Psychological Review*, *102*, 567–593. 15

Smith, R. A., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and Type II error rates in multiway ANOVA. *Human Communication Research*, *28*, 515–530. 230

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. 179, 189

Song, X.-Y., & Lee, S.-Y. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *Journal of Mathematical Psychology*, *56*, 135–148. 174, 240

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101. 159, 160, 161, 162, 163, 164

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussions). *Journal of the Royal Statistical Society Series B*, *64*, 583–616. 83

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., & Lunn, D. (2003). BUGS: Bayesian inference using Gibbs sampling [Computer software manual]. Retrieved from `http://www.mrc-bsu.cam.ac.uk/bugs/` 108

Spieler, D. H. (2001). Modelling age-related changes in information processing. *Europian Journal of Cognitive Psychology*, *13*, 217–234. 18

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 461–479. 14

Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 506–526. 18

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods*, *39*, 267–273. 95

Stan Development Team. (2012). Stan modeling language [Computer software manual]. Retrieved from `http://mc-stan.org/` 122, 165

Stauffer, J. M., Ree, M. J., & Carretta, T. R. (1996). Cognitive-components tests are not much more than *g*: An extension of Kyllonen's analyses. *Journal of General Psychology*, *123*, 193–206. 148

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652–654. 18

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, *11*, 439–455. 152

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, *25*, 1371–1470. 141

Sturtz, S., Ligges, U., & Gelman, A. E. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*, 1–16. 260

Sumby, W. H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, *1*, 443–450. 111

Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability— and a little bit more. *Intelligence*, *30*, 261–288. 148

Tetlock, P. E., & Mitchell, G. (2009). Implicit bias and accountability systems: What must organizations do to prevent discrimination? *Research in Organizational Behavior*, *29*, 3–38. 196

Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*, 415–429. 16

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32. 218

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, 1701–1728. 80

Toutenburg, H. (2002). *Statistical analysis of designed experiments.* New York, NY: Springer-Verslag. 229

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes.* Cambridge, UK: Cambridge University Press. 10

Tukey, J. W. (1973). *The problem of multiple comparisons.* Princeton University. 228

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. 160, 166, 174

Ulrich, R., & Miller, J. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*, 34–80. 41

Undheim, J. O., & Gustafsson, J. E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, *22*, 149–171. 148, 154, 156

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2013). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology. Manuscript accepted pending minor revision.* Oxford, UK: Oxford University Press. 181

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011–1026. 19

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*, 61–72. 19

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychologival Review*, *113*, 842–861. 148

van der Sluis, S., Dolan, C. V., & Stoel, R. D. (2005). A note on testing perfect correlations in SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 551–577. 152

van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Modeling*, *17*, 443–463. 240

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498. 70, 206

van Ravenzwaaij, D., Brown, S., & Wagenmakers, E.-J. (2011). An integrated perspective on the relation between response speed and intelligence. *Cognition*, *119*, 381–393. 170

Van Zandt, T. (2002). Analysis of response time distributions. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology* (pp. 461–516). San Diego, CA: Academic Press. 69

Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 87–94. 180

Verbruggen, F., Chambers, C. D., & Logan, G. D. (2009). Fictious inhibitory differences: How skewness and slowing distort the estimation of stopping latencies. *Psychological Science*, *24*, 352–362. 74

Verbruggen, F., & Logan, G. D. (2009). Models of response inhibition in the stop-signal and stop-change paradigms. *Neuroscience & Biobehavioral Reviews*, *33*, 647–661. 39, 74

Verbruggen, F., Logan, G. D., & Stevens, M. A. (2008). STOP–IT: Windows executable software for the stop-signal paradigm. *Behavior Research Methods*, *40*, 479–483. 39, 69

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*, 614–618. 141

Verhagen, J., & Wagenmakers, E.-J. (2014). A Bayesian test to quantify the success or failure of a replication attempt. *Journal of Experimental Psychology: General*, *143*. 191

Vickers, D., Nettelbeck, T., & Willson, R. J. (1972). Perceptual indices of performance: The measurement of "inspection time" and "noise" in the visual system. *Perception*, *1*, 263–295. 169

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220. 3, 10, 16, 17, 170, 235, 292

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228. 133

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. Retrieved from `http://www.cidlab.com/supp.php?o=wv13` 170

Wagenaar, W. A., & Boer, J. P. A. (1987). Misleading postevent information: Testing parameterized models of integration in memory. *Acta Psychologica*, *66*, 291–306. 126, 129, 130, 131, 132, 133, 134, 135, 138, 139, 140, 141, 142, 143, 144, 238

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804. 112, 180, 199, 216, 225, 233

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671. 16, 17, 160, 170, 174

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192–196. 133

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*, 641–642. 189, 216, 217

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189. 48, 141, 165, 181, 188, 199, 216

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory & Language*, *58*, 140–159. 10, 16, 22, 26, 27, 29, 235

Wagenmakers, E.-J., van der Maas, H. L. J., Dolan, C. V., & Grasman, R. P. P. P. (2008). EZ does it! Extensions of the EZ-diffusion model. *Psychonomic Bulletin & Review*, *15*, 1229–1235. 10, 43

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22. 31

Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, *50*. 126, 143

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *100*, 426–432. 137, 191, 196, 197, 199, 200, 208, 224

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. 4, 197, 199, 200, 208, 234, 240, 294

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212–213. 217

Wald, A. (1947). *Sequential analysis.* New York, NY: Wiley. 13

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference.* New York, NY: Springer. 217

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children— Revised.* New York, NY: The Psychological Corporation. 155

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics & Data Analysis*, *54*, 2094–2102. 141, 165, 188

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for ANOVA designs. *The American Statistician*, *66*, 104–111. 2, 68, 178, 241

Wetzels, R., Lee, M. D., & Wagenmakers, E.-J. (2010). Bayesian inference using WBDev: A tutorial for social scientists. *Behavior Research Methods*, *42*, 884–897. 220, 236, 252

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298. 180, 209

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*, 752–760. 68, 70, 178, 199, 200, 216, 217, 220, 241

Wetzels, R., & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*, 1057-1064. 68, 178, 182, 183, 184, 239, 241

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85. 31

Wickelmaier, F. (2011). Mpt: Multinomial processing tree (MPT) models [Computer software manual]. Retrieved from `http://cran.r-project.org/web/packages/mpt/index.html` 94

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, doi: 10.3389/fninf.2013.00014. 80, 236

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. 216

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., & Tannock, R. (1999). Development of inhibitory control across the life span. *Developmental Psychology*, *35*, 205–213. 34, 47, 49, 66, 70, 74

Winnie, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, *19*, 125-134. 161

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detecting of staring. *Journal of Parapsychology*, *61*, 197–207. 196

Wiseman, R., & Schlitz, M. (1998). Replication of experimenter effects and the remote detecting of staring. *Proceedings of the 12nd Annual Convention of the Parapsychological Association*, 471–479. 196

Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure-and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 523–538. 14

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1024–1039. 14

Wolfe, J. M. (2013). Registered reports and replications in Attention, Perception & Psychophysics. *Attention, Perception & Psychophysics*, *75*, 781-783. 234

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics*, *48*, 1005–1013. 228, 231, 232

Wu, H., Myung, J. I., & Batchelder, W. H. (2010). Minimum description length model selection of multinomial processing tree models. *Psychonomic Bulletin & Review*, *17*, 275–286. 135

Yap, M. J., & Balota, D. A. (2007). Additive and interactive effects on response time distributions in visual word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 274–296. 12

Yap, M. J., Balota, D. A., Cortese, M. J., & Watson, J. M. (2006). Single-versus dual-process models of lexical decision performance: Insights from response time distributional analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1324–1344. 12, 18

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*, 301–322. 177, 178, 180, 181, 189, 190, 191, 239

Zeelenberg, R., Wagenmakers, E.-J., & Rotteveel, M. (2006). The impact of emotion on perception: Bias or enhanced processing? *Psychological Science*, *17*, 287–291. 201

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. De Groot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia, Spain: University Press. 183

Samenvatting

Hoe kunnen we data verkregen uit psychologische experimenten het beste beschrijven? In dit proefschrift stelde ik dat dit bij uitstek gedaan kan worden met behulp van formele wiskundige modellen. Het doel van modelleren is om datapatronen te ontdekken en deze te beschrijven aan de hand van parameters die verschillende statistische of psychologische processen vertegenwoordigen. Er bestaan veel verschillende typen wiskundige modellen waarvan velen in dit proefschrift aan de orde zijn komen: Ik heb me gericht op beschrijvende en procesmodellen voor eenvoudige tweekeuze-responstijdtaken, modellen voor responsinhibitie zoals gemeten in het stopsignaalparadigma, multinomiale *processing-tree*-modellen voor de analyse van categorische data, en bekende statistische modellen zoals de t-toets, de variantieanalyse, correlaties en partiële correlaties, latente-variabelemodellen, en mediatie-analyse.

Nadat we een wiskundig model voor onze data hebben gekozen, moeten we de modelparameters schatten en nagaan of het gekozen model inderdaad een adequate beschrijving van de data biedt. Hoe kunnen we het beste psychologische datasets die beschreven zijn met behulp van mathematische modellen analyseren? In dit proefschrift stelde ik dat dit het beste gedaan kan worden door middel van Bayesiaanse inferentie. Ik heb beweerd dat Bayesiaanse statistiek belangrijke theoretische en praktische voordelen biedt ten opzichte van frequentistische statistiek, voordelen die Bayesiaanse procedures bij uitstek geschikt maken om de problemen uit de dagelijkse werkpraktijk van psychologisch onderzoek aan te pakken.

In rest van deze samenvatting zal ik een overzicht en een korte beschrijving geven van de vraagstukken waarmee ik me tijdens mijn promotieproject heb beziggehouden. In dit proefschrift zijn zeer verschillende onderwerpen verkend; de gemene deler is de toewijding aan mathematisch modelleren en zorgvuldige statistische inferentie.

Deel I. De Analyse van Responstijdverdelingen

Het eerste deel van dit proefschrift richtte zich op het modelleren van responstijden—zowel geobserveerde als niet geobserveerde—met behulp van de ex-Gaussian en shifted Wald-responstijdverdelingen.

In het tweede hoofdstuk onderzocht ik de validiteit van de cognitieve interpretatie van de parameters van de ex-Gaussian en shifted Wald verdelingen. De ex-Gaussian en de shifted Wald zijn veelgebruikte statistische modellen voor het beschrijven van responstijdverdelingen in snelle tweekeuzetaken, waarvan de parameters vaak geïnterpreteerd worden in termen van cognitieve processen. We hebben de validiteit van deze cognitieve interpretatie onderzocht door de parameters van de ex-Gaussian en shifted Wald-verdelingen te relateren aan de parameters van het Ratcliffdiffusiemodel (Ratcliff, 1978), een succesvol procesmodel waarvan de parameters een gegronde cognitieve interpretatie kennen (e.g.,Voss et al., 2004). De resultaten tonen aan dat de ex-Gaussian en shifted Wald-parameters niet uniek overeenkomen met de parameters van het diffussiemodel. De cognitieve interpretatie van de parameters van deze verdelingen wordt daarom afgeraden.

In het derde hoofdstuk introduceerde ik een Bayesiaanse parametrische methode voor het schatten van de tijd die mensen nodig hebben om hun respons af te breken (stopsignaalresponstijd; SSRT) in het stopsignaalparadigma. Het stopsignaalparadigma is een populair experimentele procedure die wordt gebruikt om het onderdrukken van responsen op tweekeuzetaken te onderzoeken. Gebaseerd op het racemodel (Logan & Cowan, 1984) zijn verscheidende methoden ontwikkeld om SSRT's te schatten die anders niet geobserveerd zouden kunnen worden. Geen van deze methoden is echter in staat om een accurate schatting te maken van de gehele verdeling van SSRT's, terwijl responstijdverdelingen juist waardevolle informatie kunnen bevatten voor de onderzoeker (Heathcote et al., 1991). We introduceerden een Bayesiaanse methode om deze beperking te verhelpen. Deze nieuwe aanpak doet de aanname dat SSRT's een ex-Gaussian-verdeling volgen en gebruikt Markov chain Monte Carlo (Gamerman & Lopes, 2006; Gilks et al., 1996) sampling om de posterior-verdeling van de parameters te schatten. We toonden aan dat de Bayesiaanse methode in staat is de ware waarden van de parameters terug te schatten in datasets van een realistische omvang. We raadden onderzoekers aan de voorgestelde methode consequent te gebruiken en de hele verdeling van SSRT's te beschouwen bij het analyseren van stopsignaaldata.

In het vierde hoofdstuk presenteerde ik BEESTS, een efficiënte en gebruiksvriendelijke software-implementatie van de Bayesiaanse parametrische methode die geïntroduceerd werd in hoofdstuk 3. BEEST heeft een eenvoudig te gebruiken grafische gebruikersinterface en voorziet gebruikers van kengetallen van de posterior-verdeling van de parameters, alsook verschillende diagnostische middelen om de kwaliteit van de parameterschattingen te beoordelen. We illustreerden het gebruik van BEESTS aan de hand van gepubliceerde stopsignaaldata. Deze software maakt het schatten van SSRT verdelingen ook toegankelijk voor de toegepaste wetenschap.

Deel II. Multinomiale *Processing-Tree*-Modellen

Het tweede deel van dit proefschrift richtte zich op modelselectie en het schatten van parameters voor multinomiale *processing-tree* (MPT) modellen. MPT-modellen zijn theoretisch gemotiveerde stochastische modellen voor categorische data. Als gevolg van hun eenvoud worden MPT-modellen frequent en in veel verschillende gebieden toegepast binnen de cognitieve psychologie (e.g., Batchelder & Riefer, 1999).

In het vijfde hoofdstuk introduceerde ik een Bayesiaanse aanpak voor het schatten van parameters van MPT-modellen. MPT-modellen worden gewoonlijk toegepast op geaggregeerde data, waarbij de onrealistische aanname wordt gedaan dat er geen heterogeniteit bestaat tussen de parameters (Hu & Batchelder, 1994). Onze voorgestelde Bayesiaanse aanpak houdt rekening met de heterogeniteit van de model parameters, die kan ontstaan als gevolg van individuele verschillen zowel tussen proefpersonen als items. We hebben het gebruik van de nieuwe methode geïllustreerd aan de hand van experimentele data verkregen uit de *pair-clustering*-taak (Batchelder & Riefer, 1980), een geheugentaak waarin proefpersonen semantisch gerelateerde woorden moeten onthouden. We raadden onderzoekers aan om de voorgestelde methode consequent toe te passen om de vertekening van parametersschattingen als gevolg van parameterheterogeniteit te voorkomen.

In het zesde hoofdstuk presenteerde ik verschillende procedures voor modelselectie voor MPT-modellen. Het onderwerp van kwantitatieve modelselectie krijgt van oudsher veel aandacht in de statistiek en tegenwoordig ook in de psychologie (Pitt & Myung, 2002). We richtten ons op twee populaire informatiecriteria, namelijk de AIC ("an information criterion", Akaike, 1973) en de BIC ("Bayesian information criterion", G. Schwarz, 1978), het *minimum-description-length*-principe (Grünwald, 2007), en de Bayes-factor verkregen met importance sampling (Hammersley & Handscomb, 1964). Naast de beschrijving van deze methode werd computercode geleverd die de praktische toepasbaarheid van de besproken modelselectiematen verhoogt.

Deel III. Correlaties, Partiële Correlaties en Mediatie-analyse

Het derde deel van dit proefschrift behandelde het schatten en toetsen van (partiële) correlaties.

In het zevende hoofdstuk onderzocht ik het onderscheidingsvermogen om de hypothese van perfecte correlatie te verwerpen binnen latente-variabelemodellen. In onderzoek waarbinnen hirarchische latente-variabelenmodellen worden gebruikt, wordt vaak gerapporteerd dat algemene intelligentie (g) een perfecte correlatie vertoont met lagere orde latente variabelen. Hieruit wordt vaak geconcludeerd dat g en de lagere orde latente variabele, zoals werkgeheugen, één en hetzelfde zijn. We hebben op basis van simulaties en gepubliceerde datasets onderzocht wat het onderscheidingsvermogen is om de gelijkheid van g en de lagere orde latente variabelen te verwerpen. De resultaten toonden aan dat het overgrote deel van de studies die een perfecte correlatie rapporteerden over onvoldoende onderscheidsvermogen beschikten om aan te tonen dat g en de lagere orde latente variabelen identiek zijn. We benadrukten het belang van het onderscheidingsvermogen in onderzoek naar de equivalentie van g en lagere orde latente variabelen.

In het achtste hoofdstuk behandelde ik een Bayesiaanse methode om de correlatiecoëfficiënt te corrigeren voor de onzekerheid van de observaties. De correlatiecoëfficiënt kan ernstig worden onderschat wanneer de observaties onderhevig zijn aan meetfouten. Hoewel verschillende methoden ontwikkeld zijn om hiervoor te corrigeren, worden deze amper toegepast in de psychologie. We richtten ons op een Bayesiaanse correctiemethode, ontwikkeld door Behseta et al. (2009), en toonden aan dat het toepassen hiervan tot een substantiële verhoging van de correlatie kan leiden tussen met ruis gemeten observaties. We raadden onderzoekers aan zich bewust te zijn van meetfouten en, indien mogelijk, de correlatiecoëfficiënt corrigeren voor de attenuatie die op kan treden als gevolg van de onzekerheid van de observaties.

In het negende hoofdstuk besprak ik een Bayesiaanse hypothesetoets voor mediatie. Om de relatie tussen verschillende variabelen te kunnen kwantificeren, voeren onderzoekers vaak een mediatie-analyse uit. In een dergelijke analyse verstuurt een mediator (zoals kennis van een gezond dieet) het effect van een onafhankelijke variabele (zoals instructie over een gezond dieet) naar een afhankelijke variabele (zoals de consumptie van fruit en groente). Vrijwel alle mediatie-analyses in de psychologie gebruiken frequentistische parameterschattingen en hypothesetoetsing. We ontwikkelden echter een Bayesiaanse hypothesetoets die gebaseerd is op de Jeffreys-Zellner-Siow prior (Rouder et al., 2009) en hebben de voordelen daarvan geïllustreerd aan de hand van gepubliceerde data.

Deel IV. Verbeteren van de Onderzoekspraktijk

Het vierde en laatste deel van dit proefschrift richtte zich op suboptimale onderzoekspraktijken binnen de psychologie.

In het tiende hoofdstuk introduceerde ik een nieuwe opzet voor samenwerking tussen voor- en tegenstanders van een empirische bevinding (*adversarial collaboration*). Een toenemend aantal wetenschappers suggereert dat horizontale saccadische oogbewegingen het ophalen van episodische herinneringen in geheugentaken faciliteren. Een aantal studies heeft dit verband echter niet weten te reproduceren. We hebben gepoogd deze inconsistente bevindingen op te lossen door een gezamenlijk onderzoek uit te voeren met voorstanders en sceptici. Onze aanpak combineerde elementen van een *adversarial collaboration* (Kahneman, 2003) en volledig confirmatorisch gepreregistreerd onderzoek (Wagenmakers et al., 2012). Conform de verwachtingen van de sceptici toonden de resultaten van Bayesiaanse hypothesetoetsen aan dat horizontale oogbewegingen de prestaties in geheugentaken niet verbeterden. Het toepassen van deze opzet vermindert de kans op het gebruik van *questionable research practices* en heeft de potentie om wetenschappelijk onenigheden op te lossen.

In het elfde hoofdstuk presenteerde ik een vergelijking tussen het statistisch bewijs dat wordt geleverd door p-waarden, effectgroottes en Bayes-factoren. Hierbij maakten we gebruik van 855 recent gepubliceerde t-toetsen in de psychologie. Hoewel de p-waarde en de Bayes-factor vrijwel altijd dezelfde hypothese ondersteunden, was er vaak sprake van verschil tussen de kracht van het geleverde bewijs; 70% van de p-waarden tussen 0.01 en 0.05 correspondeerden met Bayes-factoren die slechts anekdotisch bewijs leverden voor de alternatieve hypothese. Daarnaast concludeerden we dat de effectgrootte aanvullend bewijs kan leveren aan de p-waarde en de Bayes-factor.

In het twaalfde en laatste hoofdstuk, behandelde ik de meerweg-variantieanalyse (ANOVA). Veel onderzoekers realiseren zich niet dat de veelgebruikte meerweg-ANOVA onderhevig is aan kanskapitalisatie. We hebben het gebruik van sequentiele Bonferroni-correctie (Hartley, 1955) geïllustreerd. We lieten zien dat de conclusies die uit een ANOVA-design worden getrokken, vaak veranderen na toepassing van deze correctie en raadden de consequente toepassing ervan aan.

Acknowledgments

Graag wil ik beginnen met mijn dank te betuigen aan Eric-Jan voor de inspiratie, steun en begeleiding bij het schrijven van mijn proefschrift. Jouw diepgaande kennis vormde de basis, maar jouw toewijding aan de wetenschap, je tomeloze enthousiasme, je bereikbaarheid, je vriendschap en je oneindige optimisme maakten het verschil. Bovenal heb je me geleerd dat bitterballen en La Chouffe een volwaardige maaltijd zijn. Ik hoop nog lang en veel met je samen te mogen werken.

Mijn dank gaat ook uit naar Conor, mijn co-promotor; die me reality checks, een origineel perspectief en de nodige dosis humor bood. Ook wil ik Han bedanken, die mij de kans gunde om binnen zijn geweldige afdeling te mogen promoveren.

Het overgrote deel van dit proefschrift is geschreven in het gezelschap van een bonte verzameling mede-promovendi en labgenoten. Over de jaren heen veranderde de samenstelling, maar de constante factor was een groep mensen die op en buiten het werk elkaar steunden, uitdaagden, successen en tegenslagen deelden. Helen, mijn paranimf, dank voor je vriendschap, je steun, je kritische houding en je lekkere toetjes. Josine, dank voor je luisterend oor bij twijfels en onzekerheden—ik ga onze gezellige conferentieavonden missen. Alexander, dank voor je enorme behulpzaamheid, je groene curries en voor het voortzetten van een traditie van ongepaste humor in het lab. Jonathon, thank you for the Friday drinks and for listening to (and even showing some enthusiasm for) my constant babbling about stop-signal stuff. Don. Ruud, en Gilles—de oude generatie: dank voor het warme welkom in het lab, jullie kameraadschap en jullie humor. Dank ook aan de nieuwe generatie labgenoten: Ravi, Maarten en Tahira, aan jullie de taak het stokje over te nemen. And Michael, the honorary lab member, thanks for the funniest stories I have ever heard and for your advice and guidance throughout the years.

I am grateful to all the co-authors who have contributed to the papers that form the basis of this dissertation. This thesis would have been impossible without your ideas, knowledge, and advice.

Een speciale dankbetuiging voor Monique, mijn paranimf, die altijd met een wijntje klaar stond om naar mijn geklaag en twijfels te luisteren.

Ik ben dankbaar voor Carla en Gerard, mijn Nederlandse (schoon)ouders, die altijd klaar stonden en me door dik en dun gesteund hebben.

Eindeloos veel dank voor mijn moeder Edit, die me altijd in alles wat ik ondernam gesteund heeft en me heeft laten vertrekken naar een vreemd land. Köszönök mident, mama!

Harm, dank voor alles; voor de eindeloze steun, geduld, begrip, en liefde.

Publications

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. Accounting for measurement error and the attenuation of correlation: A Bayesian approach. *Manuscript in preparation.*

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. BayesMed: Default Bayesian hypothesis tests for correlation, partial correlation, and mediation. [Computer software manual]. (R package version 0.1.0). *Manuscript in preparation.*

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2013). Two birds with one stone: A preregistered adversarial collaboration on horizontal eye movements in free recall. *Manuscript submitted for publication.*

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L., & Wagenmakers, E.-J. (2013). Correction for multiple comparisons in the multiway ANOVA: Theoretical solution and practical consequences. *Manuscript submitted for publication.*

Boekel, W., Brown, S. D., Wagenmakers, E.-J., Matzke, D., & Forstmann, B. (2013). Probabilistic tractography does not reveal short-term structural changes in white matter pathways after practice on a decision making task. *Manuscript submitted for publication.*

Nuijten, M. B., Wetzels, R., Matzke, D., Dolan, C. V., & Wagenmakers, E.-J. (in press). A default Bayesian hypothesis test for mediation. *Behavior Research Methods.*

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (in press). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology.* Oxford, UK: Oxford University Press.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (in press). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika.*

Matzke, D., Lee, M. D., & Wagenmakers, E.-J. (2013). Getting started with WinBUGS. In M. D. Lee, & E.-J. Wagenmakers (Eds.), *Bayesian cognitive modeling: A practical course* (pp. 16-34). Cambridge, UK: Cambridge University Press.

Matzke, D., Lee, M. D., & Wagenmakers, E.-J. (2013). Signal detection theory: Parameter expansion. In M. D. Lee, & E.-J. Wagenmakers (Eds.), *Bayesian cognitive modeling: A practical course* (pp. 164-167). Cambridge, UK: Cambridge University Press.

Matzke, D., Lee, M. D., & Wagenmakers, E.-J. (2013). Multinomial processing trees. In M. D. Lee, & E.-J. Wagenmakers (Eds.), *Bayesian cognitive modeling: A practical course* (pp. 187-195). Cambridge, UK: Cambridge University Press.

Bakker, M., Cramer, A. O. J., Matzke, D., Kievit, R. A., van der Maas, H. L. J., Wagenmakers, E.-J., & Borsboom, D. (2013). Dwelling on the past. *European Journal of Personality, 27,* 120-144. Open peer commentary on Asendorp et al., "Recommendations for increasing replicability in psychology".

Matzke, D., Love, J., Wiecki, T., Brown, S. D., Logan, G. D., & Wagenmakers, E.-J. (2013). Releasing the BEESTS: Bayesian ex-Gaussian estimation of stop-signal reaction time distributions. *Frontiers in Quantitative Psychology and Measurement, 4:918*, doi: 10.3389/fpsyg.2013.00918.

Matzke, D., Dolan, C. V, Logan, G. D., Brown, S. D., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General, 142,* 1047-1073.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*, 291-298.

Matzke, D., Dolan, C. V., & Molenaar, D. (2010). The issue of power in the identification of "g" with lower-order factors. *Intelligence, 38*, 336-344.

Matzke, D. & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review, 16*, 798-817.