Psychological Methods
2011, Vol. 16, No. 1, 44–62

© 2011 American Psychological Association
1082-989X/11/$12.00 DOI: 10.1037/a0021765

Hierarchical Diffusion Models for Two-Choice Response Times

Joachim Vandekerckhove and Francis Tuerlinckx
University of Leuven

Michael D. Lee
University of California, Irvine

Two-choice response times are a common type of data, and much research has been devoted to the development of process models for such data. However, the practical application of these models is notoriously complicated, and flexible methods are largely nonexistent. We combine a popular model for choice response times—the Wiener diffusion process—with techniques from psychometrics in order to construct a *hierarchical diffusion model*. Chief among these techniques is the application of random effects, with which we allow for unexplained variability among participants, items, or other experimental units. These techniques lead to a modeling framework that is highly flexible and easy to work with. Among the many novel models this statistical framework provides are a *multilevel diffusion model*, *regression diffusion models*, and a large family of *explanatory diffusion models*. We provide examples and the necessary computer code.

Keywords: response time, psychometrics, hierarchical, random effects, diffusion model

Supplemental materials: http://dx.doi.org/10.1037/a0021765.supp

In his 1957 presidential address at the 65th annual business meeting of the American Psychological Association, Lee Cronbach drew a captivating sketch of the state of psychology at the time. He focused on the two distinct disciplines that then existed in the field of scientific psychology. On the one side, there was the experimental discipline, which concerned itself with the systematic manipulation of conditions in order to observe the consequences. On the other side, there was the correlational discipline, which focused on the study of preexisting differences between individuals or groups. Cronbach saw many potential contributions of these disciplines to one another and argued that the time and opportunity had come for the two dissociated fields to crossbreed: "We are free at last to look up from our own bedazzling treasure, to cast properly covetous glances upon the scientific wealth of our neighbor discipline. Trading has already been resumed, with benefit to both parties" (Cronbach, 1957, p. 675). Two decades onward, Cronbach (1975) saw the hybrid discipline flourishing across several domains.

In the area of measurement of psychological processes, there exists a schism similar to the one Cronbach pointed out in his presidential address. Psychological measurement and individual differences are studied in the domain of psychometrics, whereas cognitive processes are the stuff of the more nomothetic mathematical psychology. In both areas, statistical models are used extensively. There are common models based on the (general) linear model, such as analysis of variance (ANOVA) and regression, but we focus on more advanced, nonlinear techniques.

Experimental psychology has, for a long time, made use of process models to describe interesting psychological phenomena in various fields. Some famous examples are Sternberg's (1966) sequential exhaustive search model for visual search and memory scanning, Atkinson and Shiffrin's (1968) multistore model for memory, multinomial processing tree models for categorical responses (Batchelder & Riefer, 1999; Riefer & Batchelder, 1988), and the general family of sequential sampling models for choice response times (Laming, 1968; Link & Heath, 1975; Ratcliff & Smith, 2004). One property shared by these process models is that they give detailed accounts of underlying response processes. Such models are typically applied to data from single participants, and they are very successful in fitting empirical data.

In the correlational area, however, measurement models are dominant. Most well known among these is the factor analysis (FA) model, but models from item response theory (IRT) belong to this class as well. In the past decade, a lot of work has appeared showing the relationships between FA, IRT, and multilevel models. Rijmen, Tuerlinckx, De Boeck, and Kuppens (2003) showed that many IRT models are generalized linear mixed models and that the rest are nonlinear mixed models (NLMM; see also De Boeck & Wilson, 2004). Skrondal and Rabe-Hasketh (2004) offered an encompassing framework for FA models, IRT models, and multilevel models (called generalized linear latent and mixed

This article was published Online First February 7, 2011.

Joachim Vandekerckhove, Postdoctoral Fellow of the Research Foundation−Flanders (FWO), Department of Psychology, University of Leuven, Leuven, Belgium; Francis Tuerlinckx, Department of Psychology, University of Leuven; Michael D. Lee, Department of Cognitive Sciences, University of California, Irvine.

This research was supported by Grants GOA/00/02–ZKA4511, GOA/2005/04–ZKB3312, and IUAP P5/24 to Francis Tuerlinckx and Joachim Vandekerckhove; Grant K.2.215.07.N.01 to Joachim Vandekerckhove; and KULeuven/BOF Senior Fellowship SF/08/015 to Michael D. Lee.

The authors are indebted to Philip Smith for his insightful comments and to Gilles Dutilh, Roger Ratcliff, Jeff Rouder, and Eric-Jan Wagenmakers for sharing their data with us. This research was conducted utilizing high-performance computational resources provided by the University of Leuven. We also thank Microsoft Corporation and Dell for generously providing us with additional computing resources.

Correspondence concerning this article should be addressed to Joachim Vandekerckhove, Department of Psychology, University of Leuven, Tiensestraat 102 B3713, B–3000 Leuven, Belgium. E-mail: joachim.vandekerckhove@psy.kuleuven.be

models). The models that originated in correlational research are used to model individual differences. Often such models are less detailed and more general than the models discussed in the previous paragraph, but they are able to locate the main sources of individual differences.

Recently, some convergence between the experimental and the correlational areas has emerged. Batchelder and Riefer (1999; see also Batchelder, 1998; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002) introduced the concept of *cognitive psychometrics*. In cognitive psychometrics, models from cognitive psychology are used to capture specific interesting aspects of the data. These models typically assume that the data have been gathered with a specific paradigm (e.g., that they are binary choice response times). Although this necessarily makes the models less general than multipurpose statistical models, it provides the advantage of offering substantive insight into the data. Furthermore, ideas of hierarchical modeling have recently been introduced into the area of cognitive modeling, most notably by Rouder and colleagues (see e.g., Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder et al., 2007), who used hierarchical models as a statistical framework for inference, and also by Tenenbaum and colleagues (see e.g., Chater, Tenenbaum, & Yuille, 2006; Griffiths, Kemp, & Tenenbaum, 2008; see also Navarro, Griffiths, Steyvers, & Lee, 2006), who used hierarchical models as an account of the organization of human cognition.

Extending cognitive models to hierarchical models (or vice versa) is an important part of the trading between disciplines that Cronbach (1957) advocated. The benefits of the trade do go both ways: By extending process models hierarchically, experimental psychologists who use these models can take between-subjects variability into account and are in a better position to explain such interindividual differences. Correlational psychologists, on the other hand, could apply measurement models that are built upon firmly validated process models, often grounded in substantive theory.

In the present article, we aim to integrate both traditions further by extending hierarchically an important and popular process model: the diffusion model for two-choice response times. Even though choosing the diffusion model as our measurement level bears with it a number of implementation difficulties, we choose this model because of the interesting psychological interpretation of its parameters, which we explain in the next section. Additionally, choice response times—the combination of reaction time (RT) and accuracy data—are ubiquitous in experimental psychology, and we believe that a hierarchical extension of the diffusion model could be of considerable value to the field. In addition, a Bayesian approach is taken to fit the hierarchical extension of the diffusion model. Details on the practical implementation are provided as well.

In the sections that follow, we introduce the diffusion model for two-choice response times and then provide a detailed account of the hierarchical extension to the diffusion model. Then we describe two sample applications. We conclude with a discussion of our approach and of further possible applications.

The Diffusion Model

The diffusion model as a process for speeded decisions starts from the basic principle of *accumulation of information* (Laming, 1968; Link & Heath, 1975). When an individual is asked to make a binary choice on the basis of an available stimulus, the assumption is that evidence from the stimulus is accumulated over (continuous) time and that a decision is made as soon as an upper or lower boundary is reached. Which boundary is reached determines which response is given. The basic form of this model is often referred to as the *Wiener diffusion model with absorbing boundaries*.

Figure 1 depicts the Wiener diffusion process and shows the main parameters of the process. On the vertical axis there are the boundary separation α[1] indicating the evidence required to make a response (i.e., speed–accuracy trade-off) and the initial bias β, indicating the a priori status of the evidence counter as a proportion of α. If β is less than 0.5, this indicates bias for the response represented by the lower boundary. The absolute value of the starting position is $\alpha\beta = \zeta_{init}$, but we will generally not use this parameter. The arrow represents the average rate of information uptake, or drift rate δ, which indicates the average amount of evidence that the observer receives from the stimulus at each sampling. (The amount of variability in these samples, which makes the process stochastic, is a scaling constant that is typically set to 0.1 in the literature.) Finally, the short dashed line indicates the nondecision time τ, the time used for everything except making a decision (i.e., encoding the stimulus and physically executing the response). Table 1 gives a summary of the parameters and their classical interpretations.

The diffusion model owes much of its current popularity to the work of Ratcliff and colleagues (see e.g., Ratcliff, 1978; Ratcliff & Rouder, 1998; Ratcliff & Smith, 2004; Ratcliff, Van Zandt, & McKoon, 1999). An important contribution Ratcliff made was to incorporate trial-to-trial variance into the Wiener diffusion model, so that the parameters β, δ, and τ are not constant but vary from trial to trial. This conceptually significant extension has performed so remarkably well in the analysis of two-choice response time data that it is now sometimes referred to as the Ratcliff diffusion model (Vandekerckhove & Tuerlinckx, 2007; Wagenmakers, 2009). It has successfully been applied to data from experiments in many different fields, such as memory (Ratcliff, 1978; Ratcliff & McKoon, 1988), letter matching (Ratcliff, 1981), lexical decision (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2007), signal detection (Ratcliff & Rouder, 1998; Ratcliff, Thapar, & McKoon, 2001; Ratcliff et al., 1999), visual search (Strayer & Kramer, 1994), and perceptual judgment (Eastman, Stankiewicz, & Huk, 2007; Ratcliff, 2002; Ratcliff & Rouder, 2000; Thapar, Ratcliff, & McKoon, 2003; Voss, Rothermund, & Voss, 2004). The Ratcliff diffusion model is also one of few models that succeed in explaining all of the "benchmark" characteristic aspects of two-choice response time data—such as different response time distributions for correct and error responses, both of them positively skewed and the relation between their means dependent on parameters, with some minimum value below which there is no mass. In addition, the model has passed selective influence tests for its main parameters (see e.g., Voss et al., 2004), in which experimental manipulations are shown to affect only the relevant model parameters (e.g., changing from speed to accuracy instructions affects only the boundary separation parameter). Fitting the model to empirical data has become a topic

[1] Throughout, we use Greek letters to indicate unobserved parameters and Latin letters for running indices or observed variables.

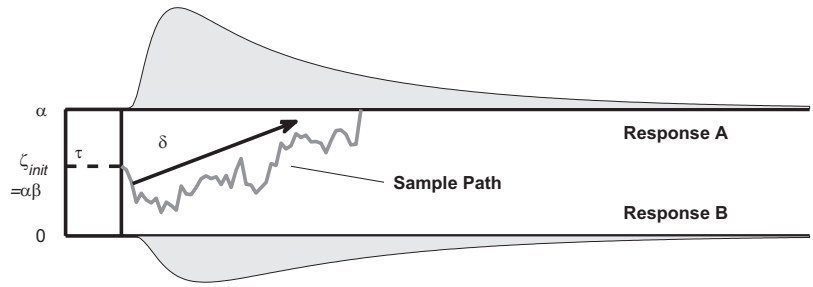


Figure 1. A graphical illustration of the Wiener diffusion model. α = boundary separation indicating the evidence required to make a response; β = initial bias indicating the a priori status of the evidence counter as a proportion of α; ζ_{init} = absolute value of the starting position; δ = average rate of information uptake; τ = time used for everything except making a decision.

of research in its own right (Donkin, Brown, Heathcote, & Wagenmakers, in press; Vandekerckhove & Tuerlinckx, 2008; Van Ravenzwaaij & Oberauer, 2009; Voss & Voss, 2007).

For our purposes, however, an important aspect of the diffusion model is that there is a mathematically tractable solution for the bivariate probability density function (PDF) of the response time and accuracy. In other words, it is possible to define explicitly a four-parameter density function, the "Wiener PDF," that describes the predictions of the model, given only the four parameters described in Table 1. The mathematical form of this PDF is given in the Supplementary Materials.

Finally, it should be kept in mind that, as with all statistical models, application of the diffusion model requires the user to assume that the process described here is the real process that brings about each individual response by a participant to a stimulus. If, for example, the experimental paradigm allows for self-correcting processes (e.g., a participant second-guessing a response), then one of the process assumptions of the diffusion model is violated and the model should not be applied.

A Hierarchical Framework for the Diffusion Model

Motivation

There are several motivations for making a hierarchical extension of a substantively generated model such as the diffusion model. The first and most important motivation is the fact that traditional applications of the diffusion model have been restricted to single participants (see e.g., Ratcliff & Rouder, 1998), and there has generally been no motivation to model interindividual differences in the decision process. The dearth of investigation into individual differences when applying process models is reminiscent of the schism between the experimental and correlational subdisciplines that Cronbach (1957, cf. supra) pointed out.

More recently, however, the diffusion model has been applied to study individual differences (see e.g., Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007; Ratcliff et al., 2004; Wagenmakers et al., 2007). The typical approach in such cases is to run multistep analyses: In a first step a specific model is fitted to data from each individual, and then inferences regarding individual differences are made on the basis of summary measures of the parameter estimates. An example of this approach can be found in Klauer et al. (2007), in which individual participants' parameter estimates are subjected to second-stage analysis using ANOVA.

However, data do not always allow for separate analyses per individual: Estimating the diffusion model's parameters typically requires a large number of data points (Wagenmakers, 2009), and in many experimental contexts it may be impractical or even impossible to obtain many data points within each participant. In particular, when studying higher level cognitive processes or emotions the stimulus material may simply not allow for the generation of hundreds of trials or for presenting stimuli more than once (see e.g., Brysbaert, Van Wijnendaele, & De Deyne, 2000; Klauer et al., 2007). Often, however, there are many participants in the sample. In cases such as these, it is natural to be interested in individual differences, but it is impossible to analyze the data separately for each participant, and the multistep procedure cannot be applied.

Another problem with the multistep procedures is that one may want to constrain parameters to be equal across participants. In this case, an analysis needs to involve all participants simultaneously, allowing some of the parameters to differ and others to be equal. However, such an approach may lead to a prohibitively large number of parameters. As will be argued in the following sections, a hierarchical approach may offer a solution by formalizing individual differences in a specific process model framework.

Table 1
Four Main Parameters of the Wiener Diffusion Model, With Their Substantive Interpretations

Symbol	Parameter	Interpretation
α	Boundary separation	Speed–accuracy trade-off (high α means high accuracy)
β	Initial bias	Bias for either response (β means bias toward Response A)
δ	Drift rate	Speed of information processing (close to 0 means ambiguous information)
τ	Nondecision time	Motor response time, encoding time (high means slow encoding, execution)

Uses of the Hierarchical Diffusion Model

In a hierarchical model (Gelman & Hill, 2007), it is assumed that participants are a randomly drawn sample from some partly specified population. Individual participants each have their own set of parameters, and because these participants are typically randomly selected from some larger population, the differences in parameter values between participants can be seen as a *random effect* in the statistical sense. A random effect occurs when experimental units are randomly drawn, interchangeable samples from a larger population. This may apply not only to participants but also to items, trials, blocks, and other units, as long as they are interchangeable samples. If the selected units comprise the entirety of the relevant population (about which one wants to make inferences), then a *fixed effect* is appropriate. In this way, individual differences can be explicitly permitted in a hierarchical model.

However, not only the person-specific parameters are important but the unknown characteristics of their population distributions are as well, characteristics such as the means, variances, and covariances, the latter two of which are indications of the magnitude (i.e., importance) of individual differences.[2] In a hierarchical framework, it is relatively easy to construct models in which some parameters are constrained to be equal across participants, whereas others may vary from individual to individual. Hierarchical models are ideally suited to handle data sets with few trials per participant (discussed earlier), even in the case in which single individuals do not provide enough information to estimate all model parameters and in which the number of data points per participant (or per cell of the design) seems absurdly small. Hierarchically extending the diffusion model leads to what we call the hierarchical diffusion model (HDM).[3]

Hierarchical models have proven useful in many areas of research. Some selected domains include psychological measurement when item response models have been used (see e.g., De Boeck & Wilson, 2004), educational measurement and school effectiveness studies (Raudenbush & Bryk, 2002), and longitudinal data analysis in psychology (Singer & Willett, 2003) and biostatistics (Molenberghs & Verbeke, 2006; Verbeke & Molenberghs, 2000).

In this article, we rely particularly on the framework proposed by De Boeck and Wilson (2004) for item response models. In their book, De Boeck and Wilson sharply distinguish between describing and explaining individual differences. Describing individual differences refers to the possibility of assuming population distributions for certain parameters and estimating some characteristics of these distributions. In such an approach, one merely acknowledges that differences between persons exist, and one quantifies the variability in the population (through the variances of the population distributions). However, in any scientific enterprise, the ultimate goal is not to simply observe differences but to attempt to explain why they occur. Individual differences can be explained by relating the person-specific parameters to predictors (see later). In doing so, we consider the variability in the population as to-be-explained, and by including a predictor in the model, we explicitly intend to decrease this unexplained variability.

It is important to emphasize that, although the previous discussion was centered on differences between persons, an HDM can equally well be applied to populations of items, trials, or indeed any experimental unit (e.g., subgroups within populations, items nested in conditions). Variability across these other experimental units can be captured in exactly the same way as is variability across persons. The sample applications make extensive use of this ability of HDMs.

The main difference between the approach of De Boeck and Wilson (2004) and our framework is that De Boeck and Wilson worked within a context of item response models: The data they considered are binary (or polytomous) responses of persons to a set of items. These item response models are logistic regression models or extensions and generalizations thereof that relate the responses (or more correctly: the probability of a certain response) to an underlying latent trait (i.e., the individual difference variable). There, the logistic regression model can be considered as the measurement model. In our case, the data are bivariate (choice response and RT) and the *measurement level* is the Wiener diffusion model, which is considerably more complex (both computationally, because the probability density function is mathematically somewhat intricate, and conceptually, because of having a process interpretation).

In the remainder of this section, we further elaborate on and apply the framework of De Boeck and Wilson (2004) to the diffusion model. This will be done by defining several basic building blocks that may be combined with the diffusion model in order to arrive at an HDM capable of describing and explaining interindividual differences. As it turns out, not only interindividual differences but other sources of variation may be tackled in such a way. Before doing so, however, we define some notation.

Notation

Suppose a person p (with $p = 1, \ldots, P$) is observed in condition i (with $i = 1, \ldots, I$) on trial j (with $j = 1, \ldots, J$) and the person's choice responses (corresponding to the absorbing boundaries) and response times are recorded, denoted by the random variables $X_{(pij)}$ and $T_{(pij)}$, respectively (realizations of these random variables are $x_{(pij)}$ and $t_{(pij)}$). Also, $Y_{(pij)}$ and $y_{(pij)}$ refer to the random vector $(X_{(pij)}, T_{(pij)})$ and the vector of realizations $(x_{(pij)}, t_{(pij)})$, respectively. Then $Y_{(pij)}$ would be distributed according to a Wiener distribution as follows:

$$Y_{(pij)} \sim Wiener(\alpha_{(pij)}, \beta_{(pij)}, \tau_{(pij)}, \delta_{(pij)}).$$

We use *Wiener distribution* as shorthand for the joint density function of hitting the boundary $X_{(pij)}$ at time $T_{(pij)}$. The distribu-

[2] Although it may seem that such an approach leads to even more parameters than when no population assumptions are made, invoking the population assumption actually reduces the number of effective parameters because it acts as a constraint on the person-specific parameters (this effect is in some cases also called *shrinkage to the mean*). A limiting case is when the variance of the population distribution is zero such that there are no individual differences and all person-specific parameters are exactly equal to the mean. Moreover, shrinkage is stronger for parameters of individuals who provide less information. For more information on hierarchical modeling and shrinkage we refer to Gelman and Hill (2007).

[3] There is some ambiguity here about the word *model*. In one sense, the diffusion model is a process model and the hierarchical extension is a statistical modeling tool. It is the combination of these two aspects, however, that makes the HDM a powerful framework.

tion is characterized by four basic parameters (explained earlier in The Diffusion Model section) that here carry a triple index, which means that, in principle, they can differ across persons, conditions, and trials. In some of the examples, we add additional indices to allow more nuanced differences. To avoid confusion with other subscripts, running indices will always be put between parentheses; for example, $\delta_{(i)}$ indicates the parameter δ that belongs to condition i, but ν_ε, $\nu_{(5)}$, and $\nu^{\text{descriptor}}$ are distinct, singular parameters.

Finally, it should be noted that we often recycle symbols for new models or new examples, so that a symbol used in one model may be redefined in another model to refer to something else.

Model Building Blocks

On the basis of the framework of De Boeck and Wilson (2004), we discern three types of useful model building blocks: levels of random variation, manifest predictors, and latent predictors. In order to render the discussion of these three aspects more concrete, we illustrate the theoretical concepts with the drift rate parameter of the diffusion model. We choose to limit the illustrations to a single parameter for reasons of clarity, but a similar story can be told for the other parameters, as will become obvious when we move to the applications later in the article.

Levels of random variation. The data may contain different levels of hierarchy. We have already implicitly referred to the most basic case when talking about individual differences: Imagine a situation in which a sample of individuals is measured repeatedly. In such a case, the data consist of two levels: At the higher level are the individuals, and at the lower level are the measurements within the persons.

As an example, consider drift rate $\delta_{(pij)}$. Assume that a set of persons p are presented with a series of stimuli j in a single condition (such that the index i for condition may be dropped). The drift rate $\delta_{(pj)}$ can then be written as follows:

$$\delta_{(pj)} = \nu_{(p)} + \varepsilon_{(pj)}, \tag{1}$$

where $\varepsilon_{(pj)} \sim N(0, \eta_\varepsilon^2)$ and $\nu_{(p)} \sim N(\mu_\nu, \sigma_\nu^2)$, with $\varepsilon_{(pj)}$ and $\nu_{(p)}$ independent. Here, the variance η_ε^2 represents trial-to-trial variability in drift rate within a person. This example is akin to the assumption of trial-to-trial variability made by Ratcliff (1978). The parameter μ_ν is the population average of individual drift rates, and σ_ν^2 is the variance of individual drift rates in the population. The importance of individual differences can be judged by comparing σ_ν^2 with η_ε^2: If σ_ν^2 is much larger than η_ε^2, this means that there are sizable individual differences, which is not the case if σ_ν^2 is much smaller than η_ε^2. Other methods of comparing the amounts of variability at different levels of hierarchy are *intraclass correlation coefficients* (see Shrout & Fleiss, 1979, for an overview).

There exist several alternative ways of writing the model in Equation 1. For instance, one could include the population average μ_ν directly into the linear decomposition (i.e., $\delta_{(pj)} = \mu_\nu + \sigma_\nu \nu_{(p)} + \eta_\varepsilon \varepsilon_{(pj)}$) and assume a mean of zero and unit variance for all random effects distributions.

Equation 1 can be extended readily to include fixed condition effects as follows:

$$\delta_{(pij)} = \gamma_{(i)} + \nu_{(p)} + \eta_\varepsilon \varepsilon_{(pij)}, \tag{2}$$

where $\gamma_{(i)}$ is a fixed condition effect. Hence, the mean drift rate in condition i for a person p depends on a fixed condition effect $\gamma_{(i)}$ and a random person effect $\nu_{(p)}$. A related model has been proposed earlier by Ratcliff (1985) and Tuerlinckx and De Boeck (2005).

Because individual differences are the main motivation for developing an HDM, we have thus far restricted the hierarchical structure to trials nested within persons (conditions are viewed as fixed effects). However, there is no reason to stop there if there is a sound reason for more complex forms of levels of random variation. For example, persons may be nested in groups and those groups nested in larger groups. In such a case, there are more than the traditional two levels in the data.

In addition, there is no reason to allow random effects only on the person side. On the condition or item side, it can make sense to allow for condition or random effects (see e.g., Baayen, Davidson, & Bates, 2008). In the types of applications we envision for the HDM, the stimulus material often consists of words or pictures (for such an application, see Dutilh, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2009). In psycholinguistics, for example, there has been some controversy over the modeling of word effects. In a seminal article, Clark (1973) strongly argued that stimulus words should be considered as randomly sampled from a population distribution as well. In such cases, the parameter $\gamma_{(i)}$ in Equation 2 can also be assumed to follow a normal distribution with mean μ_γ and variance σ_γ^2. This would yield a *crossed random effects* design (see e.g., Gonzalez, De Boeck, & Tuerlinckx, 2008; Janssen, Tuerlinckx, Meulders, & De Boeck, 2000; Rouder et al., 2007; see Vandekerckhove, Verheyen, & Tuerlinckx, 2010, for an HDM application). Similarly, conditions or items could be nested in categories that are in turn nested in larger categories.

Manifest predictors. By identifying and including levels of variation in the analyses, we describe individual differences or, if there are random item effects, differences between stimuli. We call this type of analysis *descriptive* because we are merely observing how the variability in the data is distributed among several sources. However, in a next step we want to *explain* the variability in parameters by using predictors (continuous or discrete or both). More broadly, interindividual, interstimulus, or less intuitively, intertrial variability (represented in random effects and their population variances) might be explained by regressing basic parameters on known *predictors* or *covariates*.

As an example of explaining interindividual variability, assume that the drift rate is person-specific and that there is a person covariate such as age available (with $A_{(p)}$ being the age of person p). We could then adopt the following model for the drift rate:

$$\delta_{(pij)} = \gamma_{(i)} + \beta_0 + \beta_1 A_{(p)} + \nu_{(p)} + \varepsilon_{(pij)}, \tag{3}$$

where β_0 and β_1 are the regression coefficients of the univariate linear regression of $\delta_{(pij)}$ on $A_{(p)}$ and $\nu_{(p)}$ is a person-specific error term with distribution $\nu_{(p)} \sim N(0, \sigma_\nu^2)$. The other parameters are defined as in Equation 2.

Alternatively, we may try to use covariates in order to explain some of the variability between items. For example, differences in recognizability between words may be related to their frequency of use (Vandekerckhove et al., 2010).

In sum, working with manifest predictors in the HDM means building a regression model for a random effect with known

predictors but unknown regression coefficients. Explaining variability in parameters through covariates will be an important theme in the examples in this article.

Latent predictors. De Boeck and Wilson (2004) showed that predictors do not necessarily need to be manifest; they may also be latent. That is, they may be unobserved but inferred from the data. For example, one might suspect that there exist two subgroups in the participant population, each with its own particular qualities (e.g., different people might use different strategies to respond to the stimuli). A mixture model for the diffusion model parameters may then be used to detect hidden groups or subpopulations in the data.

Latent predictors may even be continuous. Suppose that in a given experiment two (or more) dimensions of information processing are required and some conditions rely more on one dimension and other conditions more on the other dimension. Such a model for drift rate can be expressed as $\delta_{(pij)} = \gamma_{(i)} + \lambda_{(i)1}\nu_{(p)1} + \lambda_{(i)2}\nu_{(p)2} + \varepsilon_{(pij)}$, where $\lambda_{(i)1}$ and $\lambda_{(i)2}$ are the loadings of the underlying dimensions in condition i and $\nu_{(p)1}$ and $\nu_{(p)2}$ are the positions of person p on the two dimensions. Such a model can be called a *factor analysis diffusion model*.

We do not discuss latent predictor models further because they rapidly become complex and suffer from model identification issues and estimating their parameters is computationally intensive (at least, using current standard computational approaches for Bayesian inference).

Statistical Inference for HDMs

In the practical application of the HDM framework, statistical inference is performed using Bayesian statistical methods (see e.g., Box & Tiao, 1973; Gelman, Carlin, Stern, & Rubin, 2004; Gill, 2002). In this section, we provide some background on Bayesian methods that is required for interpreting the results of our analyses, as well as for the application of our software. We believe this background to be important because, although the philosophy behind Bayesian statistics is fairly straightforward and easy to explain, the computational techniques involved are not (further details about the computational challenges involved are given in the Appendix and Supplementary Materials).

Several reasons motivate our choice to use Bayesian inference. The Bayesian framework has many inherent advantages, such as the principled, consistent, and intuitive treatment of uncertainty concerning the parameters of the model. However, there are several advantages specific to the topic of the present article. Bayesian methods are most suited for flexible implementation of hierarchical models in particular (see also Gelman & Hill, 2007).

The diffusion model in itself, without any hierarchical extension, is already a computationally difficult model (see e.g., Navarro & Fuss, 2009; Tuerlinckx, 2004). These difficulties are exacerbated by even small increases in the hierarchical structure of the model (e.g., random drift; Ratcliff & Tuerlinckx, 2002; Tuerlinckx, 2004; Vandekerckhove & Tuerlinckx, 2007, 2008). Models with more extensive hierarchical structures (as discussed here) are often more interesting, but they rapidly become computationally intractable in the classical statistical framework where parameters have to be estimated using maximum likelihood methods. Take for example a crossed random effects model for drift rate (a random effect of person and of item), and assume for simplicity that the other parameters are kept constant across persons and items.

When applying such a relatively unembellished model to a data set of P persons and I items, one is confronted with a likelihood function that contains an integral of dimension $P + I$. Having P and I values both around 100 results in a 200-dimensional integral, the approximation of which is computationally prohibitive with standard numerical integration techniques (e.g., if one were to approximate each integral with a sum of 10 terms, then the total number of sums in each evaluation of the likelihood would be a disheartening 10^{200}). However, integrating over many distributions simultaneously is a basic modus operandi in Bayesian statistics, so that the addition of hierarchical structures such as random effects poses little additional burden.

Bayesian Basics

Bayesian methods depend on the computation of the posterior distribution of model parameters. That is, the probability distribution of the parameters, given the data. The posterior quantifies one's uncertainty about the model's parameters posterior to having observed the data. It is computed by updating one's prior knowledge about a parameter through observing the data (represented by the likelihood function). The posterior distribution can be obtained through Bayes' rule

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)},$$

where D and θ are generic notations referring to all the data and all the parameters, respectively; $p(D \mid \theta)$ is the likelihood; $p(\theta)$ is the prior distribution of the parameters; and $p(D)$ is the marginal probability of the data.

The posterior distribution $p(\theta|D)$ is a very intuitive measure, because it represents the uncertainty regarding the parameters after having observed the data. Its mean, the *expected a posteriori* (EAP) measure, is typically used as a point estimate of the parameter θ, with the standard deviation as a measure of uncertainty.

Often, a hypothesis can be tested by mere examination of the (marginal) posterior of one well-chosen parameter, such as a regression weight, a difference between two means, or a more complex function of one or more parameters. "Examining" a posterior distribution in this sense implies computing or estimating the probability mass of the parameter relative to some critical value. One might, for example, investigate the (signed) difference between two parameters, which should be, say, larger than zero. We then examine $p(\theta|D)$ by estimating $p(\theta > 0 \mid D)$.

Another popular method is to compute a parameter's .90 or .95 *highest posterior density region* (the smallest region of the posterior that contains the specified proportion of its mass) or its .90 or .95 *central credibility interval* (the central, contiguous region of the posterior that contains the specified proportion of posterior mass) and then evaluate whether 0 is contained within this region or not (see Box & Tiao, 1973; Gelman et al., 2004).

As straightforward as the main principle behind Bayesian statistics is, the practical task of estimating these probabilities can be somewhat daunting. For the case of the HDM, we have produced flexible software to facilitate putting it into practice. This software,

along with further information regarding the computational methods, is described in the Supplementary Materials.

Graphical Models

Advanced algorithms for sampling from a high-dimensional posterior distribution are implemented in the freely available statistical software package WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). WinBUGS can also be used easily to apply an HDM (for details, see the Supplementary Materials). In order to use WinBUGS, however, it is necessary to translate the hierarchical model into a directed acyclical graph, or *graphical model*. Graphical models (see Griffiths et al., 2008, and Lee, 2008, for accessible introductions) are a convenient formalism for describing the probabilistic relationships between parameters and data. In a graphical model, variables of interest are represented by nodes in a directed graph, with children depending on their parents. Circular nodes represent continuous variables, square nodes discrete variables, shaded nodes observed variables, and unshaded nodes unobserved variables. In addition, plates enclose parts of a graph to denote independent replications. An example of a basic graphical model is given in Figure 2. In this Figure, the data $y_{(ij)}$ are generated by a process that has parameters α, β, τ, and $\delta_{(ij)}$. $\delta_{(ij)}$ in turn is generated from a distribution with parameters $\nu_{(i)}$ and η. The graphical model is equivalent to the following set of assumptions (omitting for conciseness the prior assumptions):

$$y_{(ij)} \sim \textit{Wiener}(\alpha, \beta, \tau, \delta_{(ij)})$$

$$\delta_{(ij)} \sim N(\nu_{(i)}, \eta).$$

As soon as a model has been translated into a graphical model, it can be implemented in WinBUGS. In fact, WinBUGS code is only marginally more complex (because it includes prior information as well) than the enumeration of assumptions given earlier. In WinBUGS, it is straightforward to perform full Bayesian inference computationally, using standard Markov chain Monte Carlo methods to sample from the posterior distribution.

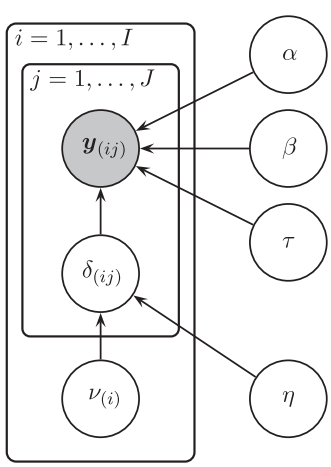


Figure 2. A sample graphical model. The shaded node $y_{(ij)}$ indicates the (bivariate) data. Nodes α, β, τ, and $\delta_{(ij)}$ are parameters of the distribution of $y_{(ij)}$. In turn, $\nu_{(i)}$ and η are parameters of the distribution of $\delta_{(ij)}$.

Evaluating Model Performance in the Bayesian Framework

After the posterior distributions of all parameters have been found, two aspects of model performance can be ascertained. To determine *relative model fit* across a series of models, the *deviance information criterion* (DIC) measure (Gelman et al., 2004) can be computed. This statistic can be considered as a Bayesian alternative to *Akaike's information criterion* (AIC; Akaike, 1973). Like the AIC, the DIC also expresses a balance between the model fit and its complexity. Smaller DICs indicate better fitting models.

Depending on context, we may choose the AIC over the DIC (Spiegelhalter, 2006). In general, if the objective of the analysis is to generalize to other populations (and hence the person- or item-specific parameters cannot be considered as given), then the AIC is more appropriate. In contrast, typically the concern is to quantify model fit for the particular data set at hand (and hence not to generalize to a larger population of persons or items), so the DIC is preferred.

In order to determine *absolute model fit*, however, we might apply *posterior predictive checks* (PPC; Gelman et al., 2004). The most basic type of PPC involves defining an interesting test statistic $G(\cdot)$ on the data and computing those statistics for the observed data (i.e., $G(y^{\text{obs}})$). Then the same statistic $G(\cdot)$ can be computed on a large number of data sets (say, 1,000) that are generated from the model, leading to a set $G(y^{\text{rep } 1}), \ldots, G(y^{\text{rep } 1000})$. Finally, the position of $G(y^{\text{obs}})$ in the distribution of $G(y^{\text{rep } 1}), \ldots, G(y^{\text{rep } 1000})$ then indicates the viability of the model with regard to the data.[4]

In general, PPCs can be applied to identify (graphically or numerically) misfit of a model to data. It can also be used to find specific loci of misfit of the model (Gelman, Goegebeur, Tuerlinckx, & Van Mechelen, 2000).[5] The method is somewhat controversial (see e.g., Bayarri & Berger, 2000; Evans & Jang, 2010), but it is very practical, easy to carry out, and highly flexible.

Application Examples

To illustrate the usefulness of the HDM framework, we now apply it to two data sets with very different designs, both of which seem usefully dealt with using HDMs. In the first application, we apply a series of HDMs to a benchmark data set concerning

[4] A more complex type of PPC can be defined as well (such that the test statistic is not pivotal but also depends on the parameters). However, this type of test statistic requires a (very time-consuming) reestimation of the model parameters for each replicated data set, which is why we do not apply it.

[5] Philip Smith (acting as reviewer) captured the importance of this practice particularly well, and we quote him here: "Conventional statistics treats the fits to the individual distributions as 'error' and then tends to ignore them—a practice that seems to carry over into the Bayesian setting. However, in the cognitive model the quality of these fits is absolutely central to the adequacy of the model, and any evidence of misfit wholly vitiates the enterprise. It follows that 'cognitive psychometrics' will require us to adopt different habits of data analysis and that the practice of what, in conventional statistics, is known as the 'analysis of residuals' will have to become the primary focus of modeling. This is why graphical contact between model and data is so important."

contrast perception and apply mainly regression-type analyses, as well as trial-to-trial variability in drift rate, initial bias, and nondecision time. We also include a basic hierarchical structure, namely, the addition of random variability over conditions. Note that in this example, we do not consider individual differences or complex hierarchical structures. We use the first application mainly to demonstrate the basic features of the diffusion model, the Bayesian modeling approach, the principles of Bayesian model selection, and the relative ease with which these otherwise involved analyses can be performed.

In the second application, the data set is different because it has more participants ($P = 9$), and we construct an HDM that permits the simultaneous analysis of data from different individuals. Thanks in part to the Bayesian framework, we are able to define a statistic that directly quantifies the effect under consideration and estimate the distribution of its size in the population.

In both of the examples, we make a large number of assumptions regarding the structure in the data. We, sometimes somewhat arbitrarily, select whether certain parameters are allowed to change between experimental units, whether effects are fixed or random, and which parametric forms are taken by population distributions or regression functions. These assumptions are often debatable, but the central point to be made here is that a wide variety of assumptions can be made explicit in the HDM framework with relative ease. For the purposes of illustrating this, it is not of crucial importance exactly which assumptions are made. In real-world applications the tools that are developed allow for the checking of assumptions.

Example 1: Fixed Effects and Nonlinear Regression

Introduction. The first application example involves a data set in a contrast discrimination task that has become something of a benchmark for RT model fitting (Ratcliff & Rouder, 1998). An important reason for this is that these data clearly show the standard RT phenomena for which any model of choice RT should be able to account (see the earlier section entitled The Diffusion Model). In the experiment, three participants saw 10 blocks of trials (after two practice blocks). Each trial consisted of a grid with 75% gray pixels and the remaining 25% either black or white. There were 33 different proportions of black versus white pixels (evenly spaced, so that the middle level was 50% black and 50% white), and the task was to determine whether this proportion was a draw from a "bright" or from a "dark" distribution. Additionally, in half of the blocks, the participants were asked to respond as accurately as possible (accuracy condition [AC]), and in the other half, to be as fast as possible (speed condition [SC]).

The research goal in this study was to study the relationship between stimulus brightness and drift rate. A link was clearly confirmed, and it was found that this link was nonlinear in nature. Here, we go two steps further. First, we formalize the nonlinear relation using a cumulative Weibull link function, which is a nonlinear function that is common in the vision literature (see e.g., Wichmann & Hill, 2001). Then we investigate the effect of the instruction (AC vs. SC) on the relation between stimulus brightness and drift rate—because it could be hypothesized that a task instruction affects the rate of information processing in the decision-making system, which would show itself in a different shape of the link function. We focus on a single participant's data.

Models. As an introductory example, we apply a basic HDM to these data. However, the features added to the Wiener diffusion are not limited to the trial-to-trial variance used by Ratcliff and Rouder (1998): We also implement a nonlinear regression and allow a difference between the instruction conditions. Specifically, let $C_{(s)} = s/32$ ($s = 0, \ldots, 32$) be a measure of intensity (i.e., brightness) and i ($i = 1$ for AC; $i = 2$ for SC) be the instruction condition. The index j ($j = 1, \ldots, J_{(si)}$) separates different trials for a given intensity-by-instruction combination. Considering only one participant (so that we can drop the index p), we have the following model for the observed response vector $\mathbf{Y}_{(sij)}$:

$$\boldsymbol{Y}_{(sij)} \sim \mathit{Wiener}(\alpha_{(sij)}, \beta_{(sij)}, \tau_{(sij)}, \delta_{(sij)}).$$

We assume that the parameter α was subject to only a fixed effect of instruction, as seen in

$$\alpha_{(sij)} = \alpha_{(i)}$$

(because it captures the speed–accuracy trade-off, which is exactly what the instructions were), whereas β (initial bias), τ (nondecision time), and δ (rate of information uptake) are made subject to random effects of trial, as seen in

$$\beta_{(sij)} \sim U(\pi_{(i)}^{low}, \pi_{(i)}^{high}),$$

$$\tau_{(sij)} \sim N(\theta, \chi^2), \text{ and}$$

$$\delta_{(sij)} \sim N(\nu_{(si)}, \eta^2).$$

The mean of the trial-to-trial distribution of δ is additionally subject to a random condition effect, as seen in

$$\nu_{(si)} \sim N(\mu, \sigma_\varepsilon^2),$$

which introduces a key ability of the HDM. Here it becomes most clear why these models are called hierarchical, because "layers" of randomness are added incrementally (in this case, one at the condition level and one at the trial level). (Note that this means that we predict $\nu_{(si)}$ to be identical if the same combination of stimulus and instruction is presented twice. However, that does not imply that the drift rate $\delta_{(sij)}$ will be identical as well, because it is still a draw from $N(\nu_{(si)}, \eta^2)$.) In this context, $\nu_{(si)}$ is the more interesting parameter, because it pertains directly to the quality of the stimulus but is not confounded by random trial-to-trial fluctuations (whose magnitude is captured by η^2).

Furthermore, the range of the trial-to-trial distribution of β is made subject to a fixed effect of condition (following Ratcliff & Rouder, 1998). The model with this set of assumptions will be called BM1 (see Figure 3 for a graphical model illustration). Note that model BM1, although acknowledging the possibility of a difference between the 66 drift rates, contains no information to quantify the differences between the conditions (i.e., nothing links it to brightness): We are ignoring the available covariate information and are assuming that all 66 drift rates are drawn from the same pool (with mean μ and variance σ_ε^2).

We model the experimental manipulation of speed-versus-accuracy instruction as a fixed effect because these levels constitute an exhaustive list of the possibilities—they are not a random selection from a larger pool of possible instructions. By contrast, the differ-

ence between trials is modeled as a random effect, because we are not (currently) interested in the effects of particular trials—we assume those to be a selection from a larger pool of possible trials. In the first model, we pretend to have no knowledge of the nature of the differences in stimulus intensity but assume them to be a random draw from the large set of possible stimulus intensities.

To continue, we can define multiple competing models. Ratcliff and Rouder's (1998) model did not restrict the across-condition drift rate distributions.[6] In contrast, we now define a second model in which we formalize the connection between stimulus intensity and drift rate with a Weibull regression. Formally, we redefine

$$\nu_{(si)} = \nu^{low} + (\nu^{high} - \nu^{low})\,\{1 - \exp[-(C_{(s)}/\nu^{\text{scale}})^{\nu^{\text{shape}}}]\} + \varepsilon_{(si)},$$

with upper and lower asymptotes ν^{high} and ν^{low}, shape parameter ν^{shape}, scale ν^{scale}, and an error term $\varepsilon_{(si)} \sim N(0, \sigma_\varepsilon^2)$. Note that although σ_ε^2 in BM1 indicated the across-condition variability in $\nu_{(si)}$, here it refers to the residual variability after accounting for the nonlinear effect of the brightness condition. Importantly, the ability to quantify residual variability after controlling for the effect of the brightness condition allowed us to investigate the magnitude of interstimulus variability that is not due to an experimental manipulation (but rather due to other manipulations or due to random, uncontrolled differences between stimuli). The choice of the Weibull function as a regression function is somewhat arbitrary and can be adapted at will. For our purposes, the Weibull is a nonlinear function with two horizontal asymptotes, a scale, and a shape. The second model, now completely specified, is called Model BM2.

However, we had originally set out to investigate the effect of the experimental instruction on the drift rates. We therefore constructed a third model in which we allowed a difference in the drift rate distributions as a function of the instruction condition, using the link function

$$\nu_{(si)} = \nu_{(i)}^{low} + (\nu_{(i)}^{high} - \nu_{(i)}^{low})\,\{1 - \exp[-(C_{(s)}/\nu_{(i)}^{\text{scale}})^{\nu_{(i)}^{\text{shape}}}]\} + \varepsilon_{(si)}.$$

Note that we added subscripts i to the Weibull's parameters to indicate their dependence on the instruction condition. This model is called BM3. The three models are displayed as graphical models in Figure 3.

Results. The means and standard deviations of the marginal posteriors for some of the parameters in each model are given in Table 2. Several results are immediately obvious. First, the parameters α_{AC} and α_{SC} are very different: The boundary separation in the SC is much lower than in the AC, in all models. This is consistent with the interpretation of that parameter. Second, the posterior standard deviations are generally small compared with the posterior means (EAPs), indicating narrow distributions and therefore reliable estimates. Third, the mean nondecision time θ is about 270 ms, and its trial-to-trial standard deviation χ is about 41 ms, which is normal in this type of application. Fourth, estimates of the parameters that do not pertain to the Weibull regression remain more or less constant between different models, indicating that the models' restrictions on the drift rate parameters do not lead to trade-off effects for other parameters. Finally, the parameter σ_ε, which indicates the amount of unexplained variability in drift rates, strongly differs between models—apparently the added covariates do explain a fair amount of variance. We therefore used the difference in unexplained stimulus variance as a quality measure of the Weibull regression, using a statistic akin to the familiar statistic

$$R^2 = 1 - \left(\frac{\sigma_{res}}{\sigma_{total}}\right)^2,$$

where in this case σ_{total} was σ_ε in BM1 and σ_{res} was σ_ε in the model with which we wanted to compare. Given a series of samples from each of these parameters, we computed a posterior mean for the proportion of variance that was explained by the addition of the nonlinear regressions.[7] In BM2, the proportion of variance explained was 96.50%, whereas in BM3 it was as high as 99.96%.

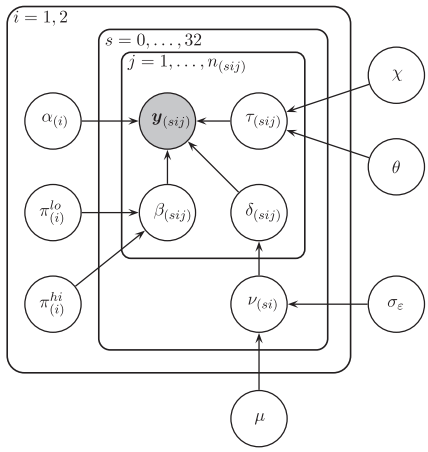
In Table 3, the parameters of the Weibull regression are shown for BM2 and BM3. It is clear from the posterior means and standard deviations that the Weibull regression function is quite different between the two instruction conditions. In particular, the upper and lower asymptotes are more extreme in the SC, and the function is somewhat steeper in that condition as well. In fact, according to the analysis, $P(\nu_{\text{SC}}^{\text{shape}} > \nu_{\text{AC}}^{\text{shape}} \mid D) \approx .9590$. Figure 4 shows the Weibull regression lines for each instruction condition and the individually estimated drift rates. To compare the performance of the three models, we computed DIC values for each model and found that BM3 performed best (DIC was 3,373.40, 2,087.60, and 642.63, for BM1, BM2, and BM3, respectively). Finally, we provide a graphical illustration of the model fit using posterior predictive checks. We generated, on the basis of 1,000 samples from the posterior, 1,000 posterior predictive response probabilities (i.e., expected probability of a "bright" response) and posterior predictive mean RTs for each of the 66 conditions. Figure 5 shows the generated probabilities (as gray dots) overlaid with the observed data (black line). As can be seen, the changes in response probability and RT over brightness conditions are well captured by the model (even the somewhat capricious behavior near the extremes is well within the posterior uncertainty of the fitted model).[8]

Conclusion. Although the model we have applied to these data is quite different from the one used by Ratcliff and Rouder (1998), our conclusions generally echo theirs, with one significant difference: We find an effect of instruction on drift rate. The Weibull link functions are manifestly different between the instruction conditions—evidently the rate of information accumulation is not entirely independent of the participants' motivations

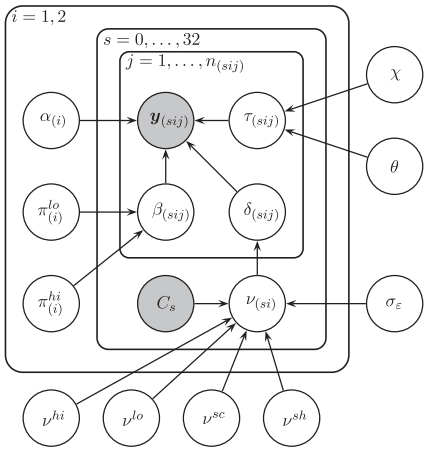
[6] Ratcliff and Rouder (1998) do mention that they could (in principle) further simplify the model by implementing a regression of mean drift rate as a linear function of the probability that the stimulus was a draw from the "bright" distribution, that is, $\nu_{(ps)} = \zeta_{(p)0} + \zeta_{(p)1}P_{(s)}$, with $P_{(s)} = N(s \mid \kappa_1, \phi)/[N(s \mid \kappa_1, \phi) + N(s \mid \kappa_2, \phi)]$ and $\kappa_1 = 5/8$, $\kappa_2 = 3/8$, and $\phi = 3/16$. However, they did not actually apply this regression. Similar nonlinear regression models of drift rate (cast in a classical statistical framework) have been investigated by Smith, Ratcliff, and Wolfgang (2004) and Palmer, Huk, and Shadlen (2005).

[7] Because we were not dealing with a linear model and were in fact comparing across models with strongly different assumptions, the R^2 statistic used here is not exactly the same as the familiar statistic. However, for the purpose of comparing model fits, we believe it is a succinct summary measure.

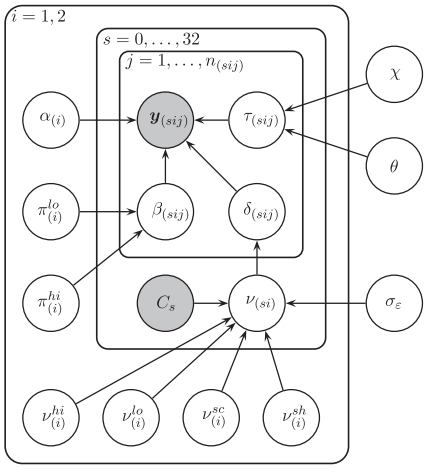
[8] Further posterior predictives for this data set and a similar model are reported in Vandekerckhove, Tuerlinckx, and Lee (2008).



(a) Model BM1



(b) Model BM2



(c) Model BM3

Figure 3. A graphical model representation of each of the models for the first application. (a) In Model BM1, τ, β, and δ are indexed for instruction conditions i, stimuli s, and trials j, indicating that we allowed these parameters to be different for each condition-by-stimulus-by-trial combination. α, on the other hand, is indexed by only i, for condition. The population parameters for β, called π^{high} and π^{low}, also depend on only the instruction condition. ν may differ between stimuli and instruction conditions, but its population distribution has invariant parameters μ and σ_ε. Finally, η, χ, and θ do not vary between conditions or stimuli either. (b) In Model BM2, the mean of ν is determined not by one but by four parameters (the parameters of the Weibull link). (c) In Model BM3, those Weibull parameters are moved inside the plate over i, so that they may differ between instruction conditions.

(see Vandekerckhove & Tuerlinckx, 2007, for analogous results). Although we do not describe them here, the results were analogous for the other two participants.

In addition to the relative ease with which it was applied (only 30 or so lines of highly redundant WinBUGS code; see the Supplementary Materials), the model just described contains two properties that are fundamentally novel in the domain. Trial-to-trial variance and constraints on parameters have already been applied (see e.g., by Vandekerckhove & Tuerlinckx, 2007), but the application of Bayesian inference and in particular the addition of random effects on the condition (stimulus) level are new. Random effects are an important modeling construct that has not previously been considered in this context. In the next example, we focus more closely on the addition of random effects.

Example 2: ANOVA and Random Person–Domain Effects

Introduction. In the previous application, we focused on single participants (mainly because the data set contained only three participants in total). However, one of the more significant advantages of the hierarchical setting is that it allows for the simultaneous analysis of many participants' choice response time

Table 2
Some Parameter Estimates for the First Application

Parameter	EAP			Posterior *SD* (× 100)			Parameter interpretation
	BM1	BM2	BM3	BM1	BM2	BM3	
α_{AC}	0.2192	0.2314	0.2199	0.4344	0.5465	0.4507	Caution (AC)
α_{SC}	0.0501	0.0511	0.0502	0.0984	0.0956	0.0990	Caution (SC)
θ	0.2791	0.2769	0.2789	0.1758	0.1726	0.1779	Mean nondecision time
χ	0.0412	0.0404	0.0410	0.0934	0.0917	0.0937	Intertrial *SD* of nondecision time
η	0.1261	0.1425	0.1273	0.7848	0.8903	0.7960	Intertrial *SD* of drift rate
π_{AC}^{low}	0.3522	0.3431	0.3515	0.9947	0.9402	1.0429	Lower limit of initial bias (AC)
π_{AC}^{high}	0.5755	0.5832	0.5757	0.8259	0.6975	0.8218	Upper limit of initial bias (AC)
π_{SC}^{low}	0.4498	0.4492	0.4495	0.9888	1.0170	0.9984	Lower limit of initial bias (SC)
π_{SC}^{high}	0.4779	0.4771	0.4776	0.9670	0.9835	0.9865	Upper limit of initial bias (SC)
σ_{ε}	0.4008	0.0732	0.0064	3.8249	1.1022	0.4323	Residual interstimulus *SD* of drift rate

Note. BM1–BM3 refer to the model names. EAP = expected a posteriori; AC = accuracy condition; SC = speed condition.

data. For example, diffusion parameters could be kept constant across items for each participant, but individual participants' parameters would be considered random draws from a population distribution. This would be a very typical hierarchical model; van der Linden (2007) would call this a *population model*. Analyzing data from different participants simultaneously results in greater stability for the statistical inferences. In particular, it becomes possible to fit the model even with relatively few data points per participant.

However, because it remains unreasonable to assume that all parameters stay exactly constant across trials, we combined mixing over trials with mixing over persons. This yielded a *multilevel random effects* design wherein the parameters of individual participants' mixing distributions were themselves draws from a population-level distribution. A graphical representation of this multilevel diffusion model is given in Figure 6. The data set to which we applied this model was taken from a change detection study (Vandekerckhove, Panis, & Wagemans, 2007, data used with permission). For a detailed description of the research questions, the reader is referred to Vandekerckhove et al. (2007). For the purposes of our demonstration, it suffices to know that the difficulty of a visual detection task was manipulated in a 2 × 2 factorial design and that there were nine participants. The independent variables of interest are called *Q*, for *quality*, and *T*, for *type*. Because the manipulations were all intended to affect higher order properties of the stimulus, we expected changes in drift rate but not in any other variable. The main research question was whether there is an effect of *T* on detection performance and whether this effect is independent of *Q*. It is hence a straightforward ANOVA-type design, and we were interested in the main effect of *T* and the *T* × *Q* interaction. The factorial design is given in the second and third columns of Table 4.

Model. We define only one model, which includes a hierarchical structure that simultaneously incorporates different participants' data. The assumptions of this population-hierarchical model (PHM) are as follows.

The core of the PHM is still the Wiener diffusion model, so that each individual data point $y_{(pij)}$ follows a Wiener distribution with four parameters, as seen in

$$y_{(pij)} \sim Wiener(\alpha_{(pij)}, \beta_{(pij)}, \tau_{(pij)}, \delta_{(pij)}),$$

with indices *p* for participants ($p = 1, \ldots, 9$), *i* for conditions ($i = 1, \ldots, 4$), and *j* for trials ($j = 1, \ldots, 80$). We assume an unbiased diffusion process: $\beta_{(pij)} = .5$.

The hierarchical structure now contains two levels of random variation: the trial level and the participant level. At the trial level, the nondecision time τ and drift rate δ are assumed to vary between trials, as seen in $\tau_{(pij)} \sim N(\theta_{(p)}, \chi^2_{(p)})$ and $\delta_{(pij)} \sim N(\nu_{(pi)}, \eta^2_{(p)})$. In contrast, we treated the boundary separation as constant across trials (for a given participant), as in $\alpha_{(pij)} = \alpha_{(p)}$.

At the participant level, although the boundary separation is assumed constant across trials, at a higher level of heterogeneity, interindividual differences arise. We treated all interindividual differences as random effects (because we knew that participants were a random sample from a larger population), as seen in $\alpha_{(p)} \sim N(\mu_\alpha, \sigma^2_\alpha)$.

Further interindividual differences are allowed; that is, the parameters of the intertrial mixing distributions (those for τ and δ, described earlier) depend on participant *p* and may depend on condition *i* (in the case of drift rate), as seen in $\theta_{(p)} \sim N(\mu_\theta, \sigma^2_\theta)$ and $\nu_{(pi)} \sim N(\mu_{\nu(i)}, \sigma^2_{\nu(i)})$.

Note that the fixed effect of condition *i* remains present in the dependence of $\mu_{\nu(i)}$ on *i*, but now it exists on the population level. It is not necessary to define the factorial structure of the conditions in the experiment at this stage; because the parameters in a linear model that quantify main effects and interactions are mere linear combinations of the data (i.e., the mean in each condition), we could compute posterior distributions for each conditional mean first and derive the posterior distributions of the ANOVA parameters later[9] (see the next section).

Finally, although it was not the primary focus of the present analysis, the trial-to-trial variability parameters were also given population distributions, as seen in $\chi_{(p)} \sim N(\mu_\chi, \sigma^2_\chi)$ and $\eta_{(p)} \sim N(\mu_\eta, \sigma^2_\eta)$. We did this primarily to formalize our knowledge that participants were a random selection from a pool yet did exhibit interindividual differences.

[9] Parameters that are not directly estimated themselves but are obtained from transformations or combinations of other parameters are sometimes called *derived parameters* or *structural parameters* (Congdon, 2003; Jackman, 2000).

Table 3
Parameter Estimates of the Weibull Regression in the First Application

Parameter	EAP		Posterior *SD* (× 100)		Parameter interpretation
	BM2	BM3	BM2	BM3	
ν_{AC}^{high}	0.4132	0.3292	2.2774	1.4160	Upper asymptote (AC)
ν_{SC}^{high}	0.4132	0.5110	2.2774	2.5016	Upper asymptote (SC)
ν_{AC}^{low}	−0.4296	−0.3516	2.4513	1.4473	Lower asymptote (AC)
ν_{SC}^{low}	−0.4296	−0.5654	2.4513	2.7277	Lower asymptote (SC)
ν_{AC}^{scale}	0.5258	0.5259	1.0179	0.5080	Location (AC)
ν_{SC}^{scale}	0.5258	0.5214	1.0179	0.6037	Location (SC)
ν_{AC}^{shape}	5.4092	4.4127	70.6052	24.1439	Steepness (AC)
ν_{SC}^{shape}	5.4092	5.2268	70.6052	42.4271	Steepness (SC)

Note. BM2 and BM3 refer to the model names, and BM2 does not allow for differences between AC and SC. EAP = expected a posteriori; AC = accuracy condition; SC = speed condition.

In all cases, population distributions are truncated to a reasonable interval (for numerical stability; see the Supplementary Materials for the intervals).

Results. We were interested in two different aspects of the results. For the experimenter, it is important to know whether a main effect of T and a $T \times Q$ interaction appear on the mean drift rates $\mu_{\nu(i)}$. From a general-interest perspective, we were additionally interested in the population-level variability of the different parameters.

Summary statistics of the obtained drift rate population distributions (per condition) are given in Table 1. It can be seen that the distributions differ strongly between conditions. In order to more precisely investigate our hypotheses, we transformed the drift rate distributions into ANOVA contrast parameters that exactly quantified the effects in which we were interested. First, the main effect of T is given by the contrast $\zeta_T = (\mu_{\nu(1)} + \mu_{\nu(3)}) - (\mu_{\nu(2)} + \mu_{\nu(4)})$, for which the posterior distribution is shown in Figure 7. It is clear from that figure that there is strong evidence for a main effect of T, averaged over levels of Q. Indeed, $P(\zeta_T < 0 \mid D) \approx 0$. Similarly, in the second panel in Figure 7, we can confirm that there is a main effect of Q, because for $\zeta_Q = (\mu_{\nu(1)} + \mu_{\nu(2)}) - (\mu_{\nu(3)} + \mu_{\nu(4)})$, $P(\zeta_Q < 0 \mid D) \approx .994$. To investigate the interaction, we computed the interaction contrast $\zeta_I = (\mu_{\nu(1)} - \mu_{\nu(2)}) - (\mu_{\nu(3)} - \mu_{\nu(4)})$. As it turned out, $P(\zeta_I < 0 \mid D) \approx .886$, providing only marginally convincing evidence for an interaction. The interaction pattern is shown in Figure 8.

The population variability in the parameters is directly quantified by their variance parameters. Although not the focus of the present experiment, each of these parameters has a unique interpretation that may be relevant in other contexts (here their main purpose was to account for extraneous variability in the data). For example, the EAP of the interindividual standard deviation of boundary separation σ_α estimates to 0.0541 and that of the interindividual standard deviation of mean nondecision time σ_α =

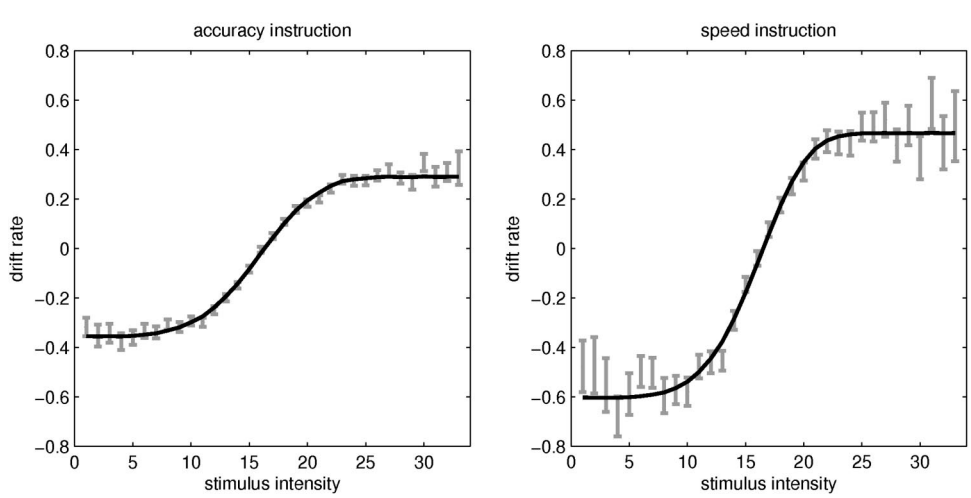


Figure 4. The Weibull regression on drift rate. Both panels, one for each instruction condition, contain individually estimated drift rates for each level of stimulus intensity ($\nu_{(si)}$ from BM1), with error bars extending one posterior standard deviation in both directions. Overlaid are the Weibull regression lines (from BM3), based on the posterior means of $\nu_{(i)}^{high}$, $\nu_{(i)}^{low}$, $\nu_{(i)}^{scale}$, and $\nu_{(i)}^{shape}$ from BM3. The Weibull function captures the effect of stimulus intensity well.

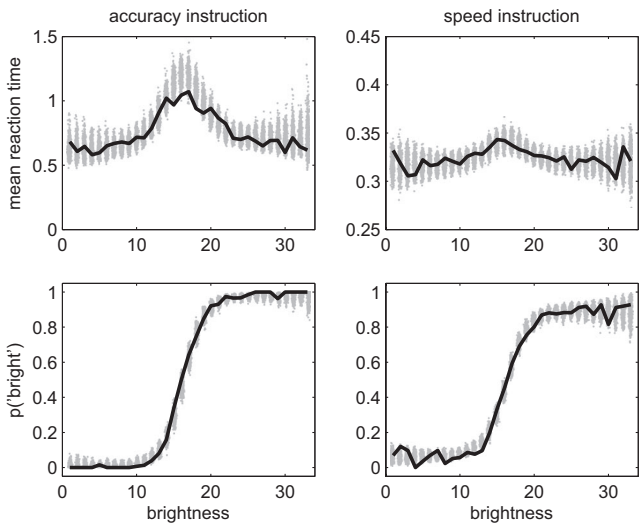


Figure 5. A comparison of posterior predictive mean reaction times (top panels) and the probability of "bright" responses (bottom panels) with the empirically observed quantities. The gray crosses indicate the model predictions (which have some variability due to posterior uncertainty), and the black lines indicate the data. Note the different scales in the two reaction time graphs.

0.0663. By themselves, these numbers mean little, but given these estimated population distribution parameters and their remaining uncertainty (i.e., the posterior variance of these parameters), we could now depict the distribution of the model parameters in the population by computing posterior predictive distributions. Take, for illustration, the population distribution of α. Given a single sample $\mu_\alpha^{(s)}$ from the posterior distribution of μ_α, and a single sample $\sigma_\alpha^{(s)}$ from the posterior distribution of σ_α, one can generate a single sample $\alpha^{(s)}$. Repeating this procedure many times yields a vector of α values that are sampled from the population distribution. Thus, a sufficiently high number of samples obtained this way represents the expected population distribution of α. Figure 9

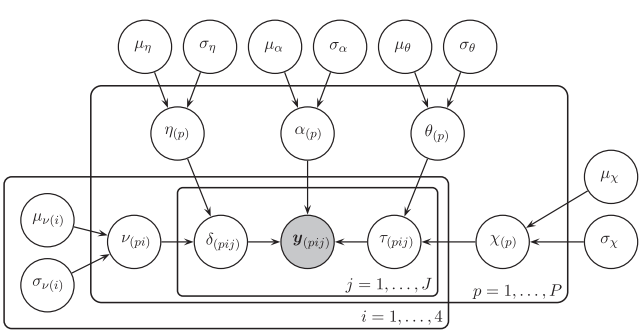


Figure 6. The graphical model representation of the population-hierarchical model. The shaded node indicates the observed data, and its direct contributors are (clockwise starting from the left) the drift rate δ, boundary separation α, and nondecision time τ. Drift rate and nondecision time have across-trial means (ν and θ, respectively) and standard deviations (η and χ, respectively). The rest of the parameters belong to prior distributions. Subscripts (and plates) indicate repetitions of the parameter across participants p, conditions i, and trials j. See text for more details.

Table 4
Posterior Distributions of the Mean Drift Rate in the Population, per Condition

Condition i	T	Q	$\mu_{\nu(i)}$	$\sigma_{\nu(i)}$
1	1	0	0.0870	0.0357
2	0	0	−0.0347	0.0555
3	1	1	0.2700	0.0558
4	0	1	0.0363	0.0448

Note. Where $T = 0$, the population distribution of the drift rate has much mass around 0. T = type; Q = quality; $\mu_{\nu(i)}$ = posterior mean; $\sigma_{\nu(i)}$ = posterior standard deviation.

shows these predicted population distributions for the α and θ parameters. The parameter estimates for the nine participants in the experiment are shown as circles under the distribution curve. Figure 9 invites several insights with substantive implications regarding the hierarchical diffusion model. First, it can be seen that population variability in α is quite large—it spans almost the entire range of α—whereas it is comparatively small for θ.[10] Also, although the θ parameters seem to follow a bell-shaped distribution, α parameters are more spread out and even appear to occur in clusters.

The large interindividual differences in boundary separation support a general argument of choice RT modeling as an improvement over models for accuracy or RT only: If there are large differences in how cautiously participants respond to stimuli, pure accuracy or pure RT data may paint a deceptive picture. The diffusion model allows one to quantify differences in participant caution, and the HDM framework can be used to model boundary separation (or indeed any other diffusion model parameter) at the population level.

To conclude the discussion of the parameter estimates, it can be interesting to compare the size of the variance on the trial level with that on the population level. In the present model this is possible for the nondecision time and for the drift rate in each condition. For the nondecision time, the interindividual standard deviation of θ, σ_θ, is estimated at 0.0659, and the average intertrial standard deviation, μ_χ, is of a similar magnitude: 0.0807. For the drift rates, however, the average intertrial standard deviation, μ_η = 0.3500, clearly exceeds the interindividual standard deviations, $\mu_{\nu(i)}$: 0.0858, 0.1496, 0.1388, and 0.1169.

Finally, in order to evaluate absolute model fit, we generated posterior predictives by simulating 1,000 new data sets from the joint posterior distribution of the parameters (i.e., we generated data as predicted by the fitted model). Then we pooled these data sets in each person-by-condition cell of the design and constructed a histogram for each of these pooled data sets. Figure 10 shows this histogram for each condition for three participants. The black line is a (smoothed) histogram of the simulated data, whereas the gray bars indicate the real data. The RTs for error responses were given a negative sign, so that the inverted distribution on the left side of the vertical axis indicates the error response distribution. Each cell contains 80 responses. The figure does not seem to betray any large systematic misfit of the model.

[10] See Matzke and Wagenmakers (2009) for plausible ranges of diffusion model parameters, as found in published studies.

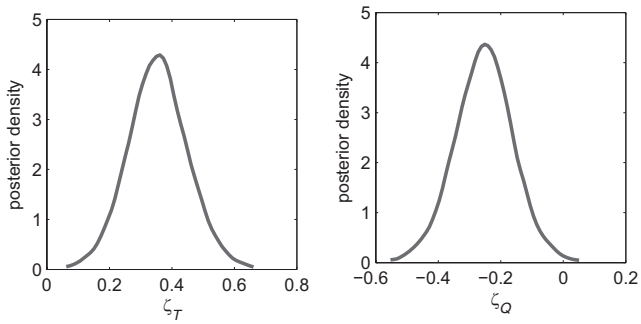


Figure 7. Posterior distributions of the main effect contrast parameters of the analysis of variance in Example 2. In the left panel, ζ_T has all its mass above zero, indicating strong evidence for a main effect of type (T), averaged over levels of quality (Q). In the right panel, a negative main effect of quality is clear, because more than 99% of the mass of the posterior distribution of ζ_Q is below zero. See text for more details.

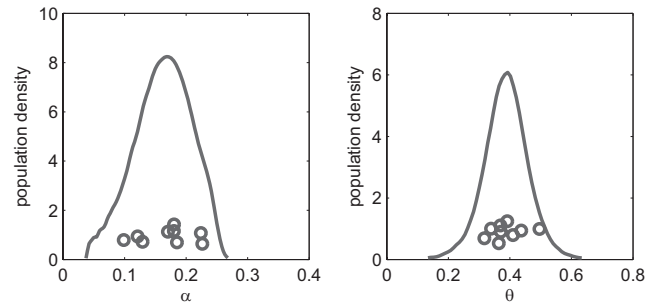


Figure 9. Posterior predictive population distributions of two parameters of the population-hierarchical model (person-specific mean nondecision time θ and person-specific boundary separation α). The population distribution of θ is bell-shaped and narrow. The population distribution of α is wider.

Conclusion. In this advanced application, we applied a population-hierarchical model to choice response time data. We computed ANOVA-style contrasts for the 2 × 2 factorial design and found two main effects of the independent variables on the drift rate parameter. The population-hierarchical diffusion model is especially noteworthy because it combines information from different participants (and conditions) in a single model, allowing for more reliable parameter estimates and hypothesis tests at the population level.

Software Implementation

We believe the hierarchical extension of the Wiener diffusion process has much potential for the field of cognitive science. However, applying this model in practice is difficult and may be prohibitively onerous for many researchers. For this reason, it is important also to publish computer software to aid in the application of the HDM.

Our software implementation is not a full software package but rather a plug-in to an existing software package, WinBUGS. A guide in the Supplementary Materials details how the software and plug-ins need to be installed and also presents some examples of usage. Note that the software is dependent on Microsoft Windows and cannot currently be made to run optimally on other systems.

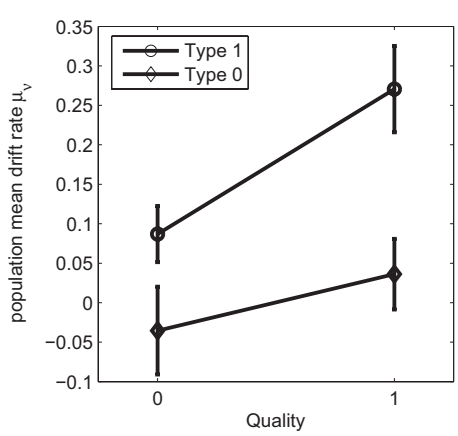


Figure 8. Drift rate parameters for the second application. Error bars extend one posterior standard deviation in either direction. The main effect of type is clearly seen. There is evidence of a superadditive interaction.

Discussion

We have introduced a hierarchical extension of the Wiener diffusion model for two-choice response times (hierarchical diffusion model, or HDM). With two examples, we have demonstrated the feasibility of the HDM. In strictly formal terms, the HDM is just another nonlinear mixed model, but it is special because it has a bivariate random variable at the measurement level. As a new data analysis approach, it is characterized by great flexibility compared with existing treatments of choice response time data. Additionally, by using the Wiener diffusion process as the measurement level, the hierarchical analysis can be performed on parameters that have well-defined substantive interpretations.

The substantively interesting process interpretation of the diffusion model parameters is important for several reasons. First, it is particularly appealing in the context of Borsboom's (2006) view that the fact that measurement models lack substantive foundation is the main reason psychometrics has had a limited impact. That is, although the analysis of choice response time data in a hierarchical framework has already been addressed in the psychometric literature (Klein Entink, Kuhn, Hornke, & Fox, 2009; van der Linden, 2007), our use of a popular process model such as the diffusion model is novel. At the same time, of course, the process interpretation limits the applicability of the present approach, because one is now required to assume that the process assumptions are not violated. As a result, certain types of data (e.g., long response times that would belie the assumption that a single information-gathering process without regulatory processes is generating the data) cannot be dealt with by the HDM.

Second, thanks to the substantively interesting process interpretation of the diffusion model parameters, the HDM framework is an instance of cognitive psychometrics, a relatively young subdiscipline of psychology. In this subdiscipline, models of cognition are extended to encompass individual differences (i.e., participants are no longer considered as mere replications of one another;

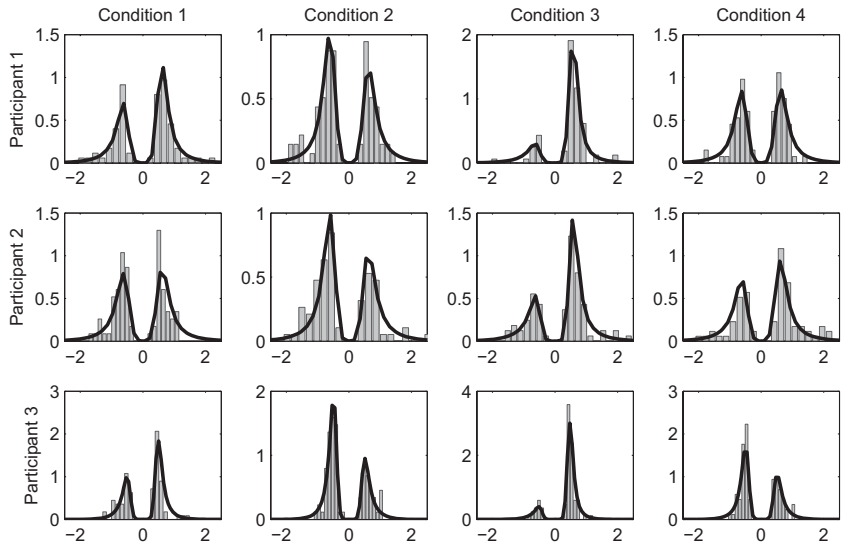


Figure 10. Posterior predictive data distributions for three participants. The flipped distributions indicate error reaction times. The full line is the model prediction, and the gray bars are the observed data. Although the correspondence is not perfect, there do not seem to be any systematic deviations of the model from the data.

Batchelder, 1998) in order to allow for population-level inferences. This strategy has gained some momentum recently, with several articles applying hierarchical models to pool data over participants (see e.g., Hoffman & Rovine, 2007; Morey, Pratte, & Rouder, 2008; Morey, Rouder, & Speckman, 2008; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010), but the approach is presently far from mainstream. We believe that cognitive psychometrics—the result of trading between subdisciplines as advocated by Cronbach (1957)—has many possible applications in a wide variety of domains where psychological measurement is used.

We have elected to implement the HDM using Bayesian statistical methods. This choice was influenced by many factors, both practical and principled. An important corollary of the Bayesian framework is that results from such an analysis have direct and often intuitive interpretations. In one of the examples, we derived posterior distributions of ANOVA contrasts, from which we could directly draw (probabilistic) inferences regarding the hypotheses at hand.

In order to facilitate the dissemination of hierarchical models (i.e., cognitive psychometrics) into mainstream cognitive science, we have provided software with which a hierarchical model for two-choice response time data can be put into practice. Although this software has some limitations (in particular, somewhat inefficient sampling due to WinBUGS's use of general-purpose rather than tailor-made sampling algorithms), we believe it may be useful for a wide audience.

References

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, Hungary: Akademiai Kiado.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–105). New York, NY: Academic Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment, 10,* 331–344.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models (with discussion). *Journal of the American Statistician Association, 95,* 1127–1142.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440.

Box, G. E. P., & Tiao, G. C. (1973). Bayesian inference in statistical analysis. Reading, MA: Addison–Wesley.

Brysbaert, M., Van Wijnendaele, I., & De Deyne, S. (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica, 104,* 215–226.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10,* 287–291.

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

Congdon, P. (2003). *Applied Bayesian models.* New York, NY: Wiley.

Cronbach, L. J. (1957, November). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1975, February). Beyond the two disciplines of scientific psychology. *American Psychologist, 30,* 116–127.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E.-J. (in press). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Psychonomic Bulletin & Review.*

Dutilh, G., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J.

(2009). A diffusion model account of practice. *Psychonomic Bulletin & Review, 16,* 1026–1036.

Eastman, K., Stankiewicz, B., & Huk, A. (2007). Optimal weighting of speed and accuracy in a sequential decision-making task. *Journal of Vision, 7,* 428.

Evans, M., & Jang, G. H. (2010). Invariant p-values for model checking and checking for prior-data conflict. *Annals of Statistics, 38,* 512–525.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., & Van Mechelen, I. (2000). Diagnostic checks for discrete-data regression models using posterior predictive simulations. *Applied Statistics, 49,* 247–268.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, MA: Cambridge University Press.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton, FL: CRC Press.

Gonzalez, J., De Boeck, P., & Tuerlinckx, F. (2008). A double-structure structural equation model for three-mode data. *Psychological Methods, 13,* 337–353.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods, 39,* 101–117.

Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. *American Journal of Political Science, 44,* 375–404.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Klauer, K. C., Voss, A., Schmitz, F., & Teige-Mocigemba, S. (2007). Process components of the implicit association test: A diffusion-model analysis. *Journal of Personality and Social Psychology, 93,* 353–368.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., & Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods, 14,* 54–75.

Laming, D. R. J. (1968). *Information theory of choice-reaction times.* London, England: Academic Press.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15,* 1–15.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika, 40,* 77–105.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review, 16,* 798–817.

Molenberghs, G., & Verbeke, G. (2006). *Models for discrete longitudinal data.* New York, NY: Springer.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology, 52,* 376–388.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near-liminal performance. *Journal of Mathematical Psychology, 52,* 21–36.

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

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.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision, 5,* 376–404.

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

Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review, 88,* 552–572.

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

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.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). Diffusion model account of lexical decision. *Psychological Review, 111,* 159–182.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review, 95,* 385–408.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356.

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.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367.

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.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review, 9,* 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review, 102,* 261–300.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review, 95,* 318–339.

Riefer, D. M., Knapp, B., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations. *Psychological Assessment, 14,* 184–201.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods, 8,* 185–205.

Robert, C. P., & Casella, G. (2003). *Monte Carlo statistical methods.* New York, NY: Springer.

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

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

Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika, 72,* 621–642.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Singer, J. D., & Willett, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York, NY: Oxford University Press.

Skrondal, A., & Rabe-Hasketh, S. (2004). *Generalized latent variable*

modeling: Multilevel, longitudinal, and structural equation models. Boca Raton, FL: Chapman and Hall/CRC Press.

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research, 44,* 1297–1320.

Spiegelhalter, D. (2006). *Some DIC slides* [PowerPoint slides]. Available from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/DIC-slides.pdf

Sternberg, S. (1966, August 5). High speed scanning in human memory. *Science, 153,* 652–654.

Strayer, D., & Kramer, A. (1994). Strategies and automaticity: I. Basic findings and conceptual framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 318–341.

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.

Tuerlinckx, F. (2004). The efficient computation of the distribution function of the diffusion process. *Behavior Research Methods, Instruments, & Computers, 36,* 702–716.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70,* 629–650.

Vandekerckhove, J., Panis, S., & Wagemans, J. (2007). The concavity effect is a compound of local and global effects. *Perception & Psychophysics, 69,* 1253–1260.

Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review, 14,* 1011–1026.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods, 40,* 61–72.

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

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

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72,* 287–308.

Van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology, 53,* 463–473.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York, NY: Springer.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32,* 1206–1220.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39,* 767–782.

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.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2007). A diffusion model account of criterion manipulations in the lexical decision task. *Journal of Memory and Language, 58,* 140–159.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the expectancy valence model of the Iowa gambling task. *Journal of Mathematical Psychology, 34,* 14–27.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling and goodness-of-fit. *Perception & Psychophysics, 63,* 1293–1313.

(*Appendix follows*)

Appendix

Technical Details

Technical details concerning some of our analyses are provided here in order to avoid cluttering the main text. We begin with some details regarding the Bayesian statistical framework, then discuss the examples from the main text within this more technical context.

The Bayesian Framework

Computation in the Bayesian Framework

As mentioned in the main text, Bayesian methods require the computation of the posterior distributions of model parameters, $p(\theta \mid y) = p(y \mid \theta)p(\theta)/p(y)$. Typically, however, these distributions are mathematically complex, and it is nontrivial to compute summary statistics of the (often high-dimensional) posterior $p(\theta \mid y)$. In many cases it is comparatively much easier to generate randomly drawn samples from this posterior distribution—in no small way because this does not require the evaluation of $p(y)$. Summary statistics of a sufficiently large sample can then be used to accurately represent the posterior distribution. A class of general methods for sampling from a complex distribution are *Markov chain Monte Carlo* (MCMC; Robert & Casella, 2003) techniques and in particular the Metropolis–Hastings algorithm, in which samples are drawn from some initial distribution that is not the posterior, and this initial distribution is changed as successive samples are drawn in such a way that after a number of iterations the samples drawn are truly samples from the posterior.

Importantly, this means that the first samples are not representative of the posterior distribution. That is, the convergence of the MCMC algorithm to a stationary distribution is not immediate and may in fact take some time. It is therefore necessary to check whether convergence has occurred before the output of the algorithm may be considered to be draws from the posterior distribution. Our preferred method of assessing convergence is to run a small number of different chains (say, six). After discarding a certain number of iterations, called the *burn-in*, we test whether the remaining draws from the different chains are in fact draws from the same distribution. For this, we use the $\hat{R}$ criterion (Gelman et al., 2004), a statistic that is similar to the F statistic in analysis of variance: $\hat{R}$ is large if the between-chain variance is larger than within-chain variance, and it approaches unity when the different chains have converged to the same distribution. $\hat{R}$ values lower than 1.1 are typically considered satisfactory. In our applications, we always ran six independent chains and obtained $\hat{R}$ values under 1.05 for all parameters.

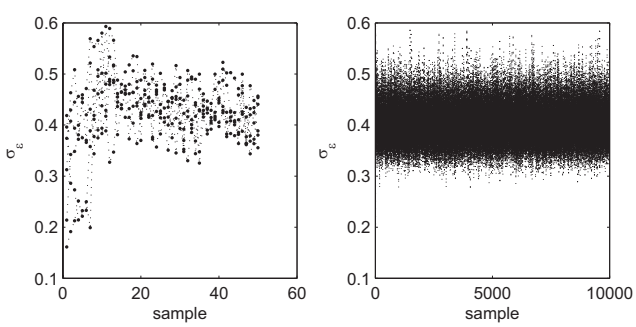


Figure A1. An example of a sample chain. The parameter depicted is σ_ε, the condition-level variability, in model BM1. Although the chain looks well mixed in the right panel, the left panel shows that the first few iterations are still distinctly overdispersed. However, because this overdispersion quickly dissipates and is comfortably within the burn-in zone, it does not affect the quality of the sampling.

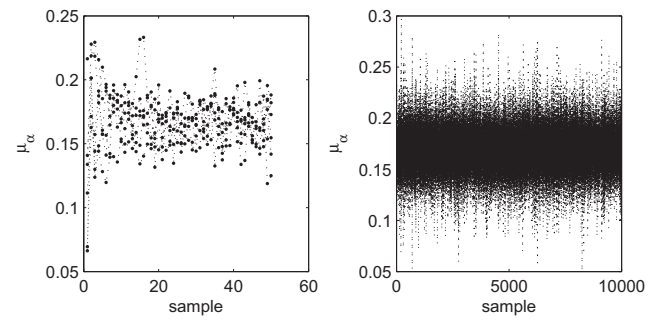


Figure A2. An example of a sample chain from model population-hierarchical model. The parameter shown is μ_α, the population mean of the boundary separations. In the left graph, the first 50 iterations are depicted, showing that the six chains rapidly converge to the same region. In the right graph, the entire chains are shown, showing that the six chains appear to be sampling from the same stationary distribution.

In this Appendix, we make brief notes regarding the technical results of our examples (because we believe this is good practice), even though they indicated good convergence and stable estimates in each of the analyses.

Priors

A final technical matter is the choice of prior distributions $p(\theta)$, which are required for the computation of the posterior distributions. Choice of priors is a somewhat controversial topic in statistics, mainly because they seem to require a somewhat subjective judgment on the part of the researcher. It is also a matter of debate

(*Appendix continues*)

whether it is possible (and desirable) to construct prior distributions that are entirely uninformative. For our purposes, we have struck a compromise: We selected theoretically informed limits on the possible range of each parameter and chose a prior distribution that reflected equiprobability of all values in that range (i.e., a uniform, or "flat," prior). Researchers who have more prior information regarding the distribution of parameters can implement this in our software easily. In practice, however, the influence of even quite informative prior distributions on the posterior distributions is small compared with the weight of the likelihood of the data.

Convergence Results for the Example Applications

Example 1

For each model, we drew samples from the posterior distribution using the software described in the Supplementary Materials. We ran six independent chains of 10,000 iterations, and 5,000 of these iterations were discarded as burn-in. We computed the $\hat{R}$ statistic for each parameter and confirmed that they were all lower than 1.05, indicating good convergence of the chains. Visual inspection of the chains indicated no issues (see Figure A1 for an example).

Example 2

We again drew 10,000 samples in each of six chains. We discarded the first 5,000 samples as burn-in and computed the $\hat{R}$ statistic for each parameter. The statistic was always lower than 1.05. Again, no issues with convergence were noticeable in a visual inspection of the sample chains (see Figure A2 for an example).

Received August 27, 2008
Revision received May 11, 2010
Accepted August 13, 2010 ■

Online First Publication

APA-published journal articles are now available Online First in the PsycARTICLES database. Electronic versions of journal articles will be accessible prior to the print publication, expediting access to the latest peer-reviewed research.

All PsycARTICLES institutional customers, individual APA PsycNET® database package subscribers, and individual journal subscribers may now search these records as an added benefit. Online First Publication (OFP) records can be released within as little as 30 days of acceptance and transfer into production, and are marked to indicate the posting status, allowing researchers to quickly and easily discover the latest literature. OFP articles will be the version of record; the articles have gone through the full production cycle except for assignment to an issue and pagination. After a journal issue's print publication, OFP records will be replaced with the final published article to reflect the final status and bibliographic information.