*Journal of Behavioral Decision Making*
*J. Behav. Dec. Making*, **24**: 456–481 (2011)
Published online 23 May 2010 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/bdm.703

# The Right Tool for the Job? Comparing an Evidence Accumulation and a Naive Strategy Selection Model of Decision Making

BEN R. NEWELL[1]* and MICHAEL D. LEE[2]
[1]*School of Psychology, University of New South Wales, Sydney, Australia*
[2]*Department of Cognitive Sciences, University of California, Irvine, USA*

ABSTRACT

Analyses of multi-attribute decision problems are dominated by accounts which assume people select from a repertoire of cognitive strategies to make decisions. This paper explores an alternative account based on sequential sampling and evidence accumulation. Two experiments varied aspects of a decision environment to examine competing models of decision behavior. The results highlighted the intra-participant consistency but inter-participant differences in the amount of evidence considered in decisions. This pattern was best captured by a sequential evidence accumulation model (SEQ) which treated pure Take-The-Best (TTB) and pure ''rational'' (RAT) models as special cases of a single model. The SEQ model was also preferred by the minimum description length (MDL) criterion to a naive strategy-selection model (NSS) which assumed that TTB or RAT could be selected with some probability for each decision. Copyright © 2010 John Wiley & Sons, Ltd.

KEY WORDS decision making; strategy selection; take-the-best; evidence accumulation; sequential sampling

## INTRODUCTION

A problem faced commonly by decision makers is determining how much information to incorporate into a decision. Some decisions are trivial (e.g., choosing a breakfast cereal), but some more important (e.g., choosing a mate), and consequently the amount of information or evidence examined prior to deciding will vary. One way to model this variance is to suggest that people sample evidence *sequentially* and adjust the amount of evidence they consider according to a *decision threshold*. Inherent in this conception is that thresholds will vary not just between decisions but also between individuals (some of us need more information than others when choosing clothes, for example). Newell (2005) suggested an ''adjustable spanner'' (or wrench) to capture this idea; a spanner in which the width of the jaws represents the amount of evidence a person accumulates before making a decision. An alternative way to model the variance is to

* Correspondence to: Ben R. Newell, School of Psychology, University of New South Wales, Sydney 2052, Australia.
E-mail: ben.newell@unsw.edu.au

suggest that people have access to a repertoire of decision strategies or ''tools'' which are suited for particular tasks. Such explanations are often invoked in the literature on both preferential choice (Beach & Mitchell, 1978; Christensen-Szalanski, 1978; Payne, Bettman, & Johnson, 1993) and inference (Gigerenzer & Todd, 1999), and recent work has begun to examine how decision makers might learn to select different strategies (Rieskamp, 2006; Rieskamp & Otto, 2006).

This paper provides an initial step in comparing these alternative theoretical conceptions. The performance of a sequential sampling model first proposed by Lee and Cummins (2004) is compared with a strategy selection model in two experiments. The experiments vary factors that have been shown to impact the adoption of decision strategies and the amount of evidence people consider in inference tasks. In Experiment 1 the format (image or text) of stimulus materials was varied, and in Experiment 2, the cost of cue information was manipulated. These manipulations provide ''test-beds'' in which to observe the behavior of the models under consideration. A further aim is to examine the validity of some of the assumptions underlying the sequential sampling model; assumptions which have been questioned by some researchers (e.g., Bergert & Nosofsky, 2007). We begin by introducing the models under consideration and explain why these particular models were chosen for the comparison.

## THE SEQUENTIAL SAMPLING MODEL

Sequential sampling processes have been extensively studied as models of human decision-making (e.g., Busemeyer & Rapoport, 1988; Busemeyer & Townsend, 1993; Busemeyer & Johnson, 2004; Laming, 1968; Nosofsky & Palmeri, 1997; Ratcliff, 1978; Vickers, 1979; Wallsten & Barton, 1982) particularly in relation to elementary psychophysical tasks, such as judging which of two lines is longer. Although there are many variants, the basic assumption of these models is that stimuli are searched for information until sufficient evidence has been accumulated to favor one decision, or no more information is available. These models have not often been applied to multiple-cue inference tasks, but their extension to such tasks is straightforward and provides a valuable alternative to the more commonly proposed strategy selection perspective.

Lee and Cummins (2004) presented a formal instantiation of an evidence accumulation model based on sequential sampling ideas for a two-alternative, multiple-cue task. In such tasks participants learn through trial-by-trial experience how to predict which of two objects is more likely to be higher on a given criterion. For example, participants might learn which of two companies is the better investment on the basis of indicators or cues such as employee turnover, or market share (e.g., Newell, Weston, & Shanks, 2003). Each cue in the environment has a predictive validity and participants must attempt to learn these validities in order to improve their performance. Thus one way of conceptualizing performance in these tasks is that participants consider the evidence provided by the different cues and make choices accordingly.

Lee and Cummins proposed that a model which considered cues sequentially in the order of their validity (i.e., highest to lowest) could be interpreted as a unification of ''take-the-best'' (TTB) (Gigerenzer & Goldstein, 1996) and ''rational'' models (RAT) because these two models represent the two extremes of evidence accumulation. TTB is a frugal heuristic which considers cues in the order of their validity like the sequential sampling model but stops search as soon as a cue which discriminates between alternatives is found. For example, one might discover that one company is a multi-national and the other is not and infer that the former is a better investment. TTB is a *noncompensatory* strategy because the ''best'' cue cannot be outweighed by any combination of less valid cues. TTB contrasts with *compensatory* strategies such as a linear weighted additive strategy (WADD) or the ''rational'' strategy (RAT) which we consider in this paper, because these strategies integrate cue values and choose the alternative with the higher weight of evidence. Such compensatory strategies are often considered the ''gold standard'' and rational or optimal because they weight appropriately and integrate the information available in a decision environment (Gigerenzer & Goldstein, 1996; Keeney & Raiffa, 1976; Payne et al., 1993). There is a long-standing interest in examining



DOI: 10.1002/bdm

the circumstances under which people's judgments and decisions accord with the frugal models exemplified by TTB or the more comprehensive models exemplified by WADD (e.g., Simon, 1956; see Newell, Lagnado, & Shanks, 2007 for a review). Thus these models are important ones to include in our comparison. The key advance of our perspective is that these models need not be considered as discrete, but rather points on a continuum of evidence accumulation.

Figure 1 provides a graphical illustration of the sequential sampling process used by Lee and Cummins (2004). This is the simplest and most popular form of sequential sampling, known as a *random walk* in which information is accumulated in a single tally as cues are observed, and a decision is made as soon as there is a threshold amount of evidence in favor of one alternative. (If all the cues are exhausted, and the threshold is not reached, we follow previous sequential sampling modeling (e.g., Lee & Corlett, 2003), in assuming that the alternative with the greatest evidence is chosen.)

Figure 1 illustrates a situation where the first cue provides strong evidence (measured on a standard log-odds scale) in favor of decision A, but all of the subsequent lower validity cues favor decision B. The log-odds scale is used because it is symmetric about the origin and additive: log-odds of zero mean that each decision is equally favored, and equal positive or negative increments represent equal amounts of evidence in favor of the two alternative decisions. (The *Method* section of Experiment 1 explains how cue validities are transformed into log-odds.) Once all cues have been observed, there is more evidence for decision B than A. Accordingly, for low thresholds (the value two is shown as a concrete example) decision A will be made; for higher threshold values (the value three is shown as a concrete example) decision B will be made.

In general, low thresholds that guarantee sampling terminates as soon as evidence favoring one option is found will model TTB decisions, while high thresholds that guarantee exhaustive sampling of all cues will model RAT decisions. The sequential sampling approach views these alternatives as special cases of a single evidence accumulation model corresponding to low (TTB) and high (RAT) evidence thresholds. Lee and Cummins (2004) tested the model in a multiple-cue judgment task and found that what they called the "unified" model[1] (unifying TTB and RAT) accounted for a higher proportion of participants' decisions

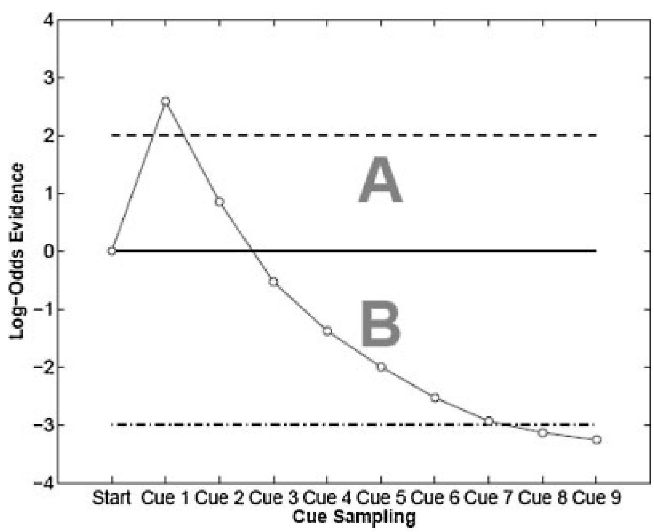


Figure 1. The sequential sampling model, showing the accumulated evidence as nine cues are sampled in validity order, and Take-The-Best (TTB)-consistent (top) and rational (RAT)-consistent (bottom) decision thresholds

[1]We acknowledge that other models which posit a single underlying representation (such as exemplar models) could also be considered "unified".



DOI: 10.1002/bdm

(84.5%) than either a pure TTB (36%) or RAT (64%) model. This better account of the data was not due simply to the unified model being more complex; the model was also favored by a minimum description length (MDL) model selection criterion that took into account the additional complexity of the unified model.

Lee and Cummins (2004) presented their model as an alternative conceptualization of behavior often interpreted as the selection of different decision making strategies. However, Lee and Cummins only compared the unified model with models that assumed *all* participants used either TTB or RAT. They did not compare the model with a *strategy selection* approach which assumes that, for each participant as they make each decision there is some probability of selecting TTB or RAT. Including a comparison with such an approach is an essential next step because current influential models of strategy selection propose this type of selection mechanism (e.g., *Strategy Selection Learning* (SSL) model, Rieskamp, 2006; Rieskamp & Otto, 2006).

## COMPARING THE MODELS

Before presenting the experiments we provide descriptions of the four models we consider, and present a demonstration of our methodology for their comparison, using hypothetical data. The environment used in our experiments is the same one originally developed by Lee and Cummins (2004), and subsequently used by Bergert and Nosofsky (2007). It involves a training phase, in which participants learn how the cues relate to outcomes, and a test phase designed to differentiate different models of human decision-making.

The environment itself is described in detail later, but to introduce the model comparison method three key features need to be noted: (1) the TTB and RAT strategies make identical choice-predictions in the training phase of the experiments removing the possibility that feedback will lead to one strategy being favored over another during training; (2) TTB and RAT make opposite choice-predictions in the test trials allowing inferences to be drawn about the strategy selected/evidence accumulated on each test trial; (3) Feedback is provided during training but not at test, preventing further learning from occurring during the test trials.

Using this experimental approach, we consider four different models of human decision-making. The key to evaluating the models is to specify what decisions they predict on the test trials, after training is finished.

*Pure TTB*: this model assumes that all participants adopt the TTB strategy for all test items. The TTB model states that once cue validities have been learned, for a given pair of objects the cue with the highest validity is consulted. If this cue discriminates between the stimuli (i.e., one object has the cue and the other does not, or only one object has a positive value of this cue), the favored stimulus is chosen and no further cues are examined. If the cue does not discriminate, the next cue in the validity order is considered and this process continues until cues are exhausted. Accordingly, for the test trials, which have by design a TTB-consistent and a RAT-consistent option, the TTB model simply predicts the TTB-consistent option will always be chosen.

*Pure RAT*: this model assumes that all participants adopt the RAT strategy for all test items. The RAT model states that the log-odds for each alternative are found by summing the evidence provided by each cue. The alternative with the higher weight of evidence is chosen; if the sums are equal a random choice is made. Accordingly, the RAT model simply predicts the RAT-consistent option will always be chosen at test.

*Sequential sampling model (SEQ)*: This model considers how the sequential sampling account best captures the TTB- and RAT-consistent decisions at the individual participant level. Formally, this can be achieved by allowing the model to assume a different evidence threshold for each participant. This captures the intuition that different people faced with the same decision (e.g., choosing a restaurant to dine at) might consider different levels of evidence (e.g., only the name of the restaurant, or the type of food, the cost, the location, etc.). For the environment we use, with its specific training and test questions and cue validities, it turns out that the SEQ model answers all the test questions consistent with the RAT approach once the threshold is above a critical value, but consistent with the TTB approach below that value. This means that the



DOI: 10.1002/bdm

SEQ model predicts each participant will either give RAT-consistent or TTB-consistent answers to all of the test questions, but not a mix of both.

*Naïve Strategy Selection (NSS)*: This model is premised on the notion that people have a repertoire of cognitive strategies available and that people initially select a strategy based on prior expectation of its success for a given decision problem (cf. Rieskamp & Otto, 2006). The subjective expectations of the appropriateness of these strategies then change through reinforcement learning as a function of the strategy's success in the environment. A strategy that performs well (i.e., makes correct predictions) is reinforced over trials in an experiment; a strategy that performs poorly is not reinforced and thus becomes less likely to be selected (cf., Rieskamp & Otto, 2006). Because TTB and RAT strategies make identical choice-predictions in the training trials of our experiments, the feedback provided during this stage should not differentially reinforce either the TTB or RAT decision strategy. This means that, at the beginning of test, the probability that TTB or RAT will be used is still given by a value close to their prior probabilities at the beginning of training. Since the test stage provides no feedback to reinforce one strategy over the other, this probability should remain approximately constant throughout the test stage. Accordingly, our conception of how NSS applies to the task involves a single parameter for each subject, which gives the probability that the TTB rather than RAT strategy will be used for each test question by that subject.

It is important to note that our NSS model is naive in comparison to Rieskamp and Otto's *SSL* model because in their model strategies can also be reinforced by a combination of accuracy and *effort*. Thus in their model when strategies' accuracies' are equated, a strategy which requires less effort will be reinforced. In the training phases of our experiments all cue information is available without cost. This means that there is little disincentive for people not to use all the information and thus behave in a RAT-consistent manner. Alternatively, one could argue that there will be a bias toward TTB because it is simpler in terms of cognitive effort (Lee & Cummins, 2004). Because we cannot infer strategies from choices in the training phases of our experiments (both strategies predict identical choices) and because we did not record cue-search and acquisition data from the training phases we are unable to distinguish these possibilities. Therefore, we focused on pure accuracy reinforcement in our naive model which in our environment leads both strategies to be equivalently reinforced during training.

## Minimum description length as a model selection criterion

The SEQ model has an evidence threshold parameter, and the NSS model has a strategy-selection parameter. This makes both models more complicated than the parameter-free TTB and RAT models. In addition, even though it has the same number of parameters as the NSS model (one per participant) the SEQ model is more constrained in the patterns of test trial decision-making it can predict. For the SEQ model, there can be inter-individual differences, but there must be intra-individual consistency (i.e., each participant is predicted to choose all RAT-consistent or all TTB-consistent options over their test questions). In the language of model selection theory, these constraints mean that the SEQ model has simpler *functional form* complexity than the NSS model.

It is important that these differences in both the number of parameters and functional form aspects of model complexity be taken into account when assessing the fit of the models to data. To achieve this, we used the MDL model selection criterion, which is sensitive to both goodness-of-fit and model complexity. The basic idea behind the MDL measure is to compress data by using a model, which identifies regularities in the data, and so allows a description that uses less information than would be required to list a data set literally. Formally, an MDL criterion is the length of this encoding of the data using the model, measured in bits, which are standard units of information in the fields of computer science and information theory. We used a version of the MDL criterion based on the "entropification" method developed by Grünwald (1999), and previously applied to evaluating the current decision-making models by Lee and Cummins (2004; see especially Lee, 2004 or the appendix in Lee, 2006 for a technical tutorial and Grünwald, 2007 for an extensive discussion of the MDL principle and its relationship to Bayesian and other model evaluation methods).



DOI: 10.1002/bdm

The chief advantage of the MDL measure in the current setting is that it can naturally be applied to deterministic models. More familiar model selection methods that balance goodness-of-fit with complexity, such as Bayes Factors (e.g., Kass & Raftery, 1995) and their approximation by information criteria such as the AIC and BIC, are only defined for probabilistic models. While most models of decision-making, like most models of cognition generally, are probabilistic, that is not the case for the simple TTB, RAT, and SEQ models we are considering. The TTB and RAT models simply assert that all participants will use one strategy for every decision they make. The SEQ model allows different participants to use different strategies, but as explained above predicts that all the test questions in the current experiments will be answered using only one of these strategies by any individual participant.

Taken literally, the deterministic nature of these models means they would be falsified automatically by almost any experimental data. For example, if just one participant on just one trial chooses the TTB alternative, the RAT model is falsified. Given the inherent noise in decision-making experimentation, this is an undesirable state of affairs. Our approach to this problem is based on the MDL criterion and preserves the simple deterministic nature of the heuristics, but treats the data as being inherently "noisy". There is always variation in data that is not fully explained by any set of cognitive models, and this additional variation is explicitly modeled by the MDL approach using an error rate parameter for the data. One intuitive way of understanding the MDL approach in modeling terms is that it effectively augments each of the RAT, TTB, SEQ, and NSS decision-making models with the same simple but principled error theory, so that they can all be compared directly to noisy experimental data.

NSS represents a point of calibration for the MDL measure, showing how a saturated model (one that makes probabilistic predictions, and so could potentially explain a large range of data patterns) fares against the inherently deterministic (or rule-based) TTB, RAT, and SEQ models. In this sense, we accept that the MDL measure for NSS is not a perfect representation of the complexity of NSS theory, but does provide a quantitative point of comparison to help understand the MDL measures for the alternative models being considered. Furthermore, this MDL approach makes it easy to compare the models over the whole range of different possible assumptions about the noisiness of the data. In addition, in the repeated-measures designs we consider, it is possible to estimate the level of noise, or error directly from the behavioral data (by examining the consistency of choices across repeated pairs of test items), and so inform the MDL analysis directly. We demonstrate these properties in the hypothetical example following.

**A demonstration with hypothetical data**

A concrete illustration of our MDL approach to evaluating the four models is provided by considering two hypothetical data sets. The first data set shown in Table 1A and then modeled in the left panel of Figure 2 shows data characterized by a high-degree of both inter- and intra-individual *inconsistency* in choices. The second data set shown in Table 1B and then modeled in the right panel of Figure 2 shows data characterized by a much higher degree of *intra*-individual *consistency* in choices. It is informative to consider how the models fare in capturing these patterns of consistency in choices within (*intra*) and across (*inter*) individuals. Table 1A shows the raw data for 10 hypothetical participants who have been through a training phase and then make 10 forced-choice decisions between pairs of objects (e.g., an inference about which of two companies is more profitable). For each decision, one object is the predicted choice of the TTB strategy (denoted with a T in the table) and the other is the predicted choice of the RAT strategy (denoted with an R). The data show that every participant tends to make several R and several T choices (i.e., a high degree of inter- and intra-individual *inconsistency*). The left panel of Figure 2 shows the performance of the four models in capturing these hypothetical data. The figure plots MDL on the *y*-axis and the error-rate on the *x*-axis. Error rate is a measure of the "noise" in the data (e.g., the degree to which a participant makes the *same* choice when faced with the *same* pair of objects). An error rate of 0 indicates perfect, error-free, data and an error-rate of 0.5 indicates random data (i.e., where the measurement is as likely to be correct as it is to be incorrect). Lower



DOI: 10.1002/bdm

Table 1. (A and B) Two hypothetical data sets for 10 participants (P1–10) answering 10 forced-choice test questions (Q1–10) about objects which discriminate RAT (R) and TTB (T) choices

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | |
| P1 | R | R | T | R | T | T | R | R | T | T |
| P2 | T | R | R | R | T | T | R | T | R | T |
| P3 | R | R | T | T | R | T | R | T | R | T |
| P4 | R | T | R | R | T | R | R | T | T | T |
| P5 | R | T | R | T | R | R | T | R | T | T |
| P6 | R | T | R | T | R | T | R | T | R | R |
| P7 | T | R | T | R | R | T | T | R | R | R |
| P8 | R | T | R | R | T | R | R | R | T | T |
| P9 | R | R | T | R | R | T | T | R | R | R |
| P10 | R | T | R | R | R | T | R | T | R | R |
| B | | | | | | | | | | |
| P1 | T | T | T | T | T | T | T | T | T | T |
| P2 | T | T | T | T | T | T | T | T | T | T |
| P3 | T | T | T | T | T | T | T | T | R | T |
| P4 | T | T | R | R | T | R | T | T | T | R |
| P5 | R | T | T | T | R | R | T | T | R | R |
| P6 | R | R | R | R | T | T | R | T | R | R |
| P7 | R | R | R | R | R | R | R | R | R | R |
| P8 | R | R | R | R | R | R | R | R | R | R |
| P9 | R | R | R | R | R | R | R | R | R | R |
| P10 | R | R | R | R | R | R | R | R | R | R |

The data in Table 1A display high inter- and intra- individual *inconsistency* in decision behavior. The data in Table 1B display high inter-individual *inconsistency* but high intra-individual *consistency* in decision behavior.
R refers to a choice of the object predicted the rational (RAT) strategy; T refers to a choice of the object predicted by theTake-the-Best (TTB) strategy.

values of MDL indicate a more likely model, so the figure shows that the NSS model provides the best fit to these data for all levels of error up to approximately 0.17. After this point the three other models provide better fits, with the RAT model providing the best fit to the data.

It is worth noting how the assumed error rate relates to the balance between goodness-of-fit and complexity in comparing the models. When the error rate is zero, the data are considered exact, and so, for the data in Table 1A, only the NSS model can account for the data. In the left panel of Figure 2, this is clear because the NSS model has the lowest MDL value (in fact, the other models have infinitely large MDL values, because they mis-predict at least one decision). As the error rate is assumed to be larger, however, model complexity is emphasized and goodness-of-fit is de-emphasized by the MDL criterion. At the extreme error rate of 0.5, where the data are essentially arbitrary (i.e., each datum could equally well be a TTB- or RAT-consistent decision), the MDL measures just reflect complexity, favoring the simplest RAT and TTB models, followed by the SEQ and then the NSS models.

Table 1B shows the data from a second set of hypothetical participants given the same task. These data are characterized by a much higher degree of *intra*-individual *consistency.* Although the overall number of R and T choices is the same as in Table 1A most participants' choices are dominated by one or other strategy. The right panel of Figure 2 shows that when the data are virtually error-free NSS is the preferred model, but as soon as there is some error the SEQ model provides the best fit and remains the preferred model until the data are almost random.

In addition to the fits of each model one can calculate the proportion of decisions predicted by the models. The naive NSS model, because it is a saturated model can predict all patterns of decisions with varying levels



DOI: 10.1002/bdm

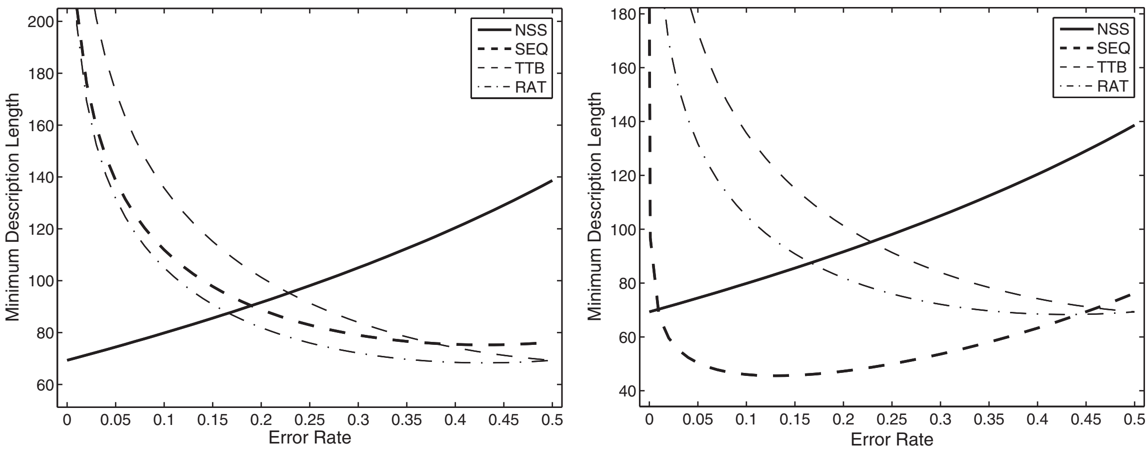


Figure 2. Minimum Description Length (MDL) results for the demonstration with hypothetical data. The curves show the MDL criteria for each of the four decision-making models, as a function of the assumed error rate of the experimental data shown from Table 1A (left panel) and Table 1B (right panel). Lower values of MDL indicate a more likely model

of probability. In contrast, the ability of the SEQ model to explain decisions is affected by the regularities in the data. The data in Table 1A lack regularity both at the inter and intra-individual level and thus SEQ can only predict 57% of the decisions—equivalent to the performance of RAT, which is in turn simply a reflection of the higher number of R decisions in the data set (57/100). TTB can predict the remaining 43% of decisions. However, given the regularities in the data in Table 1B, SEQ benefits from the intra-individual consistency and is able to predict 87% of the decisions. The 13 predictions it gets incorrect are those which are inconsistent with the majority of an individual's decisions (i.e., Q9 of Participant 3; Q3, 4, and 10 of Participant 4; half of Participant 5's responses, and Q5, 6, and 8 of Participant 6). The performance of TTB and RAT remain the same as for the Table 1A data.

Taken together, these demonstrations show that the MDL measure can find evidence for each of the models depending on the regularities in the empirical data. If most decisions are TTB- or RAT-consistent, the simplicity of those heuristics will be preferred. If there are many TTB- and RAT-consistent decisions over all participants, but each individual participant tends to use only one approach, the SEQ model will be preferred. If this intra-individual consistency is also absent, the NSS model will be preferred. We now turn to the experiments which provided the test-beds for our model comparisons.

## EXPERIMENT 1: A COMPARISON OF THE FOUR MODELS IN ENVIRONMENTS WITH TEXT AND GRAPHICAL CUE FORMATS

The main aim of Experiment 1 was to compare the performance of the SEQ, RAT, TTB, and NSS models. To achieve this aim we used a multiple-cue inference task that had the same statistical properties as that used previously by Lee and Cummins (2004 see also Bergert & Nosofsky, 2007). An additional aim of Experiment 1 was to examine the effect of different cue format instantiations. The two previous studies that have used the cue environment we adopted have displayed cue values in graphical or picture formats (Bergert & Nosofsky, 2007; Lee & Cummins, 2004). In contrast, the majority of multiple cue-inference tasks present stimulus information to participants as text lists (i.e., word descriptions of the values of cues). Some research suggests that the format in which cue information is presented (text or image) affects the adoption of different types of



DOI: 10.1002/bdm

decision strategies (Bröder & Gaissmaier, 2007; Bröder & Schiffer, 2003, 2006). Given that previous studies with the current experimental environment have only used image-based cues, it is of interest to see if format impacts on the ability of the models to account for the data. For example, might presenting discrete text-based cues promote the single discriminating cue stopping-rule of TTB? (e.g., as in Bröder & Schiffer, 2003). Thus cue information was presented either in text or image format (between-subjects) in training and at test items were presented separately in text and image formats (within-subjects).

## METHOD

### Participants

Forty-eight undergraduate students (23 male, 25 female; mean age = 18 years) from the University of New South Wales participated in the experiment in return for course credit. There was an error storing the data for one participant, giving a final total of 47 participants.

### Stimuli

The statistical structure of the stimulus environment used in the experiments was developed by Lee and Cummins (2004) and adapted from a real-world environment examined by Czerlinski, Gigerenzer, and Goldstein (1999); Lee and Cummins (2004) give a full account of its construction (see also Bergert & Nosofsky, 2007). The environment comprises 16 objects described by six binary cues. Table 2 displays this environment showing the cue patterns for each stimulus, the associated decision variable for each pattern (see Procedure section for explanation of what the decision variable refers to), and the validity of each cue where cue validity is defined using a Bayesian measure in which the validity of cue $i$ ($v_i$) is defined as

$$v_i = \frac{\text{number of correct decisions made by the } i\text{th cue} + 1}{\text{number of decisions made by the } i\text{th cue} + 2} \tag{1}$$

The evidence value of each cue is the log-odds of the Bayesian cue validity. For example, Cue 1 makes 59 correct decisions out of the 60 decisions in which it discriminates (i.e., predicts a unique choice); by Equation 1 this leads to a Bayesian validity of 60/62 or .968, which has a log-odds value of 3.40 (ln(.968/(1 − .968). In the training phase, participants received 119 training trials, which constitute all but one possible pairings of the 16 objects in Table 2. The exception, as with Lee and Cummins (2004), is the pairing of the second and seventh stimuli. This pairing was omitted because the TTB and RAT models make opposing predictions. For all remaining 119 pairings, the TTB and RAT models make the same prediction. The test items are displayed in the lower section of Table 2. In contrast to the training pairs, these items are designed specifically to distinguish between choices that accord to the predictions of the RAT and TTB model. For each pair, the TTB model selects the stimulus on the left because it has a positive value for the most predictive cue. The RAT model makes the opposite prediction because the stimulus on the right always has more evidence favoring it once all the cues are assessed (i.e., the sum of the log-odds is higher).

To enable a cover-story to be used (see procedure section) the cue environment was instantiated as six pieces of clothing—baseball cap, t-shirt, handbag, skirt, stockings, and shoes—each of which could be one of two colors. The stimuli were either text descriptions of these clothing items, or schematic images of a woman wearing these items. The assignment of cue validities to clothing items was random and differed for each participant (e.g., for one participant a value of 1 indicated "green shoes" and 0 indicated "yellow shoes" for another it would be reversed).



DOI: 10.1002/bdm

Table 2. The training and test environments used in Experiments 1and 2, showing cue patterns, cue validities, decision variable values and assignment of cues to the TTB-and RAT-consistent stimuli in the test environment

| Object number | Cue 1 (.97) | Cue 2 (.90) | Cue 3 (.82) | Cue 4 (.64) | Cue 5 (.56) | Cue 6 (.55) | Decision variable |
|---|---|---|---|---|---|---|---|
| Training environment | | | | | | | |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 16 |
| 2 | 0 | 1 | 0 | 0 | 1 | 0 | 18 |
| 3 | 0 | 0 | 1 | 0 | 0 | 1 | 21 |
| 4 | 0 | 0 | 0 | 1 | 1 | 0 | 25 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 31 |
| 6 | 1 | 0 | 0 | 0 | 1 | 1 | 40 |
| 7 | 0 | 0 | 1 | 1 | 1 | 1 | 44 |
| 8 | 1 | 1 | 0 | 1 | 0 | 0 | 51 |
| 9 | 1 | 1 | 1 | 0 | 0 | 1 | 62 |
| 10 | 1 | 1 | 0 | 0 | 1 | 0 | 70 |
| 11 | 1 | 1 | 0 | 1 | 1 | 1 | 97 |
| 12 | 1 | 1 | 1 | 1 | 0 | 0 | 104 |
| 13 | 1 | 1 | 1 | 1 | 1 | 1 | 280 |
| 14 | 1 | 1 | 1 | 1 | 0 | 1 | 285 |
| 15 | 1 | 1 | 1 | 0 | 1 | 0 | 347 |
| 16 | 1 | 1 | 1 | 1 | 1 | 0 | 444 |

| Test Pair No. | TTB object | RAT object |
|---|---|---|
| Test environment | | |
| 1 | 1 0 0 0 0 1 | 0 1 1 0 0 0 |
| 2 | 1 0 0 0 1 0 | 0 1 1 0 0 0 |
| 3 | 1 0 0 0 1 1 | 0 1 1 1 0 0 |
| 4 | 1 0 0 1 1 0 | 0 1 1 1 0 0 |
| 5 | 1 0 0 1 1 1 | 0 1 1 1 1 0 |

In the experiments the values 1 and 0 indicate one of two colors for each item of clothing. The value of the decision variable was determined by the original raw data set from which these stimuli were drawn (see Lee & Cummins, 2004). For the test environment cues are arranged in validity order (left to right).

### *Training phase procedure*

Participants were told that they were an undercover agent in a fictional country and had to learn about the clothing characteristics of members of a secret society. On each of the 119 trials in the training phase the two paired stimuli were presented on screen, and participants selected the woman they thought more likely to be a secret society member. Twenty-four participants were trained using the text descriptions, and the remaining 23 participants were trained using the schematic image representation (randomly assigned). In both formats of presentation, the correct answer was determined by the stimulus with the higher decision variable (referring to the higher likelihood of being a secret society member), and feedback was given on each trial by indicating the correct choice.

### *Test phase*

Following training, participants completed a test phase, involving two blocks of 20 trials. Both blocks comprised four repetitions of the five test questions displayed in Table 2 presented in a random order for each participant. The repetitions were included to gain a more stable estimate of the decisions participants made (Bergert & Nosofsky, 2007), and also facilitated the estimation of the level of noise in the data required for



DOI: 10.1002/bdm

the MDL analysis. One of these blocks of 20 trials was presented using the text format, while the other was presented using the image format. The order of the formats was counterbalanced across participants. No feedback was given during the test phase. A "time-out" trial was recorded if participants did not respond within 15 seconds.

## RESULTS

### Training trials

There was a clear increase in accuracy across learning trials for both groups indicating that participants were able to learn from the feedback (average accuracy across the last 19 trials Text group: .79, Image group: .80). There was a significant linear trend across learning trials, $F(1,45) = 48.80$, $p < .001$, no effect of group, and no interaction between the two variables $Fs < 1$.

### Test decisions

The raw data for the test decisions from the Image and Text test phases are shown in Table 3. The data are the number of choices of the TTB-predicted object on each of the four repetitions of the 5 test questions. Thus a participant who always chose the TTB object would have a 4 in each column (e.g., Participant 5 in the Image condition data) and one who always chose the RAT object would have 0 (e.g., Participant 43 in the Image condition data). The data are arranged in order with TTB-consistent participants at the top of the columns and RAT-consistent at the bottom. Comparison of the two sets of columns suggests that the format manipulation had little effect on the distribution of TTB- and RAT- consistent participants but highlight the intra-participant consistency and inter-participant differences in decision behavior predicted by the sequential sampling model. There was some evidence of participants switching strategies between formats (e.g., participants 5 and 42); but it was not systematic (i.e., similar numbers switched in both directions—TTB-Text to RAT-Image and vice-versa).

### Model comparisons

Figure 3 displays the model comparisons for the data from the Image condition (Left Panel) and the Text condition (Right Panel). Because participants made repeated decisions (4 for each test-pair) we can estimate the amount of error in the data. The 95% confidence interval around this estimate is depicted by the gray shaded column in the figure. The point at which the lines intersect this shaded area is the region of interest for our model comparison. Both figures show similar qualitative patterns with a clear advantage in terms of MDL for the SEQ model in both conditions. The NSS model can account for or fit all the decisions in both conditions (because there is always some probability a participant will use either the TTB or the RAT strategy for any decision), but as the figures show the model is punished (in terms of MDL value) for its additional complexity. In contrast, the more constrained SEQ model predicted 86% and 88% of decisions in the Image and Text conditions, respectively, compared to 52 and 54% for TTB and 48 and 46% for RAT.

## DISCUSSION

Consistent with previous studies, Experiment 1 provides considerable evidence for inter-individual differences but intra-individual consistency in a multi-attribute inference task. The sequential sampling model (SEQ) that allowed for TTB and RAT-users provided the best fit to the data. Importantly, under the MDL framework we adopted, this model provided a better fit than a naive strategy selection model (NSS) at



DOI: 10.1002/bdm

Table 3. Raw data from the Image (upper section) and Text (lower section) Test Phases of Experiment 1 showing the number of choices of the Take the Best (TTB)-predicted object for each participant for the four repeats of the five test questions (see Table 2 for test questions)

| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Image test phase | | | | | |
| P5 | 4 | 4 | 4 | 4 | 4 |
| P12 | 4 | 4 | 4 | 4 | 4 |
| P19 | 4 | 4 | 4 | 4 | 4 |
| P23 | 4 | 4 | 4 | 4 | 4 |
| P24 | 4 | 4 | 4 | 4 | 4 |
| P27 | 4 | 4 | 4 | 4 | 4 |
| P30 | 4 | 4 | 4 | 4 | 4 |
| P31 | 4 | 4 | 4 | 4 | 4 |
| P38 | 4 | 4 | 4 | 4 | 4 |
| P39 | 4 | 4 | 4 | 4 | 4 |
| P45 | 4 | 4 | 4 | 4 | 4 |
| P16 | 4 | 4 | 3 | 4 | 4 |
| P41 | 4 | 4 | 4 | 3 | 4 |
| P15 | 4 | 4 | 4 | 4 | 2 |
| P36 | 4 | 4 | 3 | 3 | 4 |
| P37 | 4 | 3 | 4 | 3 | 4 |
| P9 | 4 | 4 | 3 | 3 | 3 |
| P10 | 4 | 3 | 4 | 2 | 4 |
| P20 | 3 | 2 | 3 | 4 | 4 |
| P40 | 4 | 3 | 4 | 4 | 1 |
| P8 | 2 | 3 | 2 | 2 | 3 |
| P47 | 1 | 3 | 2 | 4 | 2 |
| P2 | 0 | 2 | 3 | 2 | 4 |
| P3 | 2 | 1 | 2 | 2 | 3 |
| P6 | 1 | 3 | 3 | 0 | 3 |
| P46 | 1 | 2 | 0 | 3 | 3 |
| P13 | 1 | 3 | 2 | 0 | 2 |
| P26 | 2 | 3 | 2 | 0 | 1 |
| P29 | 2 | 1 | 2 | 1 | 1 |
| P14 | 1 | 3 | 1 | 0 | 1 |
| P33 | 0 | 0 | 1 | 4 | 0 |
| P7 | 2 | 0 | 0 | 0 | 2 |
| P35 | 3 | 0 | 1 | 0 | 0 |
| P21 | 1 | 0 | 2 | 0 | 0 |
| P11 | 0 | 0 | 0 | 0 | 1 |
| P25 | 0 | 0 | 0 | 0 | 1 |
| P32 | 0 | 1 | 0 | 0 | 0 |
| P34 | 0 | 1 | 0 | 0 | 0 |
| P44 | 0 | 1 | 0 | 0 | 0 |
| P1 | 0 | 0 | 0 | 0 | 0 |
| P4 | 0 | 0 | 0 | 0 | 0 |
| P17 | 0 | 0 | 0 | 0 | 0 |
| P18 | 0 | 0 | 0 | 0 | 0 |
| P22 | 0 | 0 | 0 | 0 | 0 |
| P28 | 0 | 0 | 0 | 0 | 0 |
| P42 | 0 | 0 | 0 | 0 | 0 |
| P43 | 0 | 0 | 0 | 0 | 0 |

(*Continues*)



DOI: 10.1002/bdm

Table 3. (Continued)

| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| Text test phase | | | | | |
| P4 | 4 | 4 | 4 | 4 | 4 |
| P6 | 4 | 4 | 4 | 4 | 4 |
| P9 | 4 | 4 | 4 | 4 | 4 |
| P12 | 4 | 4 | 4 | 4 | 4 |
| P15 | 4 | 4 | 4 | 4 | 4 |
| P16 | 4 | 4 | 4 | 4 | 4 |
| P23 | 4 | 4 | 4 | 4 | 4 |
| P24 | 4 | 4 | 4 | 4 | 4 |
| P27 | 4 | 4 | 4 | 4 | 4 |
| P30 | 4 | 4 | 4 | 4 | 4 |
| P37 | 4 | 4 | 4 | 4 | 4 |
| P38 | 4 | 4 | 4 | 4 | 4 |
| P39 | 4 | 4 | 4 | 4 | 4 |
| P41 | 4 | 4 | 4 | 4 | 4 |
| P22 | 4 | 3 | 3 | 4 | 4 |
| P26 | 3 | 4 | 4 | 3 | 4 |
| P42 | 4 | 2 | 4 | 4 | 4 |
| P7 | 3 | 4 | 4 | 4 | 2 |
| P28 | 4 | 4 | 4 | 2 | 2 |
| P36 | 4 | 4 | 3 | 4 | 1 |
| P40 | 2 | 3 | 3 | 2 | 4 |
| P13 | 1 | 4 | 2 | 2 | 4 |
| P46 | 3 | 1 | 3 | 3 | 3 |
| P2 | 3 | 2 | 2 | 3 | 1 |
| P19 | 1 | 3 | 2 | 3 | 2 |
| P31 | 3 | 1 | 1 | 2 | 4 |
| P34 | 3 | 0 | 0 | 4 | 4 |
| P29 | 2 | 3 | 2 | 1 | 2 |
| P3 | 3 | 1 | 3 | 0 | 2 |
| P21 | 1 | 2 | 3 | 0 | 1 |
| P10 | 1 | 0 | 0 | 2 | 1 |
| P20 | 1 | 2 | 1 | 0 | 0 |
| P14 | 0 | 1 | 1 | 0 | 0 |
| P25 | 2 | 0 | 0 | 0 | 0 |
| P33 | 0 | 1 | 0 | 1 | 0 |
| P35 | 0 | 0 | 0 | 2 | 0 |
| P32 | 1 | 0 | 0 | 0 | 0 |
| P1 | 0 | 0 | 0 | 0 | 0 |
| P5 | 0 | 0 | 0 | 0 | 0 |
| P8 | 0 | 0 | 0 | 0 | 0 |
| P11 | 0 | 0 | 0 | 0 | 0 |
| P17 | 0 | 0 | 0 | 0 | 0 |
| P18 | 0 | 0 | 0 | 0 | 0 |
| P43 | 0 | 0 | 0 | 0 | 0 |
| P44 | 0 | 0 | 0 | 0 | 0 |
| P45 | 0 | 0 | 0 | 0 | 0 |
| P47 | 0 | 0 | 0 | 0 | 0 |

The shaded regions include all participants classified as TTB-consistent (top half of the tables) or rational RAT-consistent (bottom half of the tables) according to an 80% consistency rule (at least 16/20 decisions as predicted by the model*).

The underlined numbers correspond to participant-question pairs where only three (not the planned four) repeated answers were obtained, because the participant "timed out" on one trial.

*Binomial probability of 16/20 responses in favor of one model is 0.998.





DOI: 10.1002/bdm

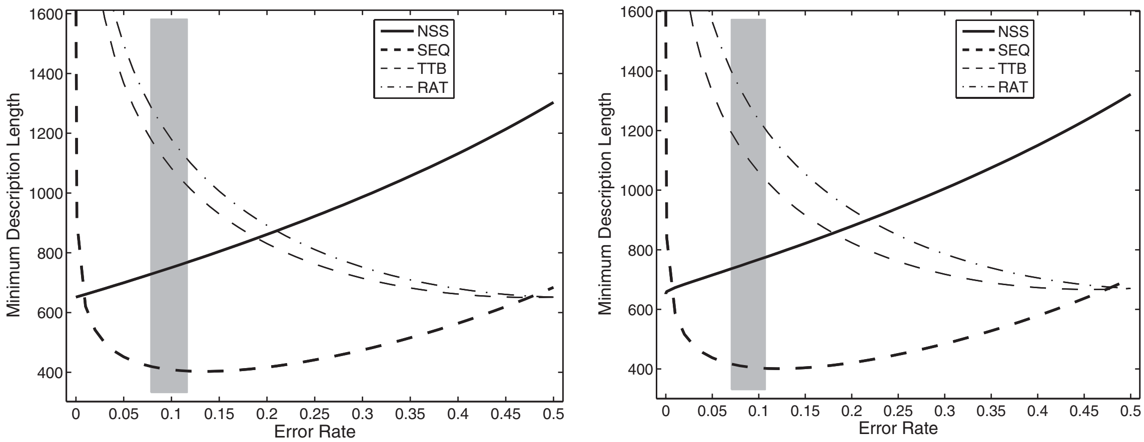


Figure 3. Minimum Description Length (MDL) results for the Text versus Image format comparison of Experiment 1. The curves show the MDL criteria for each of the four decision-making models, as a function of the assumed error rate of the experimental data for the image presentation decisions (left panel) and the text presentation decisions (right panel). Lower values of MDL indicate a more likely model. In each panel, the shaded box spans the 95% confidence interval for the error rate, estimated from the repeated-measures in the decision data. Naive Strategy Selection Model (NSS); Sequential Sampling Model (SEQ); Take-The-Best (TTB); Rational (RAT)

the estimated level of noise for the data and, indeed, for the whole range of plausible assumptions about the variability in the data.

Format had very little systematic effect on performance, suggesting that the effects found previously with this cue environment were not dependent on using image-based cues. One key difference between this study and those which have observed format effects is that the latter involve retrieval of cue information from *memory* rather than having the information available on the screen (e.g., Bröder & Schiffer, 2003, 2006). A potential follow up to Experiment 1 would be to convert the task we used to an ''inference-from-memory'' task and examine the capability of the SEQ model to account for decisions under those conditions. While this might be an interesting approach to take we decided not to because in our second experiment we wanted to incorporate process-level measures of behavior in order to test some of the underlying assumptions of the models. The process-level measures we were interested in, such as the order in which cues are searched and the number acquired are very difficult to implement in inferences from memory tasks (though see Bröder & Gaissmaier, 2007).

## EXPERIMENT 2: A COMPARISON OF THE FOUR MODELS IN ENVIRONMENTS WITH DIFFERENT INFORMATION COSTS

The principal aim of Experiment 2 was to again compare the performance of the four models TTB, RAT, SEQ, and NSS in an experimental environment in which a factor known to affect decision behavior was varied. The factor was the cost of cue information (high cost vs. low cost). This factor served our purposes for two reasons. First, forcing participants to search explicitly for and acquire cues enabled examination of the order in which cues were acquired, the number of cues acquired, and the amount of evidence accumulated at the point of terminating search. These measures can reveal whether a participant classified on the basis of their choices as consistent with one or other model, also follows the search and stopping rules of that model (cf. Bergert & Nosofsky, 2007; Newell & Shanks, 2003). Second, although the manipulation of cue cost per se



DOI: 10.1002/bdm

may not be particularly interesting, examining how the models cope with the variation in behavior induced by the cost manipulation *is* worthy of investigation. For example, is it the case that when information is expensive the appeal of the frugal TTB model is such that it provides the best account of the data? Thus, in Experiment 2 we used the same stimulus structure as in the text-based condition of Experiment 1 but placed an explicit cost on obtaining cue information during the critical test trials. The cue costs were only implemented in the test trials in order to allow participants the opportunity to learn the validities of the cues during training, without incurring costs.

## METHOD

### Participants

Forty-eight undergraduate students (10 male, 38 female; mean age = 20 years) from the University of New South Wales participated in the experiment in return for course credit. Participants were assigned randomly to either *High Relative Cost* (HRC) condition or the *Low Relative Cost* (LRC) condition resulting in 24 participants in each condition. (See the *Procedure* for an explanation of the relative costs.)

### Stimuli

Experiment 2 used the same cue environment as Experiment 1 (see Table 2) but all participants were shown the text description labels. The assignment of cue validities to clothing items was random and differed for each participant.

### Procedure

The procedure during the 119 training trials was identical to that used in the text training condition of Experiment 1(i.e., all cue values appeared simultaneously on the screen) with the addition that participants earned points for every correct answer. In the HRC condition participants earned 35 points for each correct choice and in the LRC condition they earned 70 points. There were no penalties for incorrect choices. Participants were told that at the end of the experiment all points earned would be converted to actual money at a rate of 100 points = 10 Australian cents. The number of points earned was displayed on the screen throughout the training phase.

The 20 test trials consisted of the five critical comparisons shown in Table 2 repeated 4 times (in a random order). Cue information was no longer freely available but had to be purchased by clicking on a ‘‘buy’’ button adjacent to each pair of cues. In both conditions information cost five points representing a HRC—14% of the 35 point payoff for a correct answer—in the *HRC* condition; and a LRC—7% of the 70 point payoff for a correct answer—in the *LRC* condition. Clicking on the buy button revealed the value of that cue for each of the two stimuli. The order of cue purchase and the number of cues purchased was left to the participant, with the exception that at least one cue had to be purchased on every trial before making a choice. There was no time limit for decisions.

Participants were told that the amount received for correct answers given at test remained the same as it had been in training, but their current points balance was no longer displayed on the screen and no feedback was given in order not to reinforce a particular strategy. Participants were told that the computer was keeping track of their points spent and their earnings and that they would be paid accordingly at the end of the experiment. In order to emphasize the cost of information the number of points spent on cues on the current trial was displayed. On completion of the test trials participants were debriefed and paid their earnings from the experiment.



DOI: 10.1002/bdm

**Alternative cost manipulation**

We chose to manipulate information cost in a relative sense in Experiment 2 in order to be consistent with previous research in this area. For example, Bröder (2003) argued that keeping the nominal cost of information constant (and manipulating its relative impact on gains and losses) is preferable to a direct manipulation of cost. This is because direct impositions of high costs might simply lead to cost aversion (Bröder, 2003; see also Rieskamp & Otto, 2006). Nonetheless given the potential for ambiguity in interpreting the relative costs, we also ran a replication of Experiment 2 in which the directly expressed costs were manipulated but the pay-off for each decision remained constant. The results of this experiment accorded largely with those of Experiment 2 so we do not report them in full. We do present the results of the model comparison to reassure readers that our conclusions do not rest on a particular method for manipulating cost[2].

## RESULTS

**Training phase**

The HRC group achieved .69 correct inferences by the last block of training and the LRC group .73.There was a significant linear trend across training trials, $F(1,46) = 21.59$, $p < .001$ indicating learning; there was no effect of group, and no interaction between the two variables $Fs < 1$. The absence of a group effect indicates that the two reward schemes (35 or 70 points) were equally motivating; note that information was free in the training phase.

**Test phase: cue purchase**

At test, participants in the HRC condition purchased an average of 2.32 cues per trial compared to an average of 2.84 cues per trial in the LRC condition—a marginally significant difference, $F\ (1, 47) = 3.30$, $p < .077$. A similar marginal effect was found on the mean number of discriminating cues purchased (HRC = 1.83, LRC = 2.23), $F\ (1,\ 47) = 3.57$, $p < .066$.

In addition to the number of cues purchased we examined the level of *evidence* (i.e., the sum of the log odds of the cues acquired) at which search was terminated. We identified all the participants who were either RAT- or TTB-consistent for both the High and Low Relative Cost environments (see Table 4) and then used their observed cue searching behavior on all 20 individual decisions to calculate the level of evidence at which they chose to terminate search. This means, for example, that from the seven TTB-consistent participants in the HRC condition, we found $7 \times 20 = 140$ terminating evidence values. An ANOVA conducted on the pooled levels of terminating evidence found that participants in the HRC condition had lower values of terminating evidence than those in the LRC condition (HRC = 1.86, LRC = 2.11) $F\ (1{,}640) = 3.44$, $p < .07$; and that RAT-users had higher terminating levels of evidence than TTB-users (RAT = 2.75, TTB = 1.22) $F\ (1{,}640) = 128.19$, $p < .001$.

The preceding analyses show that the cost manipulation had a modest but predicted effect on cue acquisition and evidence accumulation (more cues, and more evidence was accumulated under lower costs). However, comparison of the upper and lower sections of Table 4 shows that, although there was a high degree

[2]The additional experiment ($N = 24$) used the same stimulus environment and had a training phase in which all participants earned 100 points for each correct decision; these points were not converted into money. At test information cost 1cent per cue in the low cost condition and 12 cents per cue in the high cost condition; the pay-off for a correct decision was 80 cents in both conditions. Any money earned in this phase was given to the participants at the end of the experiment. Consistent with Experiment 2, the cost manipulation had little overall effect on the proportion of TTB-consistent decisions, or the allocation of participants as TTB or RAT-users, but had the predicted impact on the number of cues purchased per trial (total and discriminating) with participants in the High Cost condition purchasing approximately 18% fewer cues than those in the Low Cost condition.



DOI: 10.1002/bdm

Table 4. Raw data from the High Relative Cost (upper section) and Low Relative Cost (lower section) conditions of Experiment 2 showing the number of choices of the Take-The-Best (TTB)-predicted object for each participant for the four repeats of the five test questions (see Table 2 for test questions)

| | Q1 | Q2 | Q3 | Q4 | Q5 |
|---|---|---|---|---|---|
| High relative cost condition | | | | | |
| P2 | 4 | 4 | 4 | 4 | 4 |
| P15 | 4 | 4 | 4 | 4 | 4 |
| P19 | 4 | 4 | 4 | 4 | 4 |
| P6 | 3 | 4 | 3 | 4 | 4 |
| P11 | 4 | 2 | 4 | 4 | 4 |
| P23 | 3 | 3 | 3 | 4 | 4 |
| P9 | 3 | 3 | 3 | 4 | 3 |
| P22 | 3 | 4 | 2 | 4 | 3 |
| P16 | 1 | 2 | 2 | 2 | 3 |
| P21 | 0 | 2 | 2 | 4 | 2 |
| P7 | 0 | 2 | 3 | 1 | 3 |
| P1 | 3 | 1 | 1 | 1 | 2 |
| P10 | 0 | 1 | 3 | 1 | 3 |
| P5 | 2 | 0 | 2 | 1 | 2 |
| P13 | 1 | 1 | 0 | 1 | 2 |
| P14 | 0 | 2 | 0 | 0 | 2 |
| P17 | 3 | 0 | 1 | 0 | 0 |
| P3 | 0 | 2 | 1 | 0 | 0 |
| P8 | 0 | 1 | 1 | 0 | 1 |
| P12 | 0 | 0 | 3 | 0 | 0 |
| P18 | 0 | 0 | 2 | 0 | 0 |
| P20 | 0 | 1 | 0 | 0 | 0 |
| P4 | 0 | 0 | 0 | 0 | 0 |
| P24 | 0 | 0 | 0 | 0 | 0 |
| Low relative cost condition | | | | | |
| P5 | 4 | 4 | 4 | 4 | 4 |
| P6 | 4 | 4 | 4 | 4 | 4 |
| P13 | 4 | 4 | 4 | 4 | 4 |
| P23 | 4 | 4 | 4 | 4 | 4 |
| P24 | 4 | 4 | 4 | 4 | 4 |
| P4 | 3 | 4 | 4 | 4 | 4 |
| P10 | 4 | 4 | 4 | 3 | 4 |
| P19 | 4 | 4 | 3 | 4 | 4 |
| P18 | 3 | 4 | 4 | 4 | 3 |
| P22 | 4 | 4 | 4 | 4 | 1 |
| P16 | 4 | 2 | 4 | 3 | 3 |
| P1 | 4 | 2 | 3 | 4 | 1 |
| P20 | 2 | 3 | 3 | 3 | 3 |
| P3 | 2 | 3 | 4 | 1 | 3 |
| P2 | 3 | 3 | 3 | 3 | 0 |
| P17 | 2 | 3 | 1 | 2 | 2 |
| P7 | 3 | 1 | 0 | 0 | 3 |
| P8 | 1 | 1 | 2 | 0 | 2 |
| P9 | 0 | 2 | 1 | 1 | 1 |
| P14 | 2 | 0 | 2 | 1 | 0 |
| P11 | 0 | 2 | 0 | 1 | 1 |
| P15 | 1 | 1 | 2 | 0 | 0 |
| P12 | 0 | 1 | 1 | 0 | 0 |
| P21 | 0 | 0 | 0 | 0 | 0 |

The shaded regions include all participants classified as TTB-consistent (top half of the table) or rational (RAT)-consistent (bottom half of the table) according to an 80% consistency rule (at least 16/20 decisions as predicted by the model*).
*Note*: *Binomial probability of 16/20 responses in favor of one model is 0.998.



DOI: 10.1002/bdm

of intra-participant consistency and inter-participant differences, the cost manipulation had little systematic effect on the *distribution* of TTB- and RAT-consistent participants in the two cost conditions. Contrary to prediction there were four more TTB-consistent participants in the condition in which information was cheaper (LRC) than in the higher cost condition. Similarly there were fewer RAT-consistent users in the LRC than in the HRC condition. This unexpected distribution does not invalidate the cost manipulation—participants did purchase fewer cues and search for less evidence when it was costly—but it does raise questions about the classification of participants as TTB and RAT-consistent. This issue is considered further in the section on process data.

## Model comparison

Figure 4 displays the model comparisons for the data from the HRC condition (Left Panel) and the LRC condition (Right Panel). The region of interest for our comparison is the gray shaded column which indicates the 95% confidence interval around the estimate of error in our data. Both figures show similar qualitative patterns (to each other and to those from Experiment 1) with an advantage in terms of MDL for the SEQ model. The slightly better performance of the RAT model over the TTB model in the HRC condition, and the reversal of this pattern in the LRC condition, reflects the higher number of TTB-consistent participants in the LRC condition. The SEQ model predicted 81% and 83% of decisions in the HRC and LRC conditions, respectively, compared to 46 and 63% for TTB and 54 and 36% for RAT.

Figure 5 displays the data from the alternative cost manipulation version of Experiment 2 (see Footnote 2). The smaller sample in this experiment ($N = 24$) led to somewhat noisier data (indicated by the wider shaded area) but again the dominance of the sequential sampling model can be seen in both the High Cost (left panel) and Low Cost (right panel) conditions. In this experiment pure TTB performs slightly better than pure RAT in both conditions reflecting an over-all tendency to adopt more frugal decision making in both conditions. The SEQ model predicted 85% and 80% of decisions in the High Cost and Low Cost conditions, respectively, compared to 54 and 58% for TTB and 46 and 41% for RAT.

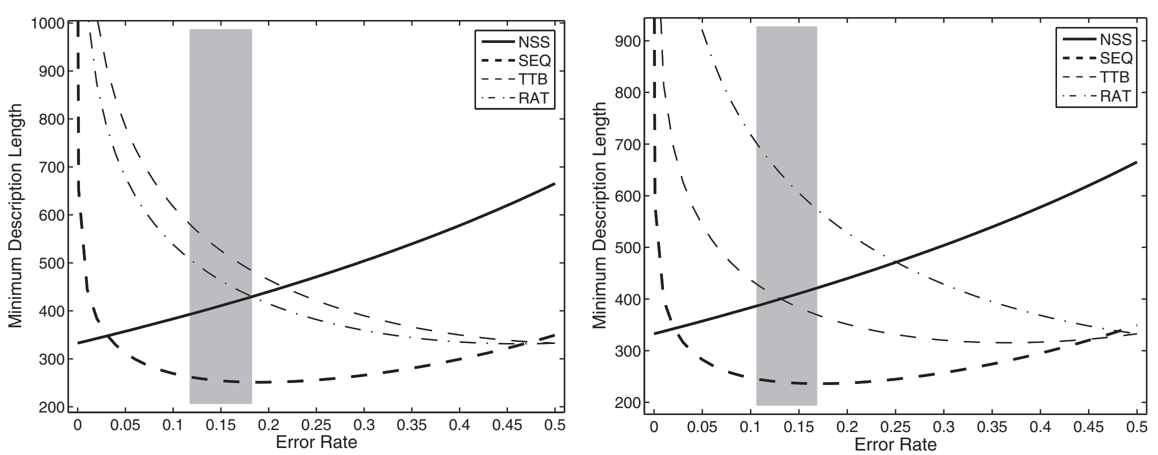


Figure 4. Minimum Description Length (MDL) results for the cost manipulation comparison of Experiment 2. The curves show the MDL criteria for each of the four decision-making models, as a function of the assumed error rate of the experimental data for the high relative cost decisions (left panel) and the low relative cost decisions (right panel). Lower values of MDL indicate a more likely model. In each panel, the shaded box spans the 95% confidence interval for the error rate, estimated from the repeated-measures in the decision data. Naive Strategy Selection Model (NSS); Sequential Sampling Model (SEQ); Take-The-Best (TTB); Rational (RAT)



DOI: 10.1002/bdm

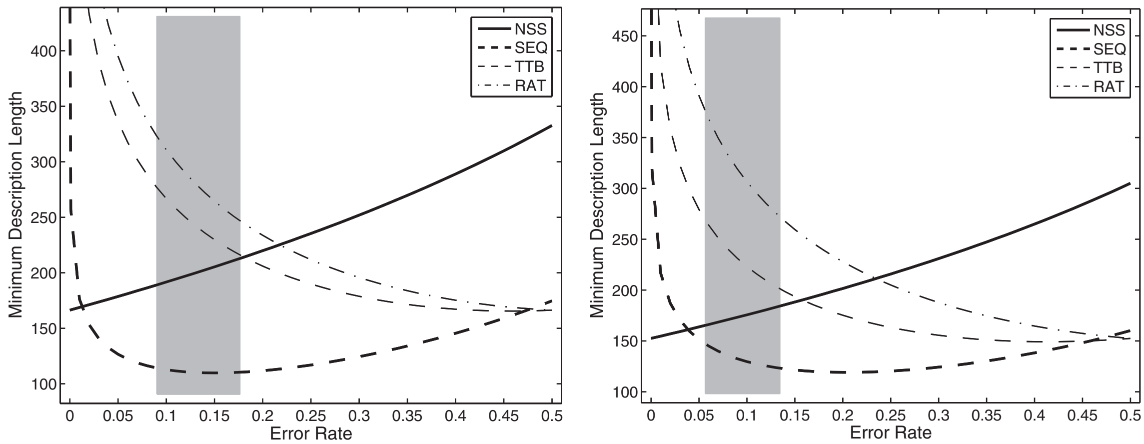


Figure 5. Minimum Description Length (MDL) results for the alternative cost manipulation comparison (see Footnote 2 and text). The curves show the MDL criteria for each of the four decision-making models, as a function of the assumed error rate of the experimental data for the high relative cost decisions (left panel) and the low relative cost decisions (right panel). Lower values of MDL indicate a more likely model. In each panel, the shaded box spans the 95% confidence interval for the error rate, estimated from the repeated-measures in the decision data. Naive Strategy Selection Model (NSS); Sequential Sampling Model (SEQ); Take-The-Best (TTB); Rational (RAT).

**Process data: testing the underlying assumptions of the models**

Table 5 displays the values for several dependent measures for each individual in Experiment 2 (excluding those in the alternative cost manipulation version). To examine search behavior a rank cue purchase order was calculated by giving a cue purchased first a rank of 1, second a rank of 2, third 3 etc., and then summing each rank and dividing by the number of occasions on which that cue was purchased. For example, a person who examined cue 1 (objectively most valid) first, on every trial would have a summed rank of 20 for cue 1 and a mean rank of 1. The mean ranks for cues 1 to 6 respectively were 1.77, 2.28, 2.31, 2.51, 2.40, and 2.99. This means that on average higher validity cues were purchased before lower validity cues. (This pattern held for the majority of participants (see rightmost columns). We also examined the percentage of times out of the 20 opportunities (20 test trials) each cue was picked *first*. This analysis revealed that on average the highest validity cue was picked first on 42 per cent of occasions and that this was significantly higher than the average for any of the remaining cues (smallest $t(47) = 2.80$, $p = .007$). The percentages for cues 2 to 6 were, 15, 18, 9, 9, and 6 per cent, respectively. Both of these analyses lend some support for the modeling assumption that participants examine cues in the order of descending validity.

The assumption that TTB-consistent participants bought fewer cues (overall and discriminating) than RAT-consistent participants was validated in both the HRC (2.17 vs. 2.61 cues, TTB and RAT, respectively) and the LRC conditions (2.67 vs. 3.01cues, for TTB and RAT, respectively—see columns 4 and 5 of Table 5).

**Individual consistency**

The process data provides some validation for the classification method and the assumptions underlying the models considered, however, Table 5 clearly shows that with the exception of four individuals (participants 19HRC, 24HRC, 5 LRC, 24LRC; highlighted in bold) no one used a strategy which was completely consistent with both the process and outcomes predicted by either the RAT or the TTB model. Note that participant 24 HRC adopted a ‘‘take the second best’’ model by only ever looking at the second best cue and then choosing the option to which that cue pointed (the RAT option). This person is thus ‘‘misclassified’’ as



DOI: 10.1002/bdm

Table 5. Individual participant data for dependent measures collected in Experiment 2

| Participant# | Condition | Prop$^{n}$ TTB | Mean cues per trial | Mean discriminating cues per trial | Rank cue purchase order | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | C1 | C2 | C3 | C4 | C5 | C6 |
| 1 | HRC | 0.4 | 1.70 | 1.15 | 1.7 | 1.7 | 1.4 | 1.3 | 1.6 | 1 |
| 2 | HRC | 1 | 1.75 | 1.20 | 1.1 | 1.3 | 1.7 | 1.6 | 1.6 | 2.2 |
| 3 | HRC | 0.15 | 2.25 | 2.05 | 2 | 1 | 2.1 | | 2.4 | |
| 4 | HRC | 0 | 3.70 | 3.20 | 2.1 | 1 | 3.1 | 4.4 | 4.2 | 4.2 |
| 5 | HRC | 0.35 | 1.05 | 1.05 | 1 | 2 | 1 | | | |
| 6 | HRC | 0.9 | 2.20 | 1.95 | 1.5 | 2.7 | 1.8 | 2.5 | 2.7 | 2.5 |
| 7 | HRC | 0.45 | 3.20 | 2.20 | 1.5 | 2.7 | 2.2 | 2.6 | 2.1 | 2.4 |
| 8 | HRC | 0.15 | 3.85 | 3.10 | 1 | 2 | 4.3 | 4.6 | 3.3 | 5.1 |
| 9 | HRC | 0.8 | 2.15 | 1.40 | 1.6 | 1 | | 2 | 1 | |
| 10 | HRC | 0.4 | 2.60 | 2.45 | 1.6 | 1.5 | 2 | 3 | 2.3 | 2.3 |
| 11 | HRC | 0.9 | 3.40 | 2.95 | 1.8 | 2.2 | 2.4 | 1.3 | 3.4 | 2.7 |
| 12 | HRC | 0.15 | 2.80 | 1.80 | 2.8 | | 1 | 2 | 3.3 | 3.5 |
| 13 | HRC | 0.25 | 2.10 | 1.45 | 1.3 | 1.7 | 2.2 | 1.7 | 1.6 | 1.3 |
| 14 | HRC | 0.2 | 2.35 | 1.95 | 2.6 | 1.8 | 1 | 2.3 | 3 | 2.5 |
| 15 | HRC | 1 | 3.15 | 2.30 | 1.5 | 6 | 2.2 | 3.1 | 4 | 2.2 |
| 16 | HRC | 0.5 | 2.65 | 1.70 | 3.5 | 2 | 2 | 2.5 | 1.2 | 2.7 |
| 17 | HRC | 0.2 | 3.60 | 2.40 | 2.5 | 2.5 | 3.2 | 1.1 | 2.5 | 3.7 |
| 18 | HRC | 0.1 | 2.45 | 2.20 | 2 | 2.4 | 1.6 | 1.5 | 1.6 | 2.5 |
| **19** | **HRC** | **1** | **1.00** | **1.00** | **1** | | | | | |
| 20 | HRC | 0.05 | 1.50 | 1.25 | 2.7 | 1.5 | 1.2 | 3 | 2.5 | 5 |
| 21 | HRC | 0.5 | 1.60 | 1.10 | 1.2 | 1 | 1.7 | 1.8 | 1.6 | 1.8 |
| 22 | HRC | 0.8 | 1.90 | 1.55 | 1.9 | 1 | 1 | 2 | 2 | 1.1 |
| 23 | HRC | 0.85 | 1.80 | 1.55 | 1.2 | 1.8 | 2.5 | 2 | 1.6 | 2 |
| **24** | **HRC** | **0** | **1.00** | **1.00** | | **1** | | | | |
| 1 | LRC | 0.7 | 2.00 | 1.40 | 1.2 | 1 | | 1.8 | 1.9 | 2 |
| 2 | LRC | 0.6 | 3.00 | 2.25 | 5 | 3 | 2.8 | 3 | 1.9 | 1.1 |
| 3 | LRC | 0.65 | 2.95 | 2.15 | 1.7 | 2.7 | 2 | 2.7 | 3.2 | 3 |
| 4 | LRC | 0.95 | 2.95 | 2.15 | 1 | 3 | 4 | 2.6 | 2.5 | 2.4 |
| **5** | **LRC** | **1** | **1.05** | **1.00** | **1.1** | | | **1** | | |
| 6 | LRC | 1 | 4.45 | 3.25 | 1.1 | 4 | 4.1 | 4.2 | 2.2 | 3.5 |
| 7 | LRC | 0.35 | 4.35 | 3.35 | 1.1 | 3.1 | 3.6 | 2.9 | 3.9 | 5.4 |
| 8 | LRC | 0.3 | 2.90 | 2.15 | 1.1 | | 2.5 | 2.4 | 2 | |
| 9 | LRC | 0.25 | 4.60 | 3.10 | 2.8 | 3.7 | 3.6 | 1 | 2.2 | 4.5 |
| 10 | LRC | 0.95 | 4.75 | 3.55 | 1 | 4.5 | 2.6 | 2.7 | 4 | 6 |
| 11 | LRC | 0.2 | 3.35 | 2.40 | 2.4 | 2.5 | 3.7 | 2.7 | 1.8 | 3.2 |
| 12 | LRC | 0.1 | 3.50 | 2.65 | | 2.1 | 2.3 | 3.1 | 1.5 | 3.9 |
| 13 | LRC | 1 | 1.40 | 1.30 | 1.1 | 2 | | 3 | 1.5 | 3 |
| 14 | LRC | 0.25 | 3.40 | 2.85 | 2.9 | 2.3 | 1.1 | 4.5 | 3.9 | 4 |
| 15 | LRC | 0.2 | 2.40 | 1.90 | 2.5 | 1.1 | 2.6 | 3 | 2.5 | 2.8 |
| 16 | LRC | 0.8 | 1.85 | 1.65 | 1.6 | | 1.3 | 2 | 2 | |
| 17 | LRC | 0.5 | 1.35 | 1.20 | 1.1 | 1.2 | | 3 | 2 | |
| 18 | LRC | 0.9 | 3.15 | 2.70 | 1.2 | | 2.2 | 2 | 2.8 | 3 |
| 19 | LRC | 0.95 | 3.85 | 3.00 | 2.8 | 2.5 | 3.7 | 2.8 | 1 | 3.3 |
| 20 | LRC | 0.7 | 2.35 | 2.20 | 2.3 | 2 | 1.1 | 2.5 | | 1 |
| 21 | LRC | 0 | 2.80 | 1.90 | 2.5 | 5.3 | 1.1 | 2.2 | 3 | 4.2 |
| 22 | LRC | 0.85 | 3.50 | 3.05 | 1 | 2 | 4 | 4 | 3 | |
| 23 | LRC | 1 | 1.40 | 1.30 | 1 | | 2 | 3 | | 3.5 |
| **24** | **LRC** | **1** | **1.00** | **1.00** | **1** | | | | | |

HRC; High Relative Cost, LRC; Low Relative Cost.



DOI: 10.1002/bdm

RAT user because his/her cue acquisition is clearly frugal. It is these kinds of misclassifications that may have led to the unexpected distribution of RAT and TTB-consistent participants across the two cost conditions[3].

## DISCUSSION

Experiment 2 compared the four models in an environment with explicit monetary search costs. Increasing the cost of information had the predictable effect of curtailing cue search and acquisition. However, this tendency to be frugal in cue acquisition did not lead to the dominance of a ‘‘pure’’ TTB model. Rather, consistent with Experiment 1, the results provided support for the SEQ model in this experiment and in a replication using a slightly different cost manipulation. In both cases the SEQ model provided the best fit to the data, according to the MDL analysis which accounts for goodness-of-fit, model complexity, and the estimated level of noise for the empirical data. The process tracing measures highlighted that classifying participants purely on the basis of choices (outcome) results in some misclassifications (cf. Bergert & Nosofsky, 2007). However, these inconsistencies do not undermine the central claim of the model—that individuals acquire cues, sequentially, up to a particular threshold of evidence before making a decision. This conclusion received support from the *level of terminating evidence* measure which showed greater levels of evidence for RAT than TTB users and greater levels for the lower than the higher cost environment. We acknowledge, however, that these results are necessary though not sufficient evidence for the sequential sampling model.

## GENERAL DISCUSSION

The sequential sampling and accumulation of evidence to a threshold provides an intuitive and simple way to think about how decision makers choose between options. We tested a model (SEQ) premised on this intuition in two experiments using multiple-cue inference tasks. In both experiments evidence was found for a sequential sampling model over models which assumed (a) all participants used a frugal model (TTB), (b) all participants used a compensatory, ‘‘rational’’ model (RAT), and (c) each participant selected a frugal or compensatory model for each decision with some probability (NSS). We do not claim that these experiments provide conclusive evidence of the general superiority of evidence accumulation over strategy selection models. Rather, we conclude that for our specific experimental environment and using the MDL framework for model comparisons, the SEQ model provides the best account of the data.

We acknowledged that the MDL framework may not be optimal for a saturated model like NSS (i.e., one which makes probabilistic predictions). The NSS model fares poorly in the competition because it is too complex not because it cannot account for the data. The complexity lies in the ability of the model to account for a large range of data patterns, some with a higher probability than others, and this flexibility is punished by MDL. One alternative approach would be to make all the models probabilistic by adding an additional parameter which specified the probability with which a person would choose an alternative that was inconsistent with the model prediction (i.e., an application error parameter). A further alternative would be to make the NSS model deterministic and assume that one of the two strategies—TTB or RAT—was always selected with probability 1.0. Model

[3]The predictions of a TTB model in which Cue 1 and Cue 2 (or Cue 1 and Cue 3) are reversed are identical to that of the RAT model given the particular test items used in Experiments 1 and 2 (see Table 2). This limitation can lead to the kinds of misclassifications identified in the text. These test items were chosen because they are the only ones that distinguish TTB and RAT under *correct* cue ordering for this stimulus and cue structure (e.g., Bergert & Nosofsky, 2007; Lee & Cummins, 2004), but clearly their use entails limitations which qualify our conclusions. The idea of imperfect cue ordering over subjects, or over trials, or both, is a theoretically important one, which we address to some extent with the process-tracing measures, but more work along the lines of that outlined in the general discussion is necessary.



DOI: 10.1002/bdm

comparisons using these methods might lead to a less clear-cut advantage for SEQ. Nonetheless, in this "first-pass" comparison, in our experimental environment and using the MDL framework the SEQ model fares best.

Whether more sophisticated selection models, such as the SSL model, that can be reinforced by accuracy and/or effort (see Rieskamp & Otto, 2006) could provide equally good accounts of these kinds of data remains open to future investigation. The NSS model relied only on accuracy for reinforcement but still fared better than either the pure TTB or pure RAT models in all of our comparisons for the estimated level of error in our data (see Figures 3–5). Thus it is reasonable to assume that augmenting the model with an effort reinforcement mechanism, like SSL, would improve its performance. Designing environments in which selection models and evidence accumulation models make different predictions (unlike the current environment) is clearly an important next step in this research program (cf., Newell & Lee, 2009).

### Effects of the experimental manipulations

In Experiment 1 the format in which cue information was presented (text or image) had no systematic effect on the evidence threshold adopted, suggesting that such format effects are restricted to memory-based tasks (e.g., Bröder & Schiffer, 2003); a result corroborated by earlier findings (e.g., Bergert & Nosofsky, 2007; Juslin, Olsson, & Olsson, 2003). In Experiment 2 increasing the cost of cue information had the predictable effect of reducing cue purchase (Bröder, 2003; Newell & Shanks, 2003), but also demonstrated that the point at which participants made a decision—their *terminating level of evidence* (as measured by the log-odds of the cues)—decreased with increasing information cost. The cost effects were marginal, suggesting over-all reluctance on the part of our (inevitably impecunious) student participants to spend money; but the general pattern was consistent with the central evidence-accumulation notion of the SEQ model. However, other process tracing measures, such as cue-search revealed limitations of relying solely on choices as a basis for classifying and modeling behavior (cf. Bergert & Nosofsky, 2007).

### Limitations of the model: cue search

An important aspect of claiming support for the SEQ model is verifying that cue information is searched in cue-validity order (e.g., Figure 1).On the average cue search in Experiment 2 showed sensitivity to cue-validity: for most individuals there was an increase in purchase-order rank as validity decreased, and on average the most valid cue was picked first significantly more often than any of the other cues. Nonetheless, the analyses of the search data revealed inconsistencies that suggest caution in drawing strong conclusions (see Footnote 3).

Many researchers have acknowledged and demonstrated the difficulty that participants exhibit in learning cue-validities (Bergert & Nosofsky, 2007; Dieckmann & Rieskamp, 2007; Newell & Shanks, 2003; Newell, Rakow, Weston, & Shanks, 2004; Rakow, Newell, Fayers, & Hersby, 2005; Todd & Dieckmann, 2005) and have dealt with it in different ways. In experiments the problem is often surmounted by simply providing validities to participants (e.g., Dieckmann & Rieskamp, 2007; Rieskamp & Otto, 2006); while this might be expedient it rather detracts from the goal of the modeling and experimental work (especially when the provided validities are not veridical—see Rieskamp & Otto, 2006, Study 1). An alternative way to deal with deviation from deterministic search orders is to convert the models under consideration into probabilistic ones, by adding cognitive processes and free parameters to accommodate idiosyncratic cue-weightings. This is a standard and reasonable approach, and is similar to the one taken by Bergert and Nosofsky (2007) in their examination of RAT and TTB-consistent decision making.

Such an approach lies at the opposite end of the spectrum to ours (in which we preserve the deterministic nature of the models) and, we suggest, has its own disadvantages. With regard to the cue search issue, the free parameter approach allows for the possibility that participants learn *nothing* about cue validities (or indeed learn a completely reversed order) during training; an assumption which is clearly violated by our search



DOI: 10.1002/bdm

order data. To address the problem of how to model the cue-validity learning and search process adequately would require a systematic exploration of models inhabiting the space between a completely deterministic and a completely probabilistic model. Unfortunately such an endeavor is beyond the scope of the current paper. We note however, that our conclusions are not necessarily at odds with those drawn by researchers who have adopted the heavily parameterized modeling approach. Bergert and Nosofsky (2007) concluded that the majority of participants in their experiments used a ''generalized'' form of TTB (formally equivalent to Tversky's (1972) Elimination by Aspects model) but they did not attempt to fit a ''unifying'' sequential sampling model of the type we advocate. They did however note that ''such a model has promise for a rigorous joint account'' (p. 127) of performance in multi-attribute inference tasks. Thus it remains possible that the behavior of participants in Bergert and Nosofsky's (2007) studies and that of the participants in the current experiments is best captured by some version of the sequential sampling model. Future research should focus on the exact specifications of such a model, with particular emphasis on how cue validities are learned.

### Spanners and toolboxes

A limitation of our experiments is that we only used one statistical environment. It is possible that better evidence for genuine switches between qualitatively different strategies would be observed in richer choice environments with multiple options and attributes. Indeed one of the reasons an evidence accumulation process fares well in capturing our participants' decisions is because inference patterns in these experiments are discriminated only by the number of cues people use—few (TTB), many (RAT or WADD).

Payne et al. (1993) in their work on preferences have documented many situations in which strategies differ not just in the amount of information used but in the *order* in which information is sought. Perhaps the clearest example is the difference between an attribute (cue)-wise and an alternative-wise search. The evidence-accumulation model that we have formulated has an attribute-wise search rule (as shown in Figure 1, cues for each alternative are consulted consecutively); but there are cases in which people adopt alternative-wise searches (examining *all* the cues for one alternative then the other). Evidence accumulation models like the one we consider currently have no way to accommodate such search patterns.

It should be noted, however, that although there is a good deal of evidence for the use of different strategies in different conditions (e.g., Beach & Mitchell, 1978; Payne et al., 1993), the exact nature of the deliberation process (''deciding how to decide'') has been neglected—or at least not clearly explained in many of these frameworks (see Bröder & Newell, 2008, for further discussion of this issue).

### One hard problem replaced by another?

A potential criticism of the sequential sampling conception of adaptive decision-making is that it replaces a hard problem of how people choose which heuristic to apply with an equally hard problem of how people set the appropriate threshold level of evidence. We agree that the self-regulation of a threshold is a challenging and important problem, but believe that it is a much simpler and much better-defined problem, than that of choosing between a set of heuristics. To begin with, for a heuristic approach one needs to be able to set constraints on how large the set or ''toolbox'' can be (cf., Dougherty, Thomas, & Franco-Watkins, 2008; Newell, 2005). In addition to this fundamental issue of constraints, the over-arching structure of accounts based on sequential sampling means attempts to model self-regulation are perhaps more likely to see significant progress, in terms of both general theories and concrete models.

Indeed, the literature on sequential sampling models already has one promising candidate approach to self-regulation. Vickers (1979; see also Vickers and Lee, 1998) proposed and evaluated a general account



DOI: 10.1002/bdm

of how the psychological measure of confidence can act as a regulatory variable to control decision-making through the adjustment of threshold levels of evidence (see also Hausmann & Läge, 2008 for a similar conception). For the most part, Vickers' theory of self-regulating accumulator (SRA) models has been evaluated on relatively low-level perceptual decision-making problems, where it has been shown to be able to account for an impressive array of phenomena, including lags in adaptive responding to step-changes in stimulus environments, and hysteresis effects in tracking non-stationary stimulus environments. With the exception of Lee and Dry (2006), however, these self-regulating sequential sampling models have not been applied to the sorts of cognitive decision-making tasks that we have considered in this paper. The extension seems straight-forward and provides an extremely promising direction for future research.

## Conclusion

Analyses of decision making in multiple-cue inference tasks has been dominated by perspectives which assume that people select from a repertoire of cognitive strategies. The modeling and experimental results reported here, demonstrate that an alternative perspective based on sequential sampling and evidence accumulation might provide a better account of performance than selection between strategies.

## ACKNOWLEDGEMENTS

The assistance of the Australian Research Council is gratefully acknowledged (DP 0551818). The authors thank Elia Vecellio, Patrick Collins, and Anke Häbeck for assistance with stimulus construction and data collection. The authors are very grateful to Rob Nosofsky and Jörg Rieskamp, Oren Griffiths, and Tim Rakow for extremely useful comments on an earlier version of this paper. Portions of the data from Experiment 1 were published in the *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

## REFERENCES

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*, 439–449.

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, and Cognition*, *33*, 107–129.

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., & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multi-attribute decisions. *Psychonomic Bulletin and Review*, *14*, 895–900.

Bröder, A., & Newell, B. R. (2008). Challenging some common beliefs about cognitive costs: Empirical work within the adaptive toolbox metaphor. *Judgment and Decision Making 3*, 205–214.

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.

Bröder, A., & Schiffer, S. (2006). Stimulus format and working memory in fast and frugal strategy selection. *Journal of Behavioral Decision Making*, *19*, 361–380.

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. J. Koehler , & N. Harvey (Eds.), *The Blackwell handbook of judgment and decision making* (pp. 133–154). Malden, MA: Blackwell.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*, 91–134.



DOI: 10.1002/bdm

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459.

Christensen-Szalanski, J. J. J. (1978). Problem solving strategies: A selection mechanism, some implications and some data. *Organizational Behavior and Human Performance*, *22*, 307–323.

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

Dieckmann, A. & Rieskamp, J. (2007). The influence of information redundancy on probabilistic inferences. *Memory & Cognition*, *35*, 1801–1813.

Dougherty, M. R., Thomas, R., & Franco-Watkins, A. M. (2008). Postscript: Vague heuristics revisited. *Psychological Review*, *115*, 211–213.

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). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer , P. M. Todd , & the ABC Research Group, (Eds.), *Simple heuristics that make us smart* (pp. 3–34). New York: Oxford University Press.

Grünwald, P. (1999). Viewing all models as ‘‘probabilistic’’. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT '99)* (pp. 171–182). New York: ACM Press.

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT 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.

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple- cue judgment. *Journal of Experimental Psychology: General*, *132*, 133–156.

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

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and Value Trade-offs*. New York: Wiley.

Laming, D. R. J. (1968). Information *Theory and choice-reaction time*. London: Academic Press.

Lee, M. D. (2004). An efficient method for the minimum description length evaluation of cognitive models. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 807–812). Mahwah, NJ: Erlbaum.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555–580.

Lee, M. D., & Corlett, E. Y. (2003). Sequential sampling models of human text classification. *Cognitive Science*, *27*, 159–193.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the ‘‘take the best’’ and the ‘‘rational’’ models. *Psychonomic Bulletin and Review*, *11*, 343–352.

Lee, M. D., & Dry, M. J. (2006). Decision making and confidence given uncertain advice. *Cognitive Science*, *30*, 1081–1095.

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

Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). *Straight choices: The psychology of decision making*. Hove, UK: Psychology Press.

Newell, B. R., & Lee, M. D. (2009). Learning to adapt evidence thresholds in decision making. In: N. Taatgen, H. van Rijn, J. Nerbonne and L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp 473–478) Austin, TX: Cognitive Science Society.

Newell, B. R., Rakow, T., Weston, N. J., & Shanks, D. R. (2004). Search strategies in decision-making: The success of success. *Journal of Behavioral Decision Making*, *17*, 117–137.

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, 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 ‘‘takes-the-best’’. *Organizational Behavior and Human Decision Processes*, *91*, 82–96.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

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

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.

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



DOI: 10.1002/bdm

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *32*, 1335–1370.
Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236.
Simon, H. A. (1956). Rational choice and the structure of environments. *Psychological Review*, *63*, 129–138.
Todd, P. M., & Dieckmann, A. (2005). Heuristics for ordering cue-search in decision making. In L. K. Saul , Y. Weiss , & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, pp. 1393–1400). Cambridge, MA: MIT Press.
Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299.
Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.
Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of self-regulating accumulator module. *Non-Linear Dynamics, Psychology, and Life Sciences*, *2*, 189–194.
Wallsten, T. S., & Barton, C. (1982). Processing probabilistic multidimensional information for decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 361–384.

*Authors' biographies*:

**Ben R. Newell** is a senior lecturer in the School of Psychology at the University of New South Wales. His research interests include judgment and decision-making, with particular interest in the cognitive mechanisms underlying multiple cue judgment, and the implicit/explicit distinction in learning, memory, categorization, and decision making.

**Michael D. Lee** is professor of Cognitive Sciences at the University of California, Irvine. His research interests are in computational and mathematical models of higher-order cognition, including representation, memory, learning, decision-making, and problem-solving.

*Authors' addresses:*

**Ben R. Newell**, School of Psychology, University of New South Wales, Sydney 2052, Australia.

**Michael D. Lee**, Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100, USA.



DOI: 10.1002/bdm