COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 38 (2014) 1384–1405

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12119

A Hierarchical Bayesian Modeling Approach to Searching and Stopping in Multi-Attribute Judgment

Don van Ravenzwaaij,[a] Chris P. Moore,[b] Michael D. Lee,[c] Ben R. Newell[b]

[a]*School of Psychology, University of Newcastle*
[b]*School of Psychology, University of New South Wales*
[c]*Department of Cognitive Sciences, University of California Irvine*

Received 9 January 2013; received in revised form 20 May 2013; accepted 19 July 2013

Abstract

In most decision-making situations, there is a plethora of information potentially available to people. Deciding what information to gather and what to ignore is no small feat. How do decision makers determine in what sequence to collect information and when to stop? In two experiments, we administered a version of the German cities task developed by Gigerenzer and Goldstein (1996), in which participants had to decide which of two cities had the larger population. Decision makers were not provided with the names of the cities, but they were able to collect different kinds of cues for both response alternatives (e.g., “Does this city have a university?”) before making a decision. Our experiments differed in whether participants were free to determine the number of cues they examined. We demonstrate that a novel model, using hierarchical latent mixtures and Bayesian inference (Lee & Newell, 2011) provides a more complete description of the data from both experiments than simple conventional strategies, such as the take–the–best or the Weighted Additive heuristics.

Keywords: Heuristic decision making; Hierarchical Bayesian models; Probabilistic inference

In everyday life, we are constantly confronted with situations that require evaluations of different courses of actions. Examples include mundane tasks such as buying a ticket for a show in the local theater (do we go, or do we forsake this opportunity to keep saving for that new laptop?) but also more serious activities, such as buying a house or deciding to marry someone.

To deal with such diverse decision environments, people have to devise strategies with respect to the information, or *cues*, they want to gather: How many cues do they examine

Correspondence should be sent to Don van Ravenzwaaij, Department of Psychology, Aviation Building, University of Newcastle, room AVG11, Newcastle, NSW 2300, Australia. E–mail: don.vanravenzwaaij@newcastle.edu.au

and in what order? This interrogation of information can be internal (i.e., searching one's memory), external (i.e., searching the environment), or a combination of the two. Here, we focus on decisions made on the basis of cues in the external environment and examine how people deal with the inherent trade-off between exhaustive information search and the need to make decisions. For example, in deciding whether to buy a house, the following cues may be of relevance: "How affordable is this house?", "What is the area of the house in square meters?", "How many rooms does it have?", "How far is the commute?", and "Is the house situated in a good neighborhood?"

Evidently, this list of cues can be expanded to a very large number of factors that are of varying relevance to a person's decision about buying a house. Typically, decision makers cannot inspect all of the available cues. One reason is that there are obligations in life besides house buying. Another reason is that someone else might buy this house during the time it takes to examine all the cues. Thus, decision makers must choose what cues to examine, and when to stop examining cues.

A logical course of action might be to start with the cue considered most important, such as the price of the house. Depending on the outcome of this comparison, the decision maker may choose to examine additional cues (if both houses are about equally expensive) or make a decision (if one house is $50,000 cheaper than the alternative). As such, there are two factors that describe a person's cue search: the order in which cues are searched and the number of cues searched.

In experimental psychology, it is common practice to study the process of decision making using relatively simple tasks. More often than not, the task requires participants to make a decision repeatedly between one of two alternatives for which the decision problem is alike. A notable example of such a task is the German cities task (Gigerenzer & Goldstein, 1996; Goldstein & Gigerenzer, 2002) in which, over the course of consecutive trials, a participant has to decide which of two cities has the larger population. This decision can be made by extracting information on different cues in any order. An example of such a cue could be "Is this city the national capital?"

A number of different heuristics have been proposed that describe the different strategies a person might entertain when faced with a decision problem such as the German cities task. In this paper, we will examine two prominent heuristics that differ primarily in their level of frugality: the take–the–best strategy (TTB: Gigerenzer & Goldstein, 1996), which relies on a single discriminating cue, and the Weighted Additive strategy (WADD: Gigerenzer & Goldstein, 1996), which weights and adds the information conveyed by *every* relevant cue in the environment. Although the cities task was originally proposed to examine inferences from memory, numerous studies have used inferences from "givens" tasks, as we do, in which cue information is presented to participants on a computer screen. These studies have revealed evidence for both classes of strategy (e.g., Bergert & Nosofsky, 2007; Bröder, 2000; Bröder & Schiffer, 2003; Glöckner & Betsch, 2008; Newell & Shanks, 2003; Newell, Weston, & Shanks, 2003; Rieskamp & Otto, 2006; see Bröder, 2012 for a review).

In a related theoretical environment, Lee and Cummins (2004) and Newell (2005) have argued that the two extreme strategies, TTB and WADD, need not be viewed as discrete strategies that are adaptively relied upon in different situations but may rather be conceptualized as lying on a continuum of "evidence accumulation." This claim has been bolstered by recent empirical evidence (Newell & Lee, 2011) and by a hierarchical Bayesian latent mixture implementation of models that accumulate evidence (Lee & Newell, 2011). The purpose of the Lee and Newell (2011) paper was to demonstrate the viability of such models for simulated decision behavior in a German cities task environment (see also Scheibehenne, Rieskamp, & Wagenmakers, 2013). Here, we go beyond that work by applying the models to real data generated by participants in experimental environments that match those used by Lee and Newell (2011).

We continue to use the hierarchical Bayesian approach adoped by Lee and Newell (2011) to implement and evaluate our models of searching and stopping in decision making. The key attraction of this approach is that it naturally allows for the development of richly structured models, designed to account for multiple sorts of variability needed to explain human data (Lee, 2011; Lee & Wagenmakers, forthcoming). Hierarchical models are naturally able, for example, to account for individual differences. The introduction of overarching hierarchical distributions on parameters naturally captures continuous variations in parameters between people. The introduction of latent mixtures (i.e., each decision is modeled to come from one of a number of underlying decision processes) naturally captures individual differences in strategy use. In combination, the use of hierarchical latent mixture models allows for the development of models that capture structured variation in the order in which people search cues, the criteria they use to terminate search, and the interaction between these two sources of variation.

Bayesian inference provides a complete, principled, and interpretable framework for statistical inference in applying the resulting hierarchical models to behavioral data (e.g., Lee & Wagenmakers, 2005; Shiffrin, Lee, Kim, & Wagenmakers, 2008). Unlike standard model fitting methods like maximum likehood, it automatically provides information about the uncertainty associated with parameter estimates, and it allows for model selection that is sensitive to both the fit of a model and all of the aspects of the complexity of a model. Both parameter estimation and model selection are important for our hierarchical models. Inferring how much evidence a person needs before terminating search, for example, is naturally cast as a question of inferring a parameter value that controls an evidence threshold. Inferring whether a person uses one decision heuristic or another, however, is naturally cast as a model selection problem involving a latent mixture over those heuristics at the level of an individual.

Thus, our basic methodological approach is to build a structured hierarchical decision-making model that allows for the possibility of quantitatively and qualitatively different accounts of searching and stopping, and apply this model to all of the behavioral data produced by all of the people in an experiment. Bayesian inference then provides information about which components of the model are useful in accounting for the variability in behavior. If two candidate search processes are allowed for by a model, for example, but the inference is that people only use one of these processes, this can be considered a form of model selection in favor of the process people were inferred to use. The goal is not, therefore, to test two models against one another and infer which model is "true."

Rather, the goal is to identify what processes, and parameter values that control those processes, are needed to account for the observed range of human behavior.

In this way, our methodological approach complements that developed by Scheibehenne et al. (2013), who present impressive Bayesian comparisons of decision-making heuristics based on the Bayes Factor. This means that Scheibehenne et al. (2013) focus on choosing between competing heuristics as accounts of human decision making, whereas we seek to build overarching hierarchical models that include various heuristics as possible components. Scheibehenne et al. (2013) apply Bayesian model selection to choose between models, whereas we use Bayesian parameter estimation to make inferences about which components of a general hierarchical model are the important ones for describing human decision-making behavior.

In the next section, we describe the heuristics and the candidate models in greater detail. Then, we present two different experiments that showcase how these models might contribute to traditional analyses, which view all decisions as if they were generated using either TTB or WADD.

1. The heuristics

TTB is a frugal heuristic that prescribes an informativeness-based order for searching cues, and it terminates search as soon as a single discriminating cue is found, leading to the alternative favored by that cue (Gigerenzer & Goldstein, 1996). In the original conceptualization of TTB, the first step in deciding which of two cities had the higher population was to use recognition: If one city was recognized and the other was not, then choose the recognized one (Gigerenzer & Goldstein, 1996). It is only when both cities are recognized, thus disabling the use of "pure" recognition, the heuristic consults cues in memory that might be indicative of population size (e.g., "Is the city the national capital?"). In the version of the German cities task developed for the current experiments, we "disable" recognition by concealing the names of the cities and externalize information search to on-screen cues. This allows us to model the key aspects of the decision process that we are interested in: the order in which cues are searched, and the number of cues acquired for each decision.

We will discuss two variants of TTB: the original variant which is based on cue "validity" and a variant which is based on cue "discriminability." Both cue validity and cue discriminability can be relevant factors in determining the order and duration of cue search (e.g., Dieckmann & Todd, 2012; Newell, Rakow, & Weston, 2004; Rakow, Newell, Fayers, & Hersby, 2005), and there is evidence that both individual differences and the nature of the task might influence to what extent each of these factors is used (e.g., Martignon & Hoffrage, 1999).

The validity of a cue quantifies the number of times a cue points to the right answer as a ratio of the times it discriminates between the two response options. For example, in the German cities environment, the cue "Is the city the national capital?" has a validity of 1, because whenever one alternative scores positive on this cue, that will be because that alternative is the city Berlin and Berlin is the largest German city.

As an example of the original validity-based TTB, which we denote TTBv,[1] consider one trial of the version of the German cities task given to participants in our experiments: "Which German city has a higher population, A or B?" The decision maker investigates the most valid cue (cue validities were provided on screen, see method section and Fig. 3), "Is the city the national capital?", first. The answer is negative for both cities, so no decision can be based on that information. The decision maker continues with next most valid cues "Was the city once an exposition site?", for which the answer is "no" in both cases, then to "Does the city have a soccer team in the major national league?", for which the answer is again "no" in both cases, and so the decision maker must continue searching. The next cue searched, "Is the abbreviation of the city on license plates only one letter long?" differentiates between both cities: B has one letter on the license plate, whereas A does not. As such, the decision maker elects to answer B on this question, which is a correct decision. It is clear how this heuristic is frugal, as the most valid cue can never be overruled, even if all remaining unsearched cues were to point to the other response alternative.

The discriminability of a cue quantifies the number of times a cue discriminates between two response alternatives as a ratio of all possible cue comparisons for each question. The national capital cue does not discriminate very often and therefore has a discriminability close to zero, because this cue will only discriminate when one of the two response alternatives is Berlin.

The discriminability-based version of TTB, or TTBd, works in a similar way to TTBv, except that the decision maker starts with the cue that has the highest discrimination rate ("Is the city home to a university?"). When this cue does not discriminate, the decision maker examines the next highest discriminating cue and so on until a cue is examined that does discriminate between the two cities.

The alternative WADD heuristic assumes that participants examine all cues, collect all evidence that points to one of the two alternatives, and weigh them according to the importance of the particular cue. As an example, consider cities A and B for which there are two discriminating cues: "Is the city a state capital?" (A only) and "Is the abbreviation of the city on license plates only one letter long?" (B only). The participant weighs all evidence, concludes that the license plate cue provides more evidence (because it has the higher validity) and responds B. WADD is often considered to be rational and optimal as opposed to (versions of) TTB, which are considered frugal (Gigerenzer & Goldstein, 1996).

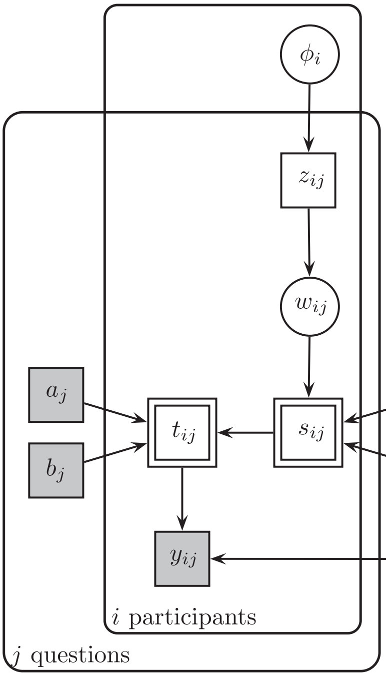
2. The Lee and Newell (2011) models

In the two models presented below, decision makers are not classified as either TTB or WADD decision makers. Rather, decision makers use both cue validity and cue discriminability and differ in terms of the weight they place on either of these, and also on how much evidence they accumulate before terminating search. By using a hierarchical Bayesian framework, we can extract information about underlying psychological processes at the level of both individuals and groups from the German cities data, and so address the deeper psychological questions about why there is variation in searching and stopping.

2.1. The Search Only model

The first model—which we call the Search Only model—is a simplification of the decision-making process. It assumes that the decision maker chooses an alternative as soon as he or she encounters the first cue that discriminates. This allows us to focus on modeling the order of search, without modeling search termination (see also Lee & Newell, 2011). The model assumes that the order in which each decision maker searches through the cues is based solely on cue validity, as in TTBv, or on some combination of cue validity and discriminability.

Fig. 1 shows the Search Only model as a *graphical model*. A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data (Koller, Friedman, Getoor, & Taskar, 2007). 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



$$\phi_i \sim \text{Beta}(1,1)$$

$$z_{ij} \sim \text{Bernoulli}(\phi_i)$$

$$\begin{cases} w_{ij} \leftarrow 1 & \text{if } z_{ij}=1 \\ w_{ij} \sim \text{Gaussian}(0.5, 10000)_{\mathcal{I}(0,1)} & \text{if } z_{ij}=0 \end{cases}$$

$$s_{ij} \leftarrow \text{Rank}\big(w_i v + (1-w_i)d\big)$$

$$t_{ij} \leftarrow \text{TTB}_{s_{ij}}(a_j, b_j)$$

$$\gamma \leftarrow 0.9999$$

$$y_{ij} \sim \begin{cases} \text{Bernoulli}(\gamma) & \text{if } t_{ij}=a \\ \text{Bernoulli}(1-\gamma) & \text{if } t_{ij}=b \\ \text{Bernoulli}(0.5) & \text{otherwise} \end{cases}$$

Fig. 1. Graphical model of the Search Only model.

deterministic variables (included for conceptual clarity) are indicated by double-bordered nodes. Encompassing plates are used to denote independent replications of the graph structure within the Search Only model.

The advantage of graphical model implementation is that it automatically allows a fully Bayesian analysis for inferences relating the model to data, and it can handle large and complicated joint distributions of parameters, as needed, for example, to examine combinatorial spaces of search orders. Further details and tutorials aimed at cognitive scientists are provided by Lee (2008), Shiffrin et al. (2008), and Lee and Wagenmakers, forthcoming.

In the Search Only model, the *i*th participant has an individual cue search order s_{ij} for each trial j. This search order s_{ij} is determined by a weighting w_{ij} of cue validity v and cue discriminability d for each cue. When $w_{ij} = 1$, search order is based exclusively on cue validity v, and when $w_{ij} = 0$, search order is based exclusively on cue discriminability d. As an example of an intermediate value, when $w_{ij} = 0.8$, cue validity v is weighted four times as strongly as cue discriminability in the determination of the search order.

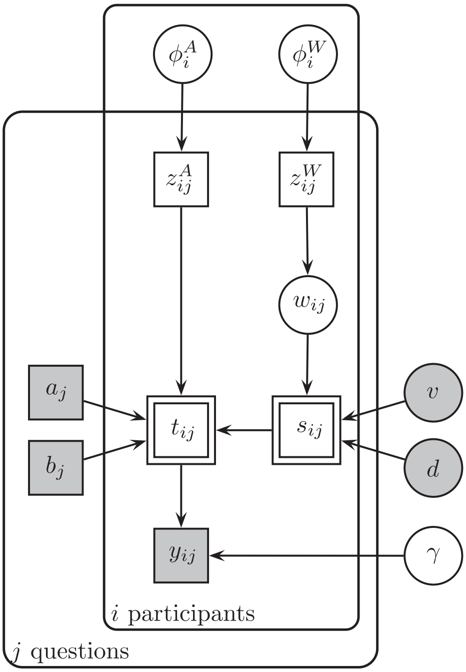
The prior on the weight parameter is determined by parameter z_{ij}. If $z_{ij} = 1$, the weight parameter w_{ij} is fixed on 1 to reflect a search order that is consistent with TTBv. Alternatively, if $z_{ij} = 0$, the weigh parameter w_{ij} is drawn from an approximately uniform distribution ranging from 0 to 1.

Values of z_{ij} are drawn from a Bernouilli distribution with a ϕ_i parameter for each participant. The search stops as soon as a discriminating cue is found, at which time the optimal response t_{ij}=1 if the cue points at city a_j and the optimal response $t_{ij} = 0$ if the cue points at city b_j. Accuracy of execution parameter γ determines the probability that the final response y_{ij} is consistent with the optimal response given the search order.

The Search Only model receives as input the choice response data and uses that to estimate a ϕ_i parameter for each participant and a z_{ij} parameter and weight parameter w_{ij} for each participant and each trial. In the original model by Lee and Newell (2011), γ was freely estimated, but for the experiments presented below, we fixed $\gamma = 0.9999$ because we wanted to focus on the extended models in as close as possible to deterministic form, since the original motivating models are deterministic.[2] The model specifications further deviate from the original model, in that a weight parameter is now estimated for each participant and for each trial, and by the fact that we include an estimation of the proportion of decisions that are in accordance with the TTB decision strategy in the Search Only model.

2.2. The Search and Stop model

The second model—which we call the Search and Stop model—represents the more realistic decision environment in which participants determine both their search order and the number of cues they examine before making a decision. The Search and Stop model is shown in Fig. 2, and it extends the Search Only model by including a search termination process.



$$\phi_i^A \sim \text{Beta}(1,1)$$

$$z_{ij}^A \sim \text{Bernoulli}(\phi_i^A)$$

$$\begin{cases} w_{ij} \leftarrow 1 & \text{if } z_{ij}^W = 1 \\ w_{ij} \sim \text{Gaussian}(0.5, 10000)_{\mathcal{I}(0,1)} & \text{if } z_{ij}^W = 0 \end{cases}$$

$$\phi_i^W \sim \text{Beta}(1,1)$$

$$z_{ij}^W \sim \text{Bernoulli}(\phi_i^W)$$

$$s_{ij} \leftarrow \text{Rank}(w_i v + (1 - w_i) d)$$

$$t_{ij} \leftarrow \begin{cases} \text{TTB}_{s_{ij}}(a_j, b_j) & \text{if } z_{ij}^A = 1 \\ \text{WADD}_{s_{ij}}(a_j, b_j) & \text{if } z_{ij}^A = 0 \end{cases}$$

$$\gamma \leftarrow 0.9999$$

$$y_{ij} \sim \begin{cases} \text{Bernoulli}(\gamma) & \text{if } t_{ij} = a \\ \text{Bernoulli}(1-\gamma) & \text{if } t_{ij} = b \\ \text{Bernoulli}(0.5) & \text{otherwise} \end{cases}$$

Fig. 2. Graphical model of the Search and Stop model.

In the Search and Stop model, the ith participant has an individual cue search order s_{ij} for each trial j. This search order s_{ij} is determined by a weighting w_{ij} of cue validity v and cue discriminability d for each cue, analogous to the Search Only model.

The prior on the weight parameter is determined by parameter z_{ij}^W. If $z_{ij}^W = 1$, the weight parameter w_{ij} is fixed on 1 to reflect a search order that is consistent with TTBv. Alternatively, if $z_{ij}^W = 0$, the weigh parameter w_{ij} is drawn from an approximately uniform distribution ranging from 0 to 1.

Values of z_{ij}^W are drawn from a Bernouilli distribution with a ϕ_i^W parameter for each participant. Participants either terminate their search after encountering the first discriminating cue if $z_{ij}^A = 1$, or examine all cues if $z_{ij}^A = 0$. Values of z_{ij}^A are drawn from a Bernouilli distribution with a ϕ_i^A parameter for each participant. When the participant stops searching, the optimal response $t_{ij} = 1$ if the cue points at city a_j and the optimal response $t_{ij} = 0$ if the cue points at city b_j. The accuracy of execution parameter γ determines the probability that the final response y_{ij} is consistent with the response determined by the search order.

The Search and Stop model incorporates TTBv as a special case: The combination of $z_{ij}^W = 1$ and $z_{ij}^A = 1$ means participants prioritize search order based on cue validity $z_{ij}^W = 1$ and stop once they encounter the first discriminating cue $z_{ij}^A = 1$. The Search and Stop model also incorporates WADD as a special case: $z_{ij}^A = 0$ means participants search all cues.[3] The combination of $z_{ij}^W = 0$ and $z_{ij}^A = 1$ corresponds to an expansion on existing heuristics incorporated in the model in which participants weigh the importance of cue validities and discriminabilities and make a decision as soon as they encounter a discriminating cue. In this way, the overall model tests for the usefulness of three different search accounts in explaining the behavior of participants.[4]

The Search and Stop model receives as input the choice response data and uses that to estimate a ϕ_i^W parameter and a ϕ_i^A parameter for each participant, and a z_{ij}^W parameter, a z_{ij}^A parameter, and a weight parameter w_{ij} for each participant and each trial.

3. Overview of experiments

With both the Search Only and the Search and Stop models introduced, we now introduce two experiments, each of which was aimed at testing one of the models empirically. Our goal was to determine what the models add in both explanatory power and psychological interpretation to fast and frugal heuristics, such as exclusive TTBv, TTBd, and more comprehensive strategies such as WADD. In Experiment 1, participants were free to choose their search order, but they were forced to stop after encountering the first discriminating cue. In Experiment 2, participants were free to choose their search order and were free to stop searching after examining at least one discriminating cue.

3.1. General participants

All participants were first-year undergraduate students at the University of New South Wales who participated in return for course credit. Twenty participants (Experiment 1: age 18–25, Experiment 2: age 17–26) took part in each experiment.

3.2. General material

A screenshot of a typical trial of each experiment is shown in Fig. 3. For each of the experiments, each trial presented the following question to the participant: "Which German city has a higher population?" The participant could choose between A and B, both of which represented German cities. As noted in the introduction, our key focus is on search and stopping behavior in external environmental cues and thus we effectively disabled internal memory–based search by concealing the names of each city. Thus, all decisions had to be made on the basis of the nine cues shown in Table 1. The cues were accessible via an on–screen menu throughout the experiment, and the order in which they appeared on screen was randomized across subjects.

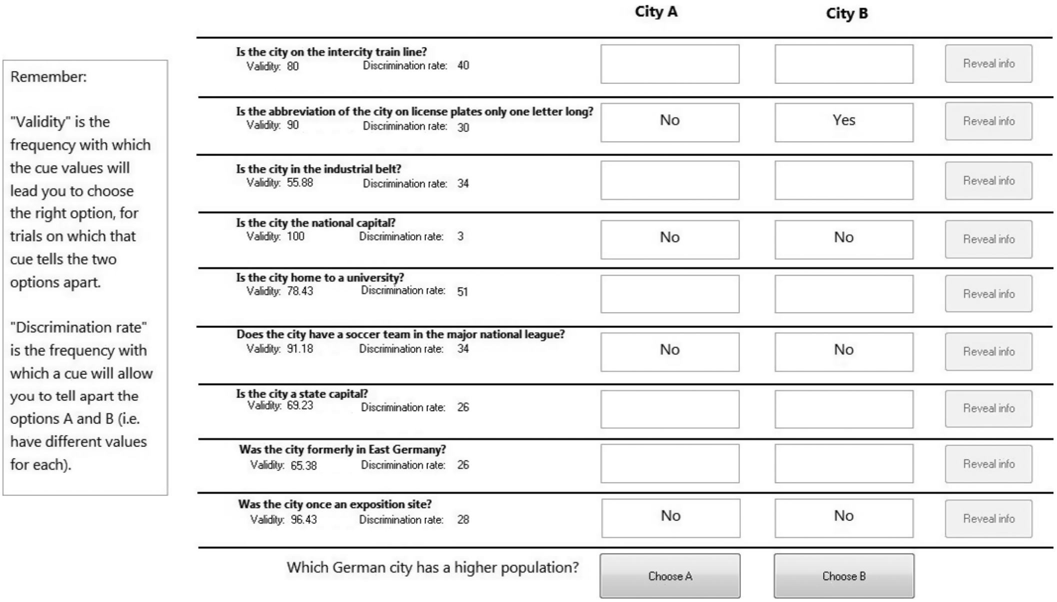
Participants examined cues by clicking on "reveal information" buttons. The order in which buttons were clicked was self-controlled. Deciding to stop examining additional cues was forced after encountering the first discriminating cue in Experiment 1, but self-controlled in Experiment 2. In all experiments, participants were provided with the cue validity and cue discriminability rates shown in Table 1. The problem set of 100 trials was the same as that used by Lee and Newell (2011) and was chosen so that every city

Table 1
The nine cues as used in the German cities environment from Gigerenzer and Goldstein (1996)

Nr.	Cue	Val	Dis
1	Is the city the national capital?	1	0.03
2	Was the city once an exposition site?	0.96	0.28
3	Does the city have a soccer team in the major national league?	0.91	0.34
4	Is the abbreviation of the city on license plates only one letter long?	0.90	0.30
5	Is the city on the intercity train line?	0.80	0.40
6	Is the city home to a university?	0.78	0.51
7	Is the city a state capital?	0.69	0.26
8	Was the city formerly in East Germany?	0.65	0.26
9	Is the city in the industrial belt?	0.56	0.34

Note. Val = Cue validity, Dis = Cue discriminability.

was included at least once and each problem pair was unique. All participants performed the same 100 trials in randomized order. Cue validity and cue discriminability rates were calculated for this subset of 100 comparisons, rather than for the whole set of possible comparisons (cf. Gigerenzer & Goldstein, 1996). The reason for this was to ensure that participants could relate the presented cue validity and cue discriminability rates as close as possible to their actual experience when performing the task. The presented information could be used by participants to base their search order on cue validity, cue discriminability, or a combination of the two. After each trial, participants received feedback with respect to the accuracy of their response. The experiment was self-paced.



Remember:

"Validity" is the frequency with which the cue values will lead you to choose the right option, for trials on which that cue tells the two options apart.

"Discrimination rate" is the frequency with which a cue will allow you to tell apart the options A and B (i.e. have different values for each).

	City A	City B
Is the city on the intercity train line? Validity: 80 Discrimination rate: 40		
Is the abbreviation of the city on license plates only one letter long? Validity: 90 Discrimination rate: 30	No	Yes
Is the city in the industrial belt? Validity: 55.88 Discrimination rate: 34		
Is the city the national capital? Validity: 100 Discrimination rate: 3	No	No
Is the city home to a university? Validity: 78.43 Discrimination rate: 51		
Does the city have a soccer team in the major national league? Validity: 91.18 Discrimination rate: 34	No	No
Is the city a state capital? Validity: 69.23 Discrimination rate: 26		
Was the city formerly in East Germany? Validity: 65.38 Discrimination rate: 26		
Was the city once an exposition site? Validity: 96.43 Discrimination rate: 28	No	No

Which German city has a higher population?

Fig. 3. Screenshot of a trial of the German cities task. See text for details.

3.3. General procedure

Participants were given instructions that they would repeatedly be making choices between pairs of two alternatives. The concepts cue validity and cue discriminability were explained. During the experiment, the following information remained on screen at all times: "Remember: 'Validity' is the frequency with which the cue values will lead you to choose the right option, for trials on which that cue tells the two options apart. 'Discrimination rate' is the frequency with which a cue will allow you to tell apart the options A and B (i.e., have different values for each)." No information was given about cue directionality (e.g., a participant was not informed if being home to a university was indicative of big cities or indicative of small cities). Participants completed a short training phase explaining the nature of the task and the meaning of the terms. They then completed two tasks in a random order, the German cities task and a different task, the results of which will not be reported here.

3.4. General analyses

All models were run in JAGS (Plummer, 2003).[5] JAGS uses Markov chain Monte Carlo (MCMC) to draw values from the posterior distribution. In our implementation, we sampled 2,000 values from the posterior; a first set of 1,500 values was discarded as burn-in. In total, we ran three different *chains*, which means we repeated this procedure three times with different starting values. We confirmed that all chains converged to the posterior distribution by means of the $\hat{R}$ statistic (Gelman & Rubin, 1992), which compares the variance over chains to that within chains.[6]

We used three criteria for model comparison. The first criterion, which we label "choice correspondence," is the correspondence between the actual answer given by each participant for each trial and the answers predicted by the respective model. A chance level of correspondence would be 50%, as there are two possible answers.

The second criterion, which we label "search correspondence," is the correspondence between the last cue examined by the participant and the last cue examined by the model. For Experiment 2, the chance level of correspondence is 11.1%, as there are nine different cues and participants are free to stop at any time. For Experiment 1, the last cue examined is equivalent to the first discriminating cue by default, because the design forces the participant to stop searching as soon as the first discriminating cue is encountered. Therefore, chance correspondence for search correspondence is higher for Experiment 1 at 45.4%.[7]

The final criterion, which we label "stop correspondence," is the correspondence between the number of cues examined by the participant and the number of cues the model predicts to be examined. A chance level of correspondence is 11.1%, as there are nine different cues. The stop correspondence analysis was only carried out for Experiment 2, as the number of cues searched was not under active control of the participant in Experiment 1.

There are a number of ways in which search and stop correspondence could be defined, each with advantages and disadvantages. Our choice, correspondence between the last cue and the total number of cues examined, has as advantage that it takes into account both the duration of the search and the critical cue that led to making the decision. The disadvantage of this criterion is that it fails to take into account the cues that were examined prior to the deciding cue.

Correspondences for each of these levels are calculated by comparing the actual data to the *posterior predictive distribution*. This is a standard Bayesian approach to evaluating the descriptive adequacy (or "fit") of a model. It measures the distribution of behavioral data the model expects based on its inferences, and it compares that distribution with the data actually observed. If the posterior predictive distribution closely resembles the empirical data, then the model fit is deemed adequate. For choice correspondence, we compared the actual choice response data with the posterior predicted distribution of decisions. For search and stop correspondence, we compared the last cue and number of cues that were examined in the actual search order data to the last cue and number of cues that would have been examined for the model generated search order s_{ij}. Averaging these values yields a proportion correspondence for each criterion of model comparison.

Correspondences for each of these levels are calculated by comparing the actual data to the *posterior predictive distribution*. This is a standard Bayesian approach to evaluating the descriptive adequacy (or "fit") of a model. It measures the distribution of behavioral data the model expects, based on the inferences about model parameters represented in the posterior distribution. Concretely, it averages the predictions the model makes across all possible model parameterizations, weighting each set of predictions by the probability of each parameterization in the posterior distribution. A posterior predictive evaluation of model can then evaluate the probability of the data actually observed. Unlike commonly used measures such as the Bayesian Information Criterior (BIC), which equate model complexity with the number of model parameters, the posterior predictive approach considers the full possible range of model parameters, and their interaction with each other, in assessing complexity. While posterior predictive analysis still has limitations, we think it is useful for evaluation of the current models. We also think, because our models are hierarchical, and have a mix of continuous and discrete parameters, that counting the number of parameters is a poor approximation to model complexity, and so the posterior predictive approach is more useful than the BIC for assessing fit.

For our analyses, the primary level of interest is choice correspondence, as the model only receives the choice response data as input. Search and stop correspondence give an indication of the extent to which the model is able to infer properties of the search process, despite not explicitly knowing the search order. All of the analyses reported were carried out on 93 questions, as the other 7 did not contain any discriminating cues.

4. Results for Experiment 1: Search Only

All participants successfully completed the task, and the accuracy across participants and trials was 68.0%. Accuracies for the responses generated by the Search Only model, TTBv, and TTBd are shown in the left column of Table 2. To examine whether some

Table 2
Response accuracies for every model for the two experiments

Model	Experiment 1	Experiment 2
SO	80.9	–
SAS	–	80.7
TTBv	79/93 = 84.9	79/93 = 84.9
TTBd	71/93 = 76.3	71/93 = 76.3
WADD	–	75/93 = 80.6

Note. SO = Search Only model, SaS = Search and Stop model.

participants might have been guessing, we used a latent mixture model that included a simple contaminant process, and we used Bayesian inference to assign each participant to the substantive or contaminant decision process (e.g., Zeigenfuse & Lee, 2010). The resulting Bayes factors (e.g., Hoijtink, Klugkist, & Boelen, 2008; Jeffreys, 1961; Kass & Raftery, 1995) suggest that participants 4 and 7 may have been guessing. More specifically, the relevant Bayes factors were 6.82 and 12.12, indicating that these participants were respectively 6.82 and 12.12 times more likely to have been guessing than to have been responding seriously. Removing these participants did not change the qualitative pattern of results, so we present the results with all participants included.[8]

For the choice correspondence analysis, the Search Only model does very well. It has a correspondence of 81.5% (where chance correspondence is 50%; see Table 3 for a summary of all analysis results). The deterministic TTBv and TTBd, respectively, have a correspondence of 75.3% and 72.6%. WADD reduces to TTBv when search is terminated at the first discriminating cue so the results for WADD are not displayed for this experiment.

Table 3
Percentage of correspondence for the three different types of analyses for the two experiments

Experiment	Model	Choice	Search	Stop
1	SO	**81.5**	55.1	–
	TTBv	75.3	53.2	–
	TTBd	72.6	44.0	–
	TTBvd	76.2	**55.6**	–
2	SaS	**89.5**	24.0	30.2
	TTBv	82.7	**27.8**	18.2
	TTBd	79.2	23.5	7.3
	WADD	83.0	10.9	**43.7**
	TTBvd/WADD	84.9	27.7	37.7

Note. The best-performing strategy is printed in bold. SO = Search Only model, SaS = Search and Stop model, Choice = Choice correspondence, Search = Search correspondence, Stop = Stop correspondence. Chance correspondence is 50% for choice correspondence for both experiments, 45.4% for search correspondence in Experiment 1, 11.1% for search correspondence in Experiment 2, and 11.1% for stop correspondence in Experiment 2.

As a stricter test of the Search Only model, we deterministically classified participants as either TTBv or TTBd depending on which pattern was most consistent with their response pattern. The resulting mix, which we label TTBvd, had a choice correspondence of 76.2%.

For the search correspondence analysis, the Search Only model does quite well. It has an agreement of 55.1% (where chance correspondence is 45.4%). TTBv, TTBd, and TTBvd have a correspondence of 53.2%, 44.0%, and 55.6%. The performance of the Search Only model in terms of search correspondence then is comparable to the deterministic mix of strategies and superior to that of each individual heuristic.

To establish whether the data warrant the added complexity presented by the Search Only model, Fig. 4 displays the proportion of data that was accounted for by TTBv ($z_{ij} = 1$) and the proportion of data accounted for by the Search Only model–specific part of weighing cue validity and discriminability for each of the participants ($z_{ij} = 0$). The important take-home message from this figure is that for every participant, some portion of the bar is colored white (representing the Search Only model–specific weighting) and some is colored black (representing TTBv). Thus, to a greater (e.g., participant 20) or lesser (e.g., participant 1) extent, the Search Only model–specific weighting adds significant explanatory power for every participant beyond that which TTBv provides. Note that the bars do not imply that search order was exactly consistent with TTBv or a weighing of

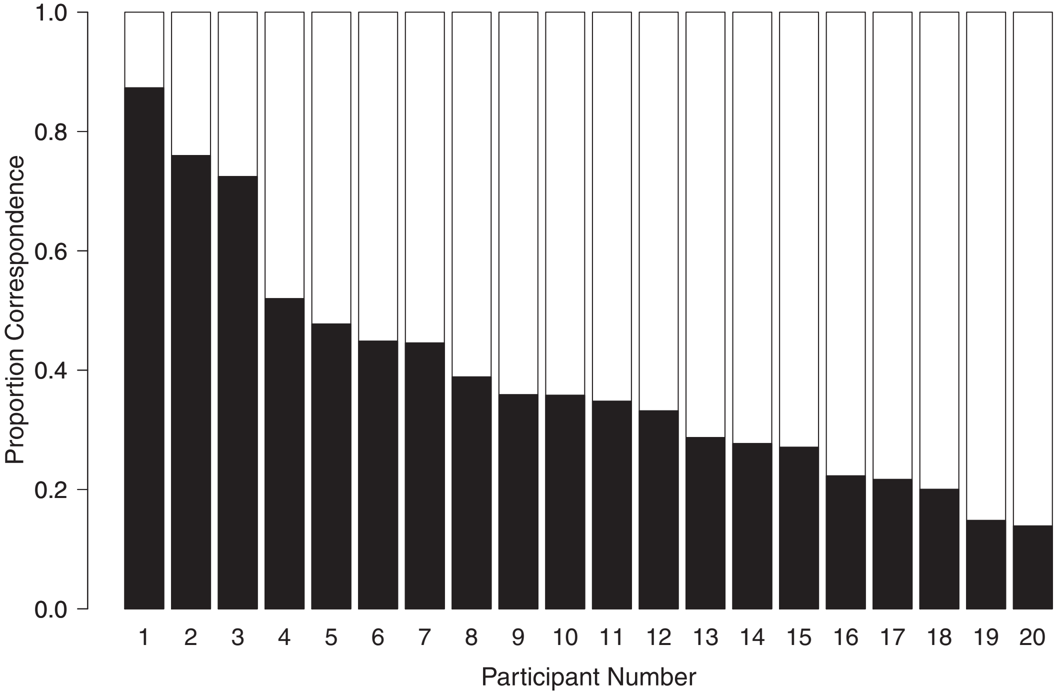


Fig. 4. Proportion of participant responses accounted for by TTBv (black) and the Search Only model–specific part that weighs cue validity and discriminability (white).

cue validity and discriminability. Rather, the bars indicate the Search Only model's best attempt at classifying the data. In other words, search orders inconsistent with either TTBv or a weighing of cue validity and discriminability are still classified as more likely to belong to either one or the other of the two processes.

In sum, it seems that the Search Only model is able to provide a more complete account of the data than the deterministic TTB heuristics. This is particularly true in terms of choice correspondence, but it is also moderately true in terms of search correspondence agreement.

5. Results for Experiment 2: Search and Stop

All participants successfully completed the task, and the accuracy across participants and trials was 75.4%. The accuracy in Experiment 2 was significantly higher than the accuracy in Experiment 1 ($t(38) = 2.38$, $BF_{01} = 0.41$), suggesting that having active control over when to stop searching leads to more accurate responding. Accuracies for the responses generated by the Search and Stop model, TTBv, TTBd, and WADD are shown in the right column of Table 2. We again undertook a contaminant latent mixture-based modeling analysis, and the Bayes factors suggest that no guessing took place.[9]

For the choice correspondence analysis, the Search and Stop model does very well. It has a correspondence of 89.5% (where chance correspondence is 50%). The deterministic TTBv, TTBd, and WADD have correspondences of 82.7%, 79.2%, and 83.0%, respectively.

As a stricter test for the Search and Stop model, we deterministically classified participants as either TTBv, TTBd, or WADD depending on which pattern was most consistent with their response pattern. The resulting mix, which we will label TTBvd/WADD, had a choice correspondence of 84.9%.

For the search correspondence analysis, the Search and Stop model has an agreement of 24.0% (where chance correspondence is 11.1%). TTBv, TTBd, WADD, and TTBvd/WADD have correspondences of 27.8%, 23.5%, 10.9%, and 27.7%, respectively, meaning the TTBv heuristic and the combination of heuristics perform better than the Search and Stop model based on search correspondence.

For the stop correspondence analysis, the Search and Stop model does moderately well. It has a stop correspondence of 30.2% (where chance correspondence is 11.1%). The deterministic TTBv and TTBd have agreements of 18.2% and 7.3%, making them substantially worse than the Search and Stop model. The deterministic WADD, which assumes that all cues are inspected at all times, has an agreement of 43.7%. Evidently, in almost half the instances, participants examined all of the cues. TTBvd/WADD has a stop correspondence of 37.7%

To establish whether the data warrant the added complexity presented by the Search and Stop model, Fig. 5 displays the proportion of data that was accounted for by TTBv ($z_{ij}^{W} = 1$ and $z_{ij}^{A} = 1$), the proportion of data accounted for by WADD ($z_{ij}^{A} = 0$), and the proportion of data accounted for by the Search and Stop model–specific part of weighing cue validity and discriminability for each of the participants ($z_{ij}^{W} = 0$ and $z_{ij}^{A} = 1$). Just

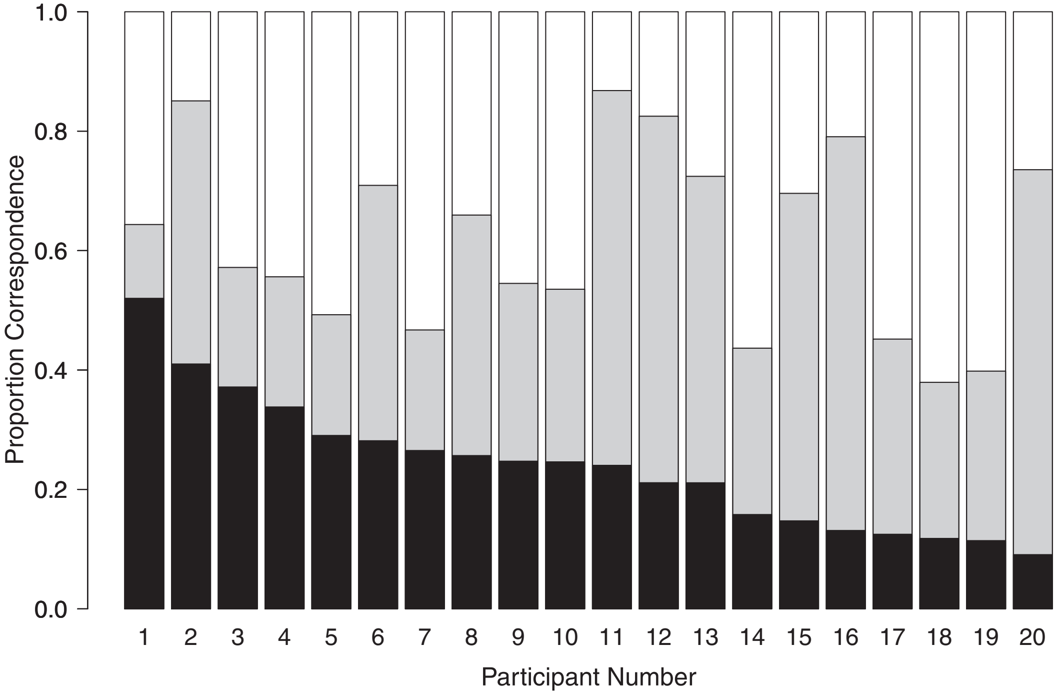


Fig. 5. Proportion of participant responses accounted for by TTBv (black), WADD (gray), and the Search and Stop model–specific part that weighs cue validity and discriminability (white).

like for Fig. 4, bars do not imply that search order was exactly consistent with TTBv, WADD, or a weighing of cue validity and discriminability. Fig. 5 shows that for all participants, the Search and Stop model adds significant explanatory power beyond that which both TTBv and WADD provide (i.e., for every participant some portion of the bar is colored white, representing the proportion of data accounted for best by the Search and Stop model–specific weighting).

In sum, it seems the Search and Stop model is able to do better than all strategies on choice correspondence. The Search and Stop model performs better than WADD, comparable to TTBd, and worse than TTBv or TTBvd/WADD on search correspondence. The Search and Stop model performs better than TTBv and TTBd, comparable to TTBvd/ WADD, and worse than WADD on stop correspondence. We address the reasons for the Search and Stop model's different patterns of performance across the three levels of analysis in the general discussion.

6. General discussion

When making a decision, people can collect information from multiple cues to help them formulate a response. Our results show that modeling people's search behavior is no trivial matter. In two different experiments, we demonstrated that describing decision makers' behavior with deterministic strategies like TTB or WADD does not do justice to the complexity of the decision-making process even in the relatively simple experimental task used here. Furthermore, we modeled the data at levels beyond choice response correspondence and examine the search process as well as the final choice.

In our first experiment, we forced participants to stop examining cues as soon as they encountered a cue that discriminated between the two response options. We showed that in terms of predicting a decision maker's response, the Search Only model, which assumes that people weigh cue validity and cue discriminability, rather than using one exclusively, is a more complete account of people's decisions than classifying them as exclusively using frugal TTB strategies. By modeling search order as a latent mixture that is either driven exclusively by cue validity or by a combination of cue validity and cue discriminability, we have in effect put a peak on a weight parameter of 1. The Search Only model takes the middle ground between assuming that all values of weighing cue validity and cue discriminability are equally likely and assuming only a weight of 1 (i.e., TTB). This is a novel and powerful perspective on participants' search strategy in multiattribute decision environments.

Our second experiment presented a more realistic scenario in which participants were free to examine cues in any order they preferred and to stop collecting evidence as soon as they felt confident enough to make a response. The Search and Stop model performed better than (combinations of) deterministic strategies in terms of predicting responses.

In terms of correctly predicting the terminating cue, the WADD strategy did no better than chance level, the Search and Stop model and a TTB strategy based on cue discriminability performed substantially better, and a TTB strategy based on cue validity performed best. The Search and Stop model could not outperform all heuristics, because it is unable to estimate precisely an evidence threshold for this decision environment. Instead, it categorizes participants in a one-reason group and a search-all-cues group. While still yielding the benefit the Search Only model provides as made clear in Experiment 1, the stop part of the Search and Stop model does not significantly add to model correspondence beyond what deterministic strategies can provide.

The Search and Stop model outperformed frugal strategies in terms of number of cues examined, but it performed somewhat worse than the Weighted Additive strategy. The agreement in number of cues searched between the data and the Weighted Additive strategy shows that participants inspected all cues on roughly half of the trials. The excellent performance of the Weighted Additive strategy seems to come from the nature of the task, in the sense that there was no explicit cost—other than time—associated with an exhaustive search.

A decision environment with more than nine cues could lead to a more sophisticated stopping criterion for the Search and Stop model, and so provide the possibility of more detailed model comparisons. More discriminating cues on each trial will increase the variation in the number of cues searched, both within and between participants. Such an environment would allow the Search and Stop model to estimate individual evidence thresholds, rather than classifying participants as either one-reason or search-everything.

This model would similarly benefit from the inclusion of time or reward penalties for each cue examined beyond the first (cf. Newell & Lee, 2011; Rakow et al., 2005; Rieskamp & Hoffrage, 2008). Such studies could also manipulate the presence or absence of information regarding cue validities and discrimination rates. In the current experiments, we chose to provide these values to participants because our goal was to model searching, stopping and choice behavior rather than the learning of cue-hierarchies. Other studies have examined the extent to which such hierarchies can be learned (Newell et al., 2004; Rakow et al., 2005) and demonstrated that participants can learn to search adaptively when given sufficient experience. A goal for the future is to extend the current model framework to incorporate such learning effects and examine how learning is affected by acquiring cue information via description (as was done in the current experiments) or trial-by-trial experience (cf. Rakow & Newell, 2010).

As noted in the introduction, our focus in this work has been on inferences from givens—examining search, stop, and decision behavior in environments where search is externalized and recognition is disabled. This focus opens us up to the criticism that our task is artificial and outside the scope of the fast and frugal heuristics program, which emphasizes memory-based inference (e.g., Gigerenzer & Goldstein, 1996). While we acknowledge that these two types of inference differ, we see no a priori reason why the same type of building blocks (i.e., searching, stopping, and decision rules) cannot apply in both situations. Many of the decisions we face in everyday life comprise both memory and external information search elements. Thus, advancing our understanding of both seems like a sensible goal. Our work follows a tradition of “preference–from–givens”-based investigations that have shed considerable light on the circumstances in which different decision strategies are adopted (e.g., Payne, Bettman, & Johnson, 1993).

For our experimental data, the increase in complexity presented by both models led to an increase in explanatory power beyond that of the fast and frugal heuristics, as emphasized in Figs. 4 and 5. It seems simple heuristics like TTB and WADD are *too* simple a tool to account for the more complex patterns of behavior exhibited by our participants. This general conclusion is in line with previous claims that more inclusive “single mechanism” models can provide better accounts of the full range of behavior observed in these types of experiments than those based on discrete, deterministic heuristics (e.g., Lee & Cummins, 2004; Lee & Newell, 2011; Newell, 2005; Newell & Lee, 2011).

Our analyses also highlight the potential of latent hierarchical mixture models, and Bayesian inference, as cognitive modeling approaches for investigations of decision making. As noted in the introduction, a key aim of the work reported here was to model real decisions and thereby go beyond the “proof of concept” provided in the parameter–recovery study of Lee and Newell (2011). Unsurprisingly, the human data were more complicated than those produced by the simulation used in Lee and Newell (2011). Nevertheless, the modeling and inference methods were able to address the complexity in the data and shed light on the underlying processes.

It is important to emphasize that the relative success of these models at each level of analysis—choice, search, and stopping—was based solely on providing the models with the choice response data as input. Extensions of this work could attempt to model the search and stopping process in situations where the Search and Stop model has access to search order data in addition to choice response data. The attraction of the current modeling approach is that such a model could be developed using the same basic components, framework, and assumptions used in our existing models.

Indeed, recent related work has shown the potential for hierarchical Bayesian models to address other key questions in this area such as the falsifiability of theories, which advocate a multiple strategy or "toolbox" view of adaptive decision making (Scheibehenne et al., 2013). The work of Scheibehenne et al. (2013) is largely complementary with the analyses presented here. Where Scheibehenne et al. (2013) focus mostly on model selection, the emphasis of our work is mostly on model development and estimation. Our work shows that the toolbox notion and the single-model notion need not be mutually exclusive, as the TTB heuristic, the WADD strategy, and a weighing of cue validity and discriminability can all be viewed as special cases of a more general model (e.g., Lee & Cummins, 2004; Lee & Newell, 2011; Marewski & Mehlhorn, 2011).

Our models represent a refinement of the fast and frugal heuristics aimed at explaining the processes underlying decision makers' search order. The parameter estimates allow one to assess to what extent the decision strategy of participants follows a TTB heuristic and to what extent something more complicated is going on (e.g., a weighting of cue validity and discriminability). For the latter, the parameter estimates also give us a magnitude of this weighting. This paper has focused on examining how such encompassing models can offer a better account of the performance of participants than a simple heuristic. The current models allow researchers to set up experiments aimed at testing specific hypotheses. For instance, one could manipulate cue validities or discriminabilities and investigate to what extent this would affect the weighting of validity and discriminability across conditions.

A logical next step is to arrive at a theoretical account of why decision makers terminate cue search. Rather than just modeling different stopping rules, we need a model that addresses why decision makers stop searching when they do. Another challenge for the future is to take into account learning across trials. How do people adapt their search order trial to trial, what motivates them to change orders, and why are there individual differences? Clearly, there is still a lot of work to be done. Nevertheless, the current work demonstrates that hierarchical Bayesian modeling of decision making may lead to a deeper understanding of the processes underlying decision making than describing behavior in terms of simple heuristics.

Notes

1. We deviate from the original TTB notation to highlight the difference with the discriminability-based version of TTB discussed later.
2. It would be interesting in future work to contrast the deterministic version with the version that uses the accuracy of executions parameter motivated and used by Lee

and Newell (2011). We found that the deterministic approach provided a better account of the current data, and this difference may have theoretical implications.

3. Note that in WADD, search order is irrelevant.
4. There are a number of ways the three alternatives could be parameterized. It would certainly be reasonable to use a ternary Categorical (also called Discrete) distribution, for example, to select between the search models. The parameterization we use was chosen because we found it a natural one in which to express reasonable priors and draw inferences. For example, separating z_{ij}^{W} from z_{ij}^{A} separates inferences about the use of TTB versus WADD from inferences about the use of validity versus other search orders within TTB.
5. Software freely available at http://mcmc-jags.sourceforge.net/.
6. As a rule of thumb, an $\hat{\mathrm{R}}$ higher than 1.05 is considered suspicious. For Experiment 1, 3,643 out of 3,656 sampled parameters had an $\hat{\mathrm{R}}$ smaller than 1.05. For Experiment 2, 5,450 out of 5,484 sampled parameters had an $\hat{\mathrm{R}}$ smaller than 1.05.
7. This value was obtained by averaging the ratios of one over the number of discriminating cues for each trial.
8. Results with participants 4 and 7 removed. For the Search Only model: Choice correspondence = 84.6%, and search correspondence = 56.2%. For TTBv: Choice correspondence = 78.6%, and search correspondence = 54.2%. For TTBd: Choice correspondence = 75.7, and search correspondence = 44.5%. For TTBvd: Choice correspondence = 79.5%, and search correspondence = 56.6%.
9. Only two participants were somewhat ambiguous: BF_{01} = 0.62 and BF_{01} = 0.22, indicating that even for the Bayes factor of 0.62 responding seriously is approximately one and a half times more likely than guessing.

References

Bergert, F. B., & Nosofsky, R. M. (2007). A response time approach to comparing generalized rational and take the best models of decision making. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*, 107–129.

Bröder, A. (2000). Assessing the empirical validity of the "Take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1332–1346.

Bröder, A. (2012). The quest for Take–the–Best: Insights and outlooks from experimental research. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Ecological rationality* (pp. 216–240). New York: Oxford University Press.

Bröder, A., & Schiffer, S. (2003). Take the Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General*, *132*, 277–293.

Dieckmann, A., & Todd, P. M. (2012). Simple rules for ordering cues in one–reason decision making. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Ecological rationality* (pp. 274–306). New York: Oxford University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–472.

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

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

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses that are of practical value for social scientists*. New York: Springer.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

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* (pp. 13–55). Cambridge, MA: MIT Press.

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 the "rational" models. *Psychonomic Bulletin & Review*, *11*, 343–352.

Lee, M. D., & Newell, B. R. (2011). Using hierarchical Bayesian methods to examine the tools of decision making. *Judgment and Decision Making*, *6*, 832–842.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lee, M. D., & Wagenmakers, E.-J. (forthcoming). Bayesian cognitive modeling: A practical course. Cambridge, UK: Cambridge University Press. Available from http://faculty.sites.uci.edu/mdlee/bgm

Marewski, J. N., & Mehlhorn, K. (2011). Using the ACT–R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, *6*, 439–519.

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.), *Simple heuristics that make us smart* (pp. 119–140). New York: 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 an evidence accumulation and a naive strategy selection model of decision making. *Journal of Behavioral Decision Making*, *24*, 456–481.

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

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

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.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, UK: Cambridge University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, A. Zeileis (Eds.) *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (pp. 20–22). Vienna: Technische Universitaet Wien.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience–based choice. *Journal of Behavioral Decision Making*, *23*, 1–14.

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., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*, 258–276.

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

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review*, *120*, 39–64.

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.

Zeigenfuse, M. D., & Lee, M. D. (2010). A general latent assignment approach for modeling psychological contaminants. *Journal of Mathematical Psychology*, *54*, 352–362.