Using hierarchical Bayesian methods to examine the tools of decision-making

Michael D. Lee[*] Benjamin R. Newell[†]

Abstract

Hierarchical Bayesian methods offer a principled and comprehensive way to relate psychological models to data. Here we use them to model the patterns of information search, stopping and deciding in a simulated binary comparison judgment task. The simulation involves 20 subjects making 100 forced choice comparisons about the relative magnitudes of two objects (which of two German cities has more inhabitants). Two worked-examples show how hierarchical models can be developed to account for and explain the diversity of both search and stopping rules seen across the simulated individuals. We discuss how the results provide insight into current debates in the literature on heuristic decision making and argue that they demonstrate the power and flexibility of hierarchical Bayesian methods in modeling human decision-making.

Keywords: hierarchical Bayesian models, Bayesian inference, heuristic decision-making, take-the-best, search rules, stopping rules.

1 Introduction

To the cognitive scientist individual differences in behavior can be both intriguing and annoying. We are all familiar with the subjects in our experiments who “don’t do what they are supposed to do.” Sometimes these different patterns of behavior are simply noise (the subject was on a cell phone during the experiment), but often they are due to legitimate responses that our theories and models failed to anticipate or cannot explain.

The field of judgment and decision making is no exception to the challenge of individual differences. As Brighton and Gigerenzer (2011) mention in passing, even a theory as important and influential as Prospect Theory (Kahneman & Tversky, 1979) typically predicts only 75%-80% of decisions in two-alternative choice tasks, and many models do much worse. How should we, as a field, treat these individual differences and the challenges they present for our models and theories?

One emerging approach for tackling these issues is to use hierarchical Bayesian methods to extend existing models, and apply them in principled ways to experimental and observational data (e.g., Lee, 2008, 2001; Nilsson, Rieskamp, & Wagenmakers, 2011; Rouder & Lu, 2005, van Ravenzwaaij, Dutilh, & Wagenmakers, 2011; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, 2010). This approach not only provide tools for interpreting individual differences, but also facilitates theory building by providing a model-based account of why individual differences might arise. We think it is an especially interesting, important, and promising approach, because it deals with fully developed models of cognition, without constraints on the theoretical assumptions used to develop the models.[1] Taking existing successful models of cognition and embedding them within a hierarchical Bayesian framework opens a vista of potential extensions and improvements to current modeling, because it provides a capability to model the rich structure of cognition in complicated settings.

To demonstrate the application of hierarchical Bayesian methods to the modeling of heuristic decision-making, we use a standard experimental setup that requires subjects to make judgments about the relative magnitudes of two objects (size, distance, fame, profitability, and so on). To perform these judgments it is often assumed that subjects search their memory, or external sources of information, for cues to help differentiate objects. For example, an inference about the relative size of two cities might be facilitated by cues

MDL and BJN acknowledge support from Australian Research Council Grant DP110100797, and MDL acknowledges support from the Air Force Office of Scientific Research FA9550-11.

[*] Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92697-5100. Email: mdlee@uci.edu.

[†] University of New South Wales.

[1] Note, in particular, that we are *not* requiring the models of decision-making we consider to be so-called “rational” models that assume people are Bayesian reasoners (Griffiths, Kemp, & Tenenbaum, 2008). We are not using Bayesian inference as a metaphor for human cognition. Rather, we are using it as a statistical and modeling tool to relate process models of cognition to behavioral data (Kruschke, 2010; Lee, 2011). In fact, it is the hierarchical (or multi-level) aspects of our modeling that provide the important theoretical capabilities, with the Bayesian inference simply providing a complete and effective approach for analyzing these models.

indicating which of the two cities is a capital, has an airport, a university, and so on (Gigerenzer & Goldstein, 1996). Judgments are then determined by rules that use the presence or absence of cues to provide estimates of the desired criterion (i.e., number of inhabitants).[2]

Such tasks, although apparently simple, incorporate several important features that need specification in theories and models that wish to describe how subjects perform them. In this paper, we present two simple case studies: the first focuses on information search, and the second focuses on stopping rules. We show how Bayesian inference allows information about these psychological processes at the level of both individuals and groups to be extracted from basic behavioral data, and how hierarchical extension of the models allow deeper psychological questions about *why* there is variation in search and stopping to be addressed.

2 Modeling Search

Our case studies rely on an environment widely used in the literature, in which 83 German cities are described by 9 different cues, and the task is decide which of two cities has the larger population (Gigerenzer & Goldstein, 1996). In a binary comparison task like this there are different properties of cues that are relevant to the likelihood of aiding judgment, and thus to the order in which one might search through cues. For example, the cue "Is the city the national capital?" is often very useful because in most cases capital cities are the largest in the country (with notable exceptions, such as Canberra), so if the capital cue is present it is highly likely that the city with that cue has more inhabitants. However, for the vast majority of cases a comparison on this cue will draw absent values for both cities, because the majority of cities in a country are not the capital. Thus, in terms of how often a cue will provide you with diagnostic information, the "national capital" cue is not at all useful.

Formally these qualities of cues can be thought of as the validity and discriminability rate of a cue in a binary comparison. Discriminability is the rate at which a cue distinguishes between two objects. Validity is the rate at which a cue, given it discriminates, indicates correctly which of two objects should be chosen. There is evidence from experimental investigations of search rules (e.g., Newell, Rakow, Weston, & Shanks, 2004; Rakow, Newell, Fayers, & Hersby, 2005) that both discriminability and validity can be relevant to search, and that individual differences and task constraints might influence the extent to which one or the other, or some combination of the two dictates search through cues (e.g., Martignon & Hoffrage, 1999). Our modeling of search in this case study uses the different emphasis people might give to discriminability and validity as a theoretical bases for individual differences in search, and shows how this theory can be formalized within the hierarchical Bayesian approach.

2.1 Data

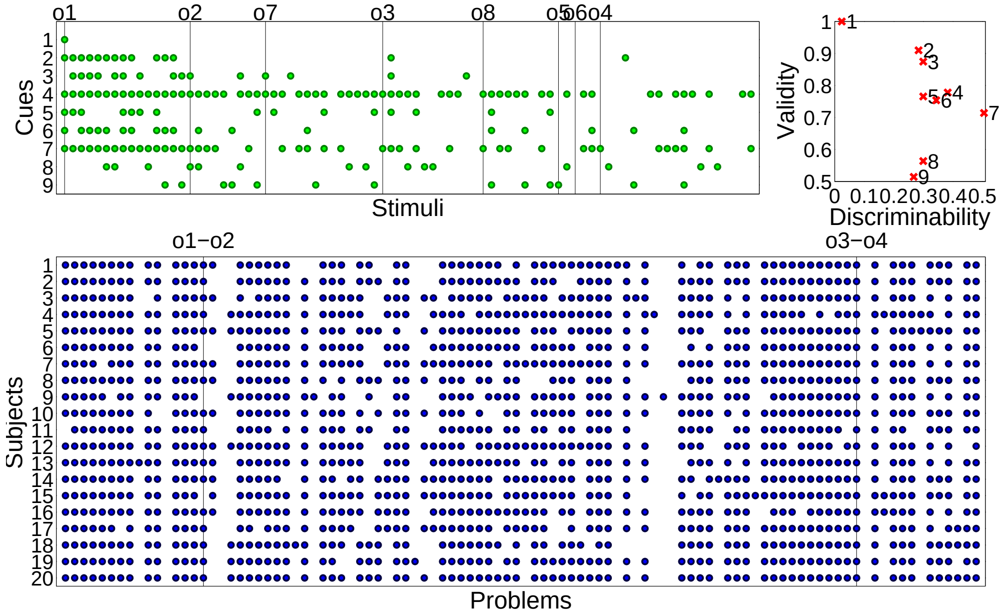
Figure 1 shows the experimental design and simulated data.The top left panel shows how the 83 city objects are defined in terms of the 9 cues. The objects are ordered (left to right) in columns from highest to lowest in terms of the population decision criterion. The cues correspond to rows, and the presence of a dot indicates that a cue belongs to an object. A few objects are depicted with vertical lines intersecting each cue that they possess (labeled o1 to o8); these are used in our subsequent worked examples to illustrate different patterns of search, stopping and choice. For example, the first object is highlighted as o1, and has cues 1, 2, 4, 5, 6 and 7. The top right panel shows how the 9 cues vary in terms of their discriminability (i.e., the proportion of stimulus pairs for which one has the cue and the other does not) and validity (i.e., the proportion of pairs for which the stimulus with the cue is higher on the criterion, given that the cue discriminates). Each of the 9 cues is represented by a red cross, showing the discriminability and validity of that cue. For example, the first cue (corresponding to "national capital" in the German cities dataset) has a very low discriminability (because for most city pairs, neither is the national capital), but a very high validity (because when one city is the national capital in a problem, one of those cities is Berlin, and it is always the largest city).

The bottom panel of Figure 1 shows the decisions made by 20 simulated subjects completing 100 two-alternative forced-choice problems. The problem set was chosen so that every object was included at least once, each problem pair was unique, and the cue validities and discriminabilities based on the problems were similar to the validities and discriminabilities obtained by considering all possible object pairs. The simulated data indicate when subjects chose the first of the presented objects. Again, some of these problem pairs are labeled, to help with later examples. For example, when objects o1 and o2 are presented as a pair, the blue dots in the highlighted column show which subjects chose o1 (i.e., all subjects except 6, 9 and 15).

These data were generated by simulating subjects who always applied a one-reason decision process, but used different search orders. That is, we used the take-the-best (TTB: Gigerenzer & Goldstein,1996) decision rule, which stops search as soon as one discriminating cue is

[2] Though see Brown and Tan (2011), Glöckner and Betsch (2008) and Juslin and Persson (1999) for alternative conceptions of the judgment process.

Figure 1: Stimuli defined in terms of cuxes (top left), with different cue discriminabilities and validities (top right). The bottom panel shows artificial decision-making data for 20 subjects on 100 problem pairs, indicating when the first stimulus in the pair was chosen. The highlighted objects o1 to o8 and the problems in which they are compared (e.g., o1–o8) are used to indicate individual differences in behavior. See main text for details.



found, but used orders other than the standard TTB one that strictly follows decreasing cue validity. To simulate the data, we assumed every subject used the same search order for all their problems, but different subjects used different orders.[3] We discuss exactly how these search orders were chosen once we have described the modeling results.

For now, some hint of the individual differences in the raw data can be gleaned from Figure 1. For example, in the highlighted problem that compares the objects labeled o1 and o2, subjects make different decisions. This could arise, for example, if some subjects (e.g., subject 1) were using a validity based order of search, and so used cue 1 to make a decision thereby choosing object o1 because only object o1 has cue 1 (see top left panel). In contrast, other subjects (e.g., subject 6) might incorporate discriminability into their search order, and so consider cue 3 before cue 1 and thus chose object o2 because only object o2 has cue 3. For other problems, however, like the highlighted o3–o4 comparison, there is consistency across all subjects with all choosing o3, presumably because the cues that o4 possesses are a subset of those possessed by o3. This subset relation can be seen clearly in the upper left panel of Figure 1, because o3 has cue 4 and cue 7 whereas o4 has only cue 7.

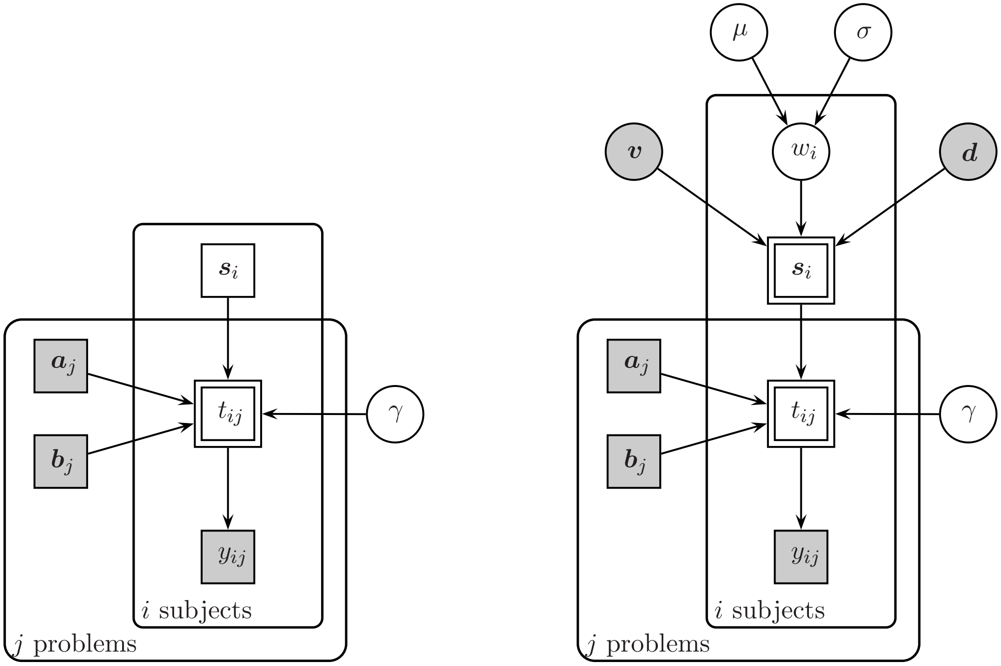
Thus, the modeling challenge is to take the information in Figure 1, and make inferences about the search orders individual subjects use, and how these search orders vary across the subjects.

2.2 Models

Figure 2 shows the two search models we apply to the decision data. On the left is a non-hierarchical model, and on the right is a hierarchically-extended model. Both models are shown using the formalism provided by graphical models, as widely used in statistics and computer science (e.g., Koller, Friedman, Getoor, & Taskar, 2007). A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data. We use the same notation as (Lee, 2008), with observed variables (i.e., data) shown as shaded nodes, and unobserved variables (i.e., model parameters to be inferred) shown as unshaded nodes. Discrete variables are indicated by square nodes, and continuous variables are indicated by circular nodes. Stochastic variables are indicated by single-bordered nodes, and deterministic variables (included for conceptual clarity) are indicated by double-bordered

[3] The current assumption that a single subject uses the same search order for all problems is a strong one, which could easily be relaxed in an extended hierarchical model. The basic idea would be for trial-by-trial variability in search orders for a subject, sampled from an overarching distribution on possible orders. This is an interesting direction for future work.

Figure 2: Graphical models for the simple search estimation model (left side), and the hierarchically extended search model (right side).



nodes. Encompassing plates are used to denote independent replications of the graph structure within the model. Further details and tutorials aimed at cognitive scientists are provided by (Lee, 2008) and Shiffrin, Lee, Kim, & Wagenmakers (2008). The advantage of graphical model implementation is that it automatically allows a fully Bayesian analysis for inferences relating the model to data, and can handle large and complicated joint distributions of parameters, as needed, for example, to examine combinatorial spaces of search orders. We achieve this using standard WinBUGS software (Spiegelhalter, Thomas, & Best, 2004), which applies Markov Chain Monte Carlo computational methods (see, for example, Chen, Shao, & Ibrahim, 2000; Gilks, Richardson, & Spiegelhalter, 1996; MacKay, 2003) to make inferences about model parameters and data. WinBUGS scripts, Matlab code, and all the relevant data for all of our models and analysis are provided as supplementary materials along with this paper on the page for this issue: http://journal.sjdm.org/vol6.8.html.

In the non-hierarchical model on the left of Figure 2, the decision made by the ith subject on the jth problem is $y_{ij} = 1$ if the first object (object “a”) is chosen, and $y_{ij} = 0$ if the second object (object “b”) is chosen. Because these data are observed, the node is shaded, and because they are discrete, it is square. The cues for the objects in the jth problem are given by the vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$, and are also known and discrete. If the ith subject searches the cues in an order given by the vector $\boldsymbol{s}_i$, then the TTB model will choose either $\boldsymbol{a}_j$ or $\boldsymbol{b}_j$, depending on which has the first discriminating cue in the search order $\boldsymbol{s}_i$, or will choose at random if there is no discriminating cue. This choice is represented by the node t_{ij}, which is double-bordered because it is a deterministic function, defined as

$$t_{ij} = \begin{cases} \gamma & \text{if } \mathrm{TTB}_{\boldsymbol{s}_i}\left(\boldsymbol{a}_j, \boldsymbol{b}_j\right) = \boldsymbol{a} \\ 1-\gamma & \text{if } \mathrm{TTB}_{\boldsymbol{s}_i}\left(\boldsymbol{a}_j, \boldsymbol{b}_j\right) = \boldsymbol{b} \\ 0.5 & \text{otherwise,} \end{cases}$$

where $\mathrm{TTB}_{\boldsymbol{s}_i}\left(\boldsymbol{a}_j, \boldsymbol{b}_j\right)$ is the TTB model, and γ is a decision parameter controlling the probability that the deterministic TTB choice is made. From this generative process, the decision data are distributed as $y_{ij} \sim \mathrm{Bernoulli}\left(t_{ij}\right)$. Using this model of the data, it is possible to infer a posterior distribution for the unknown search order for each subject. In other words, this model provides an ability to estimate search order from decision data, at the level of individual subjects.

The model on the right in Figure 2 shows how a hierarchical Bayesian approach can ask the deeper psychological question of *why* different people might have different search orders. In this model, search orders are generated by weighting information about both cue validity and discriminability. Formally, the i subject has a weight w_i that combines the validity v_k and discriminability d_k of the kth cue to give $w_i v_k + \left(1 - w_i\right) d_k$, and the order of these weighted sums gives the search order over the cues. Under this approach, purely validity-based search, as in TTB, corresponds to one extreme where $w_i = 1$, whereas purely discrimination-based research corresponds to the other extreme where $w_i = 0$. The model assumes that the weights follow a population-level normal distribution (i.e., different people have different weights, but there is a still a pattern at the population level), so that $w_i \sim \text{Gaussian}\,(\mu, \sigma)_{\mathcal{I}(0,1)}$, with weakly informative priors $\mu, \sigma \sim \text{Uniform}\,(0, 1)$.

We use a $\gamma \sim \text{Uniform}\,(0.5, 1)$ prior, reflecting the assumption that decisions will generally follow TTB. This corresponds to an assumption about the decisions that follow at the termination of search. In the literature, decision rules are perhaps less controversial, since most models simply state that one chooses the option pointed to by one (or more) of the cues. However, the extent to which such a rule should be applied deterministically or with some probability of error remains an area of contention. This is, for example, one of the issues in the debate between Brighton and Gigerenzer (2011) and Hilbig and Richter (2011). Here, we make a probabilistic assumption, consistent with the "accuracy of execution" formulation used by Rieskamp (2008).

The most interesting psychology in the modeling is that the hierarchical extension gives a theory of individual differences, as coming from different emphasis being placed on cue validity and discriminability in determining cue search order. It also naturally combines these individual differences with the idea of population-level structure, not letting the weights vary arbitrarily, but explaining them as coming from an overarching distribution. Thus, using the hierarchical model, the decision data can be used to infer group parameters like μ and σ, and individual parameters like the weights w_i and search orders $\boldsymbol{s}_i$.

2.3 Results

Figure 3 summarizes some of the main results of the modeling.[4] Each panel corresponds to a subject, and the true search order used to generate their data is shown at the top of the panel. The histograms show how close the inferred search orders are to this truth, using the standard Kendall Tau measure of distance between order (i.e., how many pair-wise swaps it takes to change the inferred order into the true order). The yellow (light) distribution corresponds to the non-hierarchical model, while the green (dark) distribution corresponds to the hierarchical model. Inset within each figure is the posterior over the w_i weight parameter for each subject, with its true data-generating value shown by the line.

Subjects 4 and 20 are good examples to focus on to discuss the general results. These two subjects give different emphases to discriminability and validity, reflected in their weights and search orders. Subject 4 tends to search cues that are higher in validity first (cues 2,3,1), according to the subject's higher value $w_i = 0.66$. In contrast, subject 20 searches the cue with the highest discrimination rate first (cue 7) since they place less emphasis on validity, as shown by the lower value $w_i = 0.48$.

The hierarchical model is able to infer these weights reasonably well (the distributions around the true value are relatively narrow), and produces excellent estimates of the search order (the greatest mass is on the true order). The modal inferred order is exactly correct, and the remainder of the inferences are within one or two swaps. It is clear for these subjects, and more generally across all subjects, that the hierarchical model inferences of search order are superior to those provided by the non-hierarchical model. This is because, in the hierarchical model, what is learned about one subject can be used to assist inferences about another, and is a good example of what is called "shrinkage" or "sharing strength" in statistics (e.g., Gelman, Carlin, Stern, & Rubin, 2004).

Not shown in Figure 3 are the inferences made about the other parameters. For both models, the expected posterior of the decision parameter $\gamma = .95$, and for the hierarchical model, expected posteriors for the overarching group distribution over the relative emphasis on validity and discrimination were $\mu_w = 0.54$ and $\sigma_w = .10$. These are all close to the true generating values of 0.5 and 0.1 for the mean and standard deviation, and 0.95 for the decision parameter. These results show how the models can make inferences about decision processes, and group-level properties of the subjects.

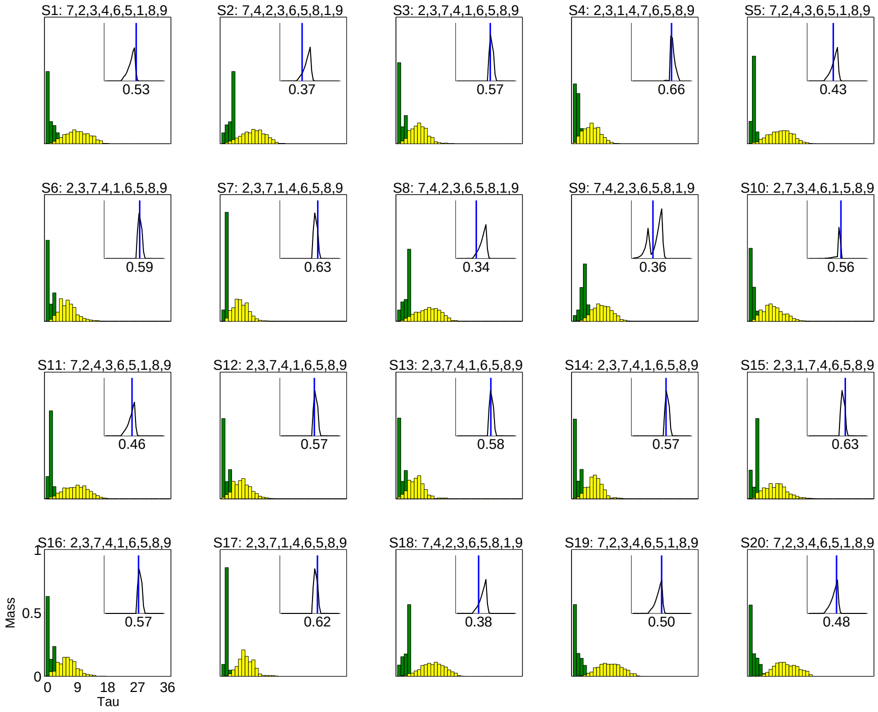
3 Modeling Stopping

In our first example all simulated subjects used a one-discriminating cue stopping rule. The accuracy and robustness of such rules has been discussed at great length in the context of heuristics like TTB and recognition (e.g., Czerlinski, Gigerenzer, & Goldstein, 1999; Gigerenzer & Goldstein,1996; Katsikopoulos, Schooler, & Hertwig, 2010). These simple rules are often contrasted against rules that require more information, including tallying and weighted-additive (WADD) rules, but do not necessarily improve the accuracy of predictions.

However, experiments that have investigated the stopping rules adopted by participants reveal mixed evidence: some participants use frugal rules often, some less so,

[4] All of our modeling inferences in this paper are based on running 2 chains of 10,000 samples, after a 100 sample burn-in (i.e., samples that are generated, but not used in inference). We checked convergence using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997) for every parameter, and always found it to be between 0.99 and 1.01, indicating convergence.

Figure 3: Performance of the two search models. Each panel corresponds to a subject, and their true cue search order is shown at the top. The histograms show the distribution of inferred search orders in terms of their tau distance from the true order. The green (dark) distribution is for the hierarchical model, and the yellow (light) distribution is for the non-hierarchical model. The inset shows the posterior distribution over the weight parameter in the hierarchical model, r



some never (e.g., Newell & Shanks, 2003). Environmental factors dictate their use, to some extent, such as the presence of costly cues increasing the use of frugal rules (Newell, Weston, & Shanks, 2003), but there are always individual differences across subjects in the same decision environment (e.g., Lee & Cummins, 2004; Newell & Lee, in press). Our second example examines how such patterns might arise.

3.1 Data

In our second example, we generated data by simulating subjects who always used the same search order (the $7, 2, 3, 4, 6, 5, 1, 8, 9$ order used by subject 20 in our search example), but used different stopping rules. In particular, we used stopping rules using minimal cue search or encouraging extensive cue search. We describe the exact nature of the simulation process once we have presented the modeling results.

For now, the data are shown in Figure 4. Again, some indication of the different stopping rules can be gleaned from the raw data. While there is consistency in problems like the highlighted o5–o6 comparison, which is not surprising given that object o6 has absent values for all cues (see top left panel of Figure 1), problems like o7–o8 show individual differences (6 subjects chose o7 with the remainder choosing o8). Looking at the cue structure of o7 and o8 (top left panel of Figure 1), it is clear that cue 7 provides some early evidence for o8 in the search order, but this evidence would later be compensated by the greater evidence cue 3 provides for o7 if a more conservative stopping rule was used to allow for longer search.

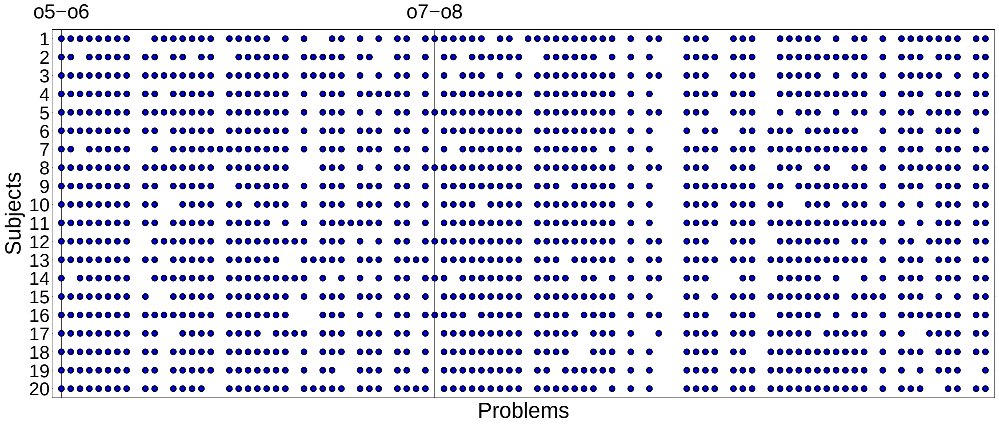
As for our search example, the modeling challenge is to take the information in Figure 4, and make inferences about the stopping rules individual subjects use, and how these rules vary across the subjects.

3.2 Models

Figure 5 shows the two models we applied to the stopping data. On the left is a simple mixture model that assumes every subject either uses a TTB strategy, in which a decision is made from the first discriminating cue, or a WADD strategy, in which all cues are used, and a decision is made based on the total evidence.[5] In this graphical model, the z_i parameter functions as an indicator vari-

[5] We measure evidence on the natural log-odds scale, so that the increment provided by a discriminating cue with validity v_k is $\log \frac{v_k}{1-v_k}$.

Figure 4: Artificial decision-making data for 20 subjects on 100 problem pairs, indicating when the first stimulus in the pair was chosen. The comparisons between objects o5– o6, and o7–o8 highlight individual differences, and are discussed in the main text.



able, with $z_i = 1$ if the ith subject uses TTB, and $z_i = 0$ if they use WADD. This indicator variable is distributed according to an (unknown) base-rate of TTB subjects in the population, so that $z_i \sim \text{Bernoulli}(\phi)$. The deterministic node θ_{ij} for the ith subject is then given by

$$\theta_{ij} = \begin{cases} \gamma & \text{if } \text{TTB}_{\boldsymbol{s}}(\boldsymbol{a}_j, \boldsymbol{b}_j) = \boldsymbol{a} \text{ and } z_i = 1 \\ 1-\gamma & \text{if } \text{TTB}_{\boldsymbol{s}}(\boldsymbol{a}_j, \boldsymbol{b}_j) = \boldsymbol{b} \text{ and } z_i = 1 \\ \gamma & \text{if } \text{WADD}_{\boldsymbol{s}}(\boldsymbol{a}_j, \boldsymbol{b}_j) = \boldsymbol{a} \text{ and } z_i = 0 \\ 1-\gamma & \text{if } \text{WADD}_{\boldsymbol{s}}(\boldsymbol{a}_j, \boldsymbol{b}_j) = \boldsymbol{b} \text{ and } z_i = 0 \\ 0.5 & \text{otherwise,} \end{cases}$$

with $y_{ij} \sim \text{Bernoulli}(\theta_{ij})$. We use a $\phi \sim \text{Uniform}(0.25, 0.75)$ prior, reflecting the assumption that we believe there are significant numbers of both TTB and WADD subjects, but we do not know the exact proportions.[6] We use a $\gamma \sim \text{Uniform}(0.5, 1)$ prior on the decision parameter, as before.

The hierarchically extended graphical model on the right of Figure 5 provides an account of why people have different stopping rules. It uses the idea that people search until they have some criterion level of evidence in favor of one stimulus over another, as per sequential sampling interpretations of decision-making in cue search tasks (Lee & Cummins, 2004; Newell, 2005). This theoretical conception is consistent with accounts of simple decision-making that emphasize desired levels of confidence as key mechanisms, as in the pioneering work of Vickers (1979) and later similar ideas in Hausmann and Lage (2008).

Formally, the criterion evidence level is e_i for the ith subject, and is assumed to come from one of two group-level Normal distributions with means and standard deviations $\mu_1, \mu_2, \sigma_1, \sigma_2$. The z_i indicator variable now controls which distribution the ith subject draws their evidence criterion value from, so that $e_i \sim \text{Gaussian}(\mu_{z_i}, \sigma_{z_i})$. We again place weakly informative priors on the means and standard deviations, and include an order constraint in the priors on the means, so that $\mu_1 \leq \mu_2$.

The deterministic decision now follows the sequential sampling model, so that

$$\theta_{ij} = \begin{cases} \gamma & \text{if } \text{SEQ}_{(\boldsymbol{s},e_i)}(\boldsymbol{a}_j, \boldsymbol{b}_j) = \boldsymbol{a} \\ 1-\gamma & \text{if } \text{SEQ}_{(\boldsymbol{s},e_i)}(\boldsymbol{a}_j, \boldsymbol{b}_j) = \boldsymbol{b}, \end{cases}$$

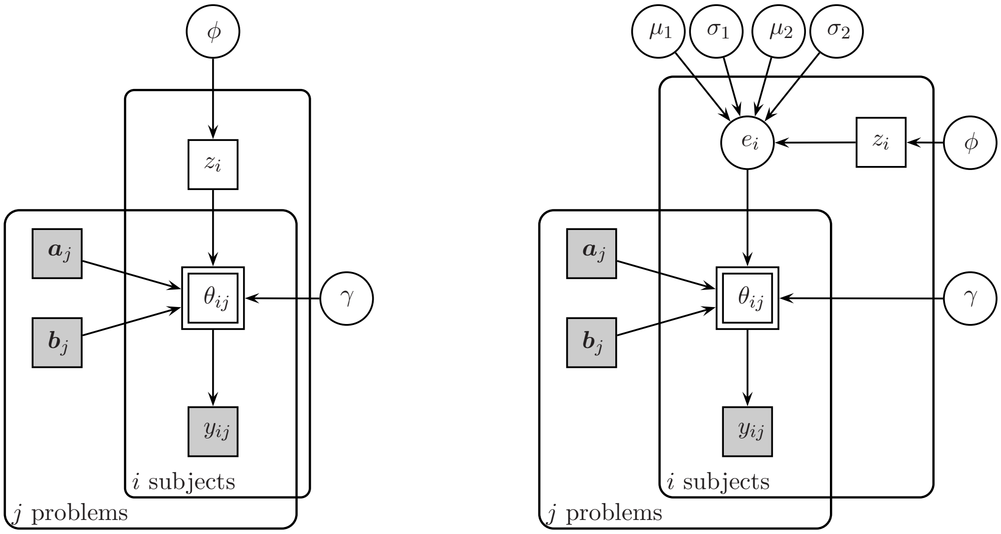
where $\text{SEQ}_{(\boldsymbol{s},e_i)}(\boldsymbol{a}_j, \boldsymbol{b}_j)$ gives the choice ($\boldsymbol{a}$ or $\boldsymbol{b}$) that the sequential sampling model makes using a search order $\boldsymbol{s}$ to criterion level of evidence e_i on stimuli $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$.

3.3 Results

Figure 6 summarizes the results of applying the two stopping models to the artificial data. The top row corresponds to the simple mixture model. The top-left panel shows the inferences about whether each subject used TTB or WADD, as given by the posterior of z_i. It is clear that 7 subjects were classified as using a TTB stopping rule, with the remaining 12 using WADD. The other panels in the top row show the posterior distributions over ϕ and γ, corresponding to inferences about the base-rate of TTB use, and the probability of following the deterministic TTB and WADD strategies in making decisions. These results show how a simple hierarchical

[6] We include this assumption not because it follows from any aspect of the current simulated data, but because it demonstrates how a reasonable substantive assumption can easily be incorporated in a Bayesian analysis.

Figure 5: Graphical models for the simple stopping estimation model (left side), and the hierarchically extended stopping model (right side).



Bayesian mixture model provides a complete and principled approach to identifying which subjects use different stopping rules, which is a common methodological challenge in the heuristic decision-making literature (Bröder & Schiffer, 2006; Newell & Lee, in press; Rieskamp & Otto, 2006).

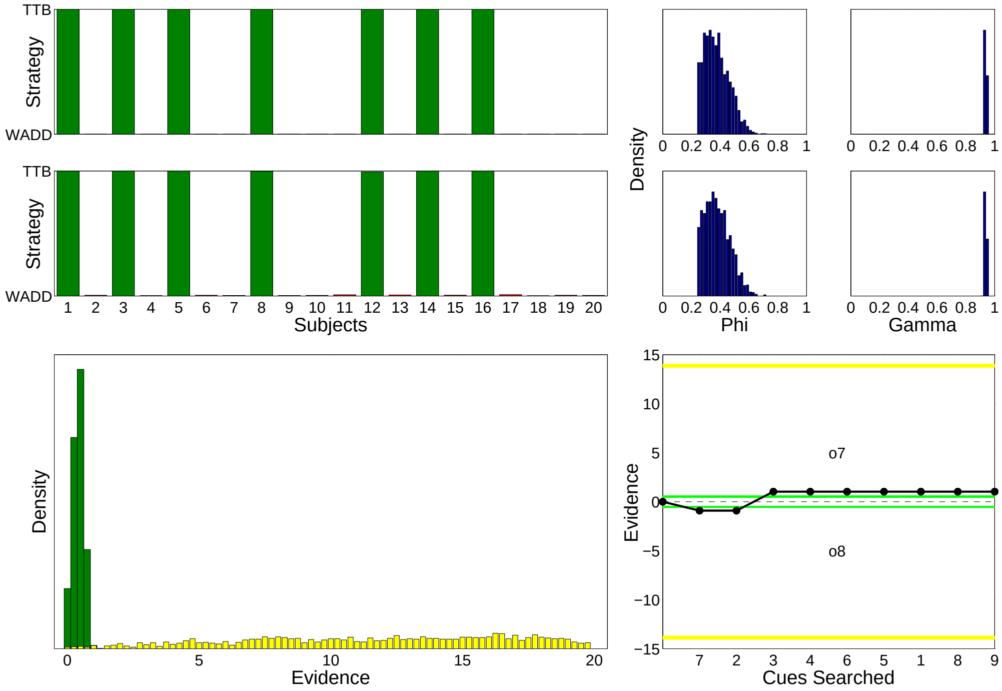
The middle row of Figure 6 presents the same analyses for the extended sequential sampling model based on evidence accumulation. The results are extremely similar, highlighting that the same information about stopping rule use can be extracted within the sequential sampling framework. The bottom row shows some of the additional psychological information gained by moving to this framework. The bottom-left panel shows the group-level distribution of evidence for the two distributions, with the green (dark) distribution corresponding to the low evidence (essentially, TTB) group, and the yellow (light) distribution corresponding to the high evidence (essentially, WADD group). These evidence values are then interpreted in sequential sampling model terms in the bottom right panel, showing the evidence bounds needed for decision-making on the o7–o8 problem's pattern of evidence accrual. The dotted black line sums the evidence provided by cues as search progresses. The green (dark) evidence threshold is relatively low, so a single discriminating cue provides sufficient evidence for the o8 decision. The yellow (light) evidence threshold is very high, so that all cues are searched, leading to o7 being chosen.

The results for the hierarchical model tell us something about different levels of evidence people may require, to explain their different stopping rules. But the results also tell us something about what we do not know, because of limitations in the environment used in the experiment. The actual distributions of evidence parameters we used to generate the data had means of 1 and 10 for the TTB and WADD groups respectively, with standard deviations of 0.2 and 3, and a decision parameter of 0.95. The inferences in Figure 6 are consistent with these generating values, but are not very precise. In particular, the diffuse distribution for the high evidence group in the lower-left panel of Figure 6 shows that, once a threshold level of about 2 is required, the sequential sampling stopping rule shown in the lower-right panel leads to all cues being examined, as per the WADD strategy. If people are using finer-grained intermediate evidence values, as they were in these simulated data, the environment used in the current experiment is not able to make this distinction. The fault here lies with the environment, rather than the inference method. The cue structure of the problems available in the German cities domain simply do not allow for diagnosis of some range of evidence values from decision-making behavior.One way to overcome this limitation would be to use environments with cue validities that allow for more compensatory decisions. In the current setting, the diffuse inference distributions are appropriate, showing both what is and is not known about underlying parameters from the available behavioral data.

4 Conclusion

We hope to have demonstrated that hierarchical Bayesian modeling can provide a principled way to understand and explain the diversity found in a standard judgment task. One of the most compelling features of the hierarchical

Figure 6: Performance of the two stopping models. The top row shows the inferences about TTB and WADD strategy use, and the base-rate (ϕ) and decision (γ) parameters for the mixture model. The middle row shows the same inferences for the hierarchically extended model. The bottom row shows the distribution of evidence values inferred by the hierarchical model (bottom left), and their interpretation as threshold levels of evidence within a sequential sampling of stopping for the problem o7–o8 (bottom right).



Bayesian approach is that it encourages deeper theorizing and the construction of more psychologically complete models Lee (2011), because the graphical modeling framework makes it easy to implement and evaluate new ideas. For example, it is natural to ask whether the extended model of search we presented, weighting both validity and discriminability, is more accurate than the original TTB validity-only approach. Both accounts are easy to implement as graphical models, and easy to compare directly, using the Bayesian model comparison approach described by Wagenmakers, Lodewyckx, Kuriyal, and Grasman (2010). As another example, it is straightforward to develop a model that incorporates both the searching and stopping processes, which we considered separately. This would constitute a more complete model of heuristic decision-making, and allow information about both searching and stopping operate, and how they interact, at both the individual and group level, to be inferred from behavioral data.

There is also, however, clearly still more that we need to understand. For example, while the models we have considered can explain why one might see individual differences in search and stopping rules (e.g., because individuals weight discrimination and validity differently), they cannot reveal how people arrive at those different search orders. In other words, while our hierarchical extensions involve theoretical accounts of searching and stopping, they are necessarily incomplete theories. A fuller account would presumably incorporate aspects of individual differences related to intelligence, personality, and so on, describing how they relate to differences in decision-making behavior (Bröder, 2003; Hilbig, 2008). More complete theories incorporating these factors could naturally be incorporated within the hierarchical Bayesian approach.

Whether one conceptualizes the tool(s) that people use for tasks of this kind as comprising numerous heuristics contained within a toolbox (e.g., Gigerenzer & Todd, 1999) or a single "tool" that can incorporate heuristics as special cases (Glöckner, Betsch, & Schindler, 2010; Lee & Cummins, 2004; Newell, 2005; Newell & Lee, in press), both need to provide accounts of how different heuristics are selected for different decision tasks (Gigerenzer & Gaissmaier, 2011; Rieskamp & Otto, 2006), or analogously, how and why new and successful parameter combinations are set for each type of problem (Marewski, 2010; Newell & Lee, 2011).

We are optimistic that using the hierarchical Bayesian methods demonstrated here will provide a window on this process and in so doing bring a new perspective to the debates between "toolbox" and "single-tool" interpretations of decision making (e.g., Gigerenzer & Gaissmaier, 2011; Glöckner et al., 2010; Hilbig, 2010; Newell, 2005; Newell & Lee, in press). More importantly we hope that other researchers see the potential for these methods to advance understanding across the wide range of higher-level cognitive phenomena that are relevant to judgment and decision making (e.g., Lee, 2008; Nilsson et al., 2011; van Ravenzwaaij et al., 2011).

References

Brighton, H., & Gigerenzer, G. (2011). Towards competitive instead of biased testing of heuristics: A reply to Hilbig & Richter (2011). *Topics in Cognitive Science*, 197–205.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 611–625.

Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 904–918.

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Brown, N., & Tan, S. (2011). Magnitude comparison revisited: An alternative approach to binary choice under uncertainty. *Psychonomic Bulletin & Review*, *18*, 392–398.

Chen, M. H., Shao, Q. M., & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Spinger-Verlag.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.) (pp. 97–118). Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis* (Second Ed). Boca Raton, FL: Chapman & Hall/CRC.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov Chain Monte Carlo in Practice*. Boca Raton (FL): Chapman & Hall/CRC.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1055–1075.

Glöckner, A., Betsch, T., & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making*, *23*, 439–462.

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.

Hausmann, D., & Läge, D. (2008). Sequential evidence accumulation in decision making: The individual desired level of confidence can explain the extent of information acquisition. *Judgment and Decision Making*, *3*, 229–243.

Hilbig, B. E. (2008). Individual differences in fast-and-frugal decision making: Neuroticism and the recognition heuristic. *Journal of Research in Personality*, *42*, 1641–1645.

Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review*, *17*, 923–930.

Hilbig, B. E., & Richter, T. (2011). Homo heuristicus outnumbered: Comment on Gigerenzer and Brighton (2009). *Topics in Cognitive Science*, *3*, 187–196.

Juslin, P., & Persson, M. (1999). Probabilities from exemplars (PROBEX): A lazy algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *95*, 1–45.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *XLVII*, 263–291.

Katsikopoulos, K. V., Schooler, L. J., & Hertwig, R. (2010). The robust beauty of ordinary information. *Psychological Review*, *117*, 1259–1266.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

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

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

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and "rational" models. *Psychonomic Bulletin & Review*, *11*, 343–352.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Marewski, J. N. (2010). On the theoretical precision and strategy selection problem of a single-strategy approach: A comment on Glöckner, Betsch, and Schindler (2010). *Journal of Behavioral Decision Making*, *23*, 463–467.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In G. Gigerenzer, P. M. Todd, & The ABC Research Group (Eds.), (pp. 119–140). Oxford University Press.

Newell, B. R. (2005). Re-visions of rationality. *Trends in Cognitive Sciences*, *9*, 11–15.

Newell, B. R., & Lee, M. D. (2011). The right tool for the job? comparing evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making*, *24*, 456–481.

Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies for decision making: The success of 'success'. *Journal of Behavioral Decision Making*, *17*, 117–130.

Newell, B. R., & Shanks, D. R. (2003). Take-the-best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 53–65.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes*, *91*, 82–96.

Nilsson, H., Rieskamp, J., & Wagenmakers, E. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology*, *55*, 84–93.

Rakow, T., Newell, B. R., Fayers, K., & Hersby, M. (2005). Evaluating three criteria for establishing cue-search hierarchies in inferential judgment. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *31*, 1088–1104.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1446–1465.

Rieskamp, J., & Otto, P. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.

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.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, *55*, 94–105.

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Wagenmakers, E., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey procedure. *Cognitive Psychology*, *60*, 158–189.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, *54*, 14–27.