An Empirical Evaluation of Models for How People Learn Cue Search Orders

Percy K. Mistry (pkmistry@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California Irvine, Irvine, CA 92697-5100 USA

Ben R. Newell (ben.newell@unsw.edu.au)
School of Psychology, University of New South Wales, Sydney 2052, Australia

Abstract

We propose simple parameter-free models that predict how people learn environmental cue contingencies, use this information to measure the usefulness of cues, and in turn, use these measures to construct search orders. To develop the models, we consider a total of 8 previously proposed cue measures, based on cue validity and discriminability, and develop simple Bayesian and biased-Bayesian learning mechanisms for inferring these measures from experience. We evaluate the model predictions against people's search behavior in an experiment in which people could freely search cues for information to decide between two stimuli. Our results show that people's behavior is best predicted by models relying on cue measures maximizing short-term accuracy, rather than long-term exploration, and using the biased learning mechanism that increases the certainty of inferences about cue properties, but does not necessarily learn true environmental contingencies.

Keywords: learning; search order; predictive models; cue contingencies

Introduction

Making a decision requires people to search for information, decide when to terminate that search, and then make a decision based on the available information (Gigerenzer, Todd, & the ABC Group, 1999). Deciding which of two cities is larger might start with finding out whether each city is a state or national capital, whether it has an airport, and so on. At some point, the information gathering must stop, and a decision made on the basis of what is known about the available cues and their relationship to the decision criterion of population size. If people receive some sort of feedback—whether implicit or explicit, or immediate or delayed—about the accuracy of their decisions, then it also becomes possible to learn the usefulness of different cues. The field of decision making is full of models for learning how cues relate to criteria, based on principles like conditioning, reinforcement, and error correction. There are some models of how people learn when to terminate search, usually in the form of adaptive sequential sampling models, and based on principles like maximizing reward rates, controlling conflict, or maintaining confidence (Lee, Newell, & Vandekerckhove, 2014).

There are fewer models of how people learn the order in which to search. Many measures have been proposed as the basis for ordering search, including those that focus on immediate benefits like the current validity or success rate of a cue (Gigerenzer & Goldstein, 1999; Newell, Rakow, Weston, & Shanks, 2004), and those that take a longer view by focusing on information gain (Nelson, 2005). Central to calculating all of these measures are the *discriminability* and *validity* properties of a cue. Discriminability is the probability that a cue takes different values for two stimuli being compared. Validity is the probability that it identifies the correct stimulus, given that it discriminates. Despite their centrality, there are few models of how validity and discriminability are learned, and it is often simply assumed they are veridically available to people. This means, in turn, that there are few models of how people learn to order search. Exceptions are Todd and Dieckmann (2004) and Martignon and Hoffrage (2002). These, however, focus on simulation studies and lexicographic rules with one-reason decision making, which are not easily extended to cases where people search beyond one discriminating cue.

A process account of how people learn these cue contingencies and decide on search orders is relevant to any model of choice that employs sequential sampling and evidence accumulation, as well as heuristics that select cues based on learned contingencies. In this paper, we develop a modeling framework that allows for different assumptions about what cue measures are important for guiding search, and can use one of two simple learning mechanisms for ordering cue search. We evaluate the resultant 16 different models against previous experimental data measuring how people search.

Experimental Data

Our data come from experiments reported by van Ravenzwaaij, Newell, Moore, and Lee (2014) in which, on each trial, participants had to select which of two cities had a larger population. The names of the cities were not provided, but various cues—such as whether the city had an airport, a university, a sports team, and so on—were available for both cities. Participants had the option to select and view as many cues as they liked, in an order of their choosing. The cues were visually presented in a circular layout, with a random ordering for each participant, to control for presentation effects on the order of search. There were two experimental conditions: in the *known* condition, cue validities and discriminabilities were provided, while in the *unknown* condition, this information was not provided. In both conditions, participants received corrective feedback after every trial.

We focus on the *unknown* condition, within which $n = 24$ participants completed two environments: 100 trials each for Italian (9 binary cues) and USA (8 binary cues) cities. On average, about four cues are used on every trial, although there is large variability between trials and participants. Interestingly, the use of individual cues is very similar between the *known* and *unknown* conditions, with a correlation of 0.94.

Modeling Assumptions

All of the models we consider come from combining a measure of cue usefulness with a method for learning that measure, based on feedback over a sequence of trials. Search orders are determined by sampling from the learned distributions of the measure for each cue.

Measures of cue usefulness

The first four measures focus on immediate reward, and are simple to define. The remaining measures involve more detailed calculations, and we provide only the intuition behind these measures (see Nelson, 2005, for details).

Validity The probability v a cue identifies the correct choice, given that it discriminates between the two stimuli. It is the basis for search in the prominent fast and frugal heuristic known as take-the-best (Gigerenzer & Goldstein, 1996).

Discriminability The probability d that a cue takes different values for the two stimuli. Searching by discriminability is an extreme case of the linear family of measures considered by Lee and Newell (2011), Lee and Zhang (2012), and Ravenzwaaij, Moore, Lee, and Newell (2014).

Additive An average of cue validity and discriminability, $\frac{1}{2}(v+d)$. Searching using this average is a special case of the linear family of measures considered by Lee and Newell (2011) and Lee and Zhang (2012).

Success rate (SR) This is defined as $dv + \frac{1}{2}(1-d)$ by Newell et al. (2004). It measures the probability of making the correct choice by combining the probability of the cue discriminating and leading to a correct decision, with the probability of guessing correctly if it does not discriminate.

Information gain A measure of the expected reduction in uncertainty— the change in entropy of the choice options— from observing the value of the cue for the two stimuli.

Probability gain A measure of increase in the expected probability of making a correct guess.

Impact A measure of the average absolute change in the probabilities of each choice being correct, as a result of observing the value of a cue for the two stimuli, weighted by the probability of the cue providing this information.

Bayesian diagnosticity A measure of the expected weight of evidence of the cue measured in terms of likelihood ratios.

Learning mechanisms

We propose that people implicitly keep track of which cues are searched, and the success of each in discriminating, and indicating the correct choice (Lagnado, Newell, Kahan, & Shanks, 2006). Formally, after t trials, we assume people know they have searched a cue on γ_t trials, that it has discriminated α_t times, and indicated the correct choice β_t times.

Standard Bayesian Learning Given this information, a straightforward way to learn cue v and d is through Bayesian belief updating. We make the simplifying assumption that people have no strong prior beliefs (an assumption we revisit in the discussion), so that initially $v_0 \sim \text{beta}(1,1)$ and $d_0 \sim \text{beta}(1,1)$. After t trials, this means that;

$$\begin{aligned} v_t &\sim \text{beta}(1+\beta_t, 1+\alpha_t-\beta_t) \\ d_t &\sim \text{beta}(1+\alpha_t, 1+\gamma_t-\alpha_t). \end{aligned}$$

Biased Bayesian learning An alternative to standard Bayesian learning is motivated by the idea of confirmation bias, or positive test strategy, whereby people are prejudiced towards aspects that have previously produced positive results (Klayman & Ha, 1987), and the idea of selective attention, whereby people tend to focus on a limited set of attributes (Wilson & Niv, 2011). Le Pelley, Beesley, and Griffiths (2011) reported results that cues with a high level of predictive power resulted in a higher attentional bias. Beesley, Nguyen, Pearson, and Le Pelley (2015) suggested that attention bias towards the exploitation of predictive cues was more robust than an attention bias towards exploratory behavior arising from increasing uncertainty about cues. To account for these sorts of biases, we consider a second learning mechanism in which the cues that are not searched on a trial are assumed to have failed. This is contrary to standard Bayesian belief updating, which only considers information about validity and discriminability from cues actually searched. Formally, this means;

$$\begin{aligned} v'_t &\sim \text{beta}(1+\beta_t, 1+t-\beta_t) \\ d'_t &\sim \text{beta}(1+\alpha_t, 1+t-\alpha_t). \end{aligned}$$

Intuitively, biased Bayesian learning increases the tendency to persist with cues that have been successfully tested in previous trials, creating an attentional or confirmational bias for exploitation over exploration.

Determining search orders

Given a cue measure and learning mechanism, we propose that people determine a search on each trial by sampling from the learned distribution of the measure for each cue. The rank order of these samples determines the search order for that trial. Since all the cue measures we consider are deterministic functions of v_t and d_t, the learning results for these two measures presented above allow any measure to be sampled.

Model Demonstration

We illustrate our modeling framework with a simple example involving three cues. Table 1 shows the counts after $t = 30$ trials, and the corresponding validity v and discriminability d for each cue. On each trial, the order of search is determined by drawing a mental sample from the inferred distribution of the cue measure, which incorporates uncertainty. These illustrative distributions are shown in Figure 1, with the two left-most panels considering the validity measure, and the two right-most panels considering the success rate measure. For the standard learning model, uncertainty is reduced as the cue

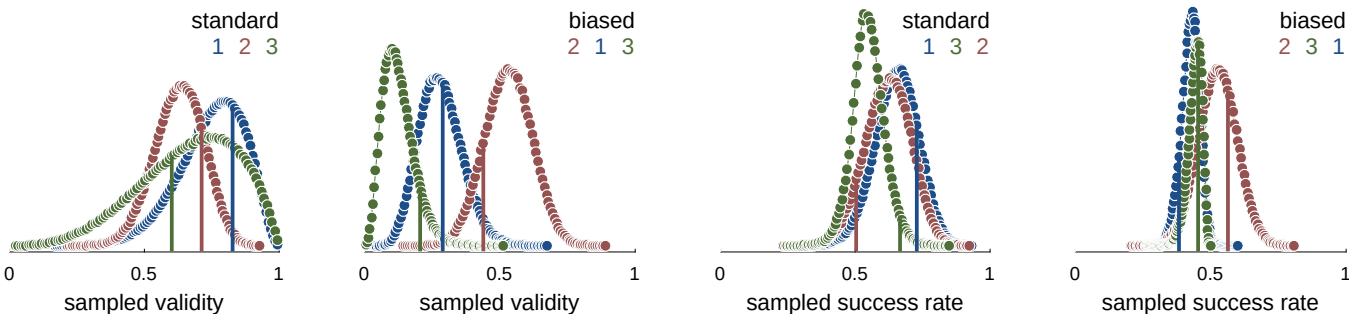

Figure 1: Demonstrations of determining search orders from learned inferences about cue measures. The two left panels relate to the validity measure, and the two right panels relate to the success rate measure. In each panel, the inferred distribution for three cues is shown, and the vertical lines indicate a sampled value that determines search order for a trial. For both the validity and success rate panel pairs, the left-most panel relates to standard learning, and the right-most panel relates to biased learning.

is sampled more often, leading to search orders more deterministically following the empirical estimates. Beliefs about v are updated only if the cue discriminates, which means the standard learning model will always result in lower certainty for v, compared to d, unless the cue always discriminates. In the biased learning model, uncertainty is reduced for both v and d on every trial, regardless of whether the cue is searched.

For this example, if search order deterministically followed the empirical rates, the validity measure would predict the cue order $[1,3,2]$, the discriminability measure would predict $[2,1,3]$, and the success rate measure would predict the order $[1,2,3]$. The other measures of usefulness might predict still different orders. The uncertainty associated with mental sampling means, however, that a distribution of search orders is predicted by each measure. Figure 1 illustrates the predicted search orders using samples from the distributions, shown by vertical lines. For the particular set of samples shown in Figure 1, the predicted search order is $[1,2,3]$ for standard learning using validity, $[2,1,3]$ for biased learning using validity, $[1,3,2]$ for standard learning using success rate, and $[2,3,1]$ for biased learning using success rate.

Table 1: Example situation after 30 trials, giving γ_{30} (number of times cue is searched), α_{30} (number of times cue is searched and discriminates), and β_{30} (number of times cue is searched, discriminates and is valid) counts for three cues.

Cue	γ_{30}	α_{30}	β_{30}	$d_{30} = \alpha_{30}/\gamma_{30}$	$v_{30} = \beta_{30}/\alpha_{30}$
1	16	10	8	0.63	0.80
2	26	25	16	0.96	0.64
3	15	4	3	0.27	0.75

Model evaluation

Generating model predictions

Because the model predictions depend on the mental samples drawn on each trial, they are inherently probabilistic. Accordingly, we generate 100 samples for each trial for each participant and measure. The predictions about cue order on each individual trial are made without the model making contact with behavioral data. The models are genuinely parameter-free, so there is no model fitting or parameter estimation involved, and thus no need to adjust for model complexity.

Evaluating model performance

A search order that involves different subsets of 8 or 9 different cues has many possible combinations (over 100,000 with 8 cues and over 900,000 with 9 cues). The actual set of unique search combinations used is fewer than 1% of these. We use a partial tau τ as a metric for the difference between observed and predicted search orders. This is a generalized version of Kendall's tau metric, and is a standard metric in statistics for the difference between two partially-ordered lists (Fagin, Kumar, Mahdian, Sivakumar, & Vee, 2006). Intuitively, τ is the number of pairwise swaps required to transform one search order into another, allowing for ties. Thus $\tau = 0$ when the observed order exactly matches the predicted order, but increases as the observed order becomes more different. The generalization of τ to the partial version we use allows rankings to have ties. This is important, because whenever a participant terminates search before examining all cues, they produce a partial order in which all of the non-searched cues can be considered as being ranked equal last.

Group-level results

We identified the cue measures, for each learning mechanism separately, that provided the best prediction (i.e. lowest τ) for each individual participant and on each trial. We then calculated $\Delta\tau$, the increase in τ of each model over this minimum τ value. The distribution of $\Delta\tau$ is shown in Figure 2. Blue (darker) lines and markers show the standard learning models, and the red (lighter) markers and lines the biased learning models. Since higher mass of the $\Delta\tau$ distribution closer to zero indicates better model performance, it is clear that validity, discriminability, additive, and success rate models, when combined with biased learning, are far better in predicting cue search orders. A series of paired sample Bayesian t-tests with

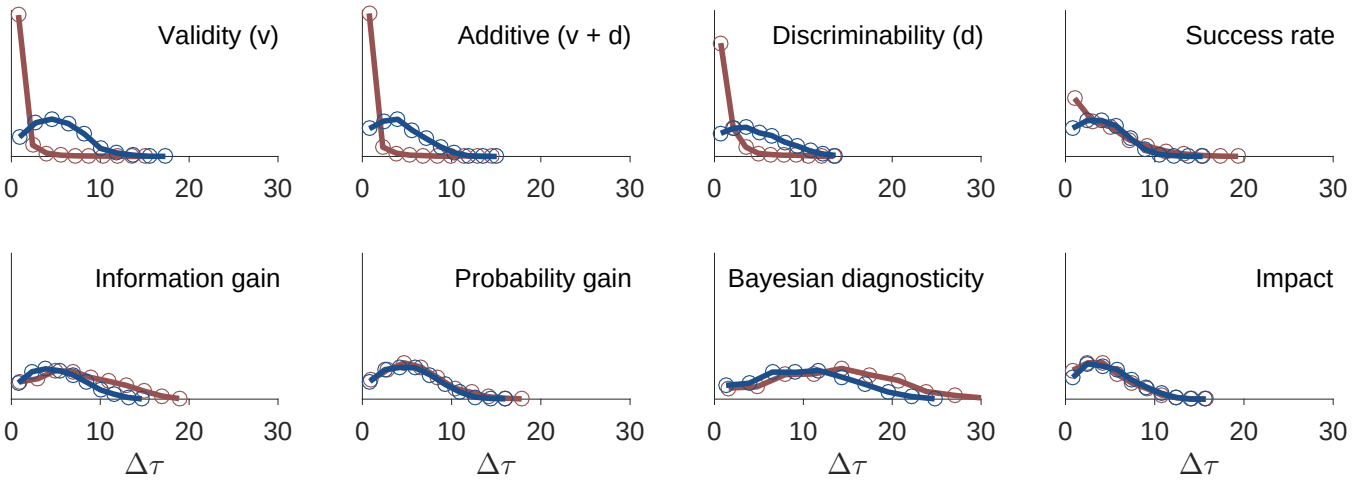

Figure 2: Distribution of $\Delta\tau$ across all trials and participants. Standard (blue) and biased (red) learning shown for each measure

default priors (Love et al., 2015) were carried out to test the one-sided alternative hypothesis that the population mean of partial taus generated by the biased-additive model is lower than for each of the remaining models. The log Bayes factors generated ranged from 6 to 28 for all pairwise comparisons for both USA and Italy datasets, except for the comparison with the biased-validity model, for which the Bayes factors were inconclusive. This provides evidence that the biased-additive and biased-validity models are significantly better than the others models with respect to the partial tau measures. Figure 3 shows τ across trials for the top four measures in the USA condition. The blue (darker) and red (lighter) circles show mean τ for standard and biased learning models, respectively. The error bars show the 95% credible interval across all participants and samples. The black line shows the mean for random sequences. In general, τ reduces across trials, suggesting behavior gradually becomes consistent with systematically predicted cue orders, either because the model becomes more accurate, people become more consistent, or both.

Individual-level results

The error bars in Figure 3 are large because of individual difference between participants. At the individual level, more confident evaluation is possible. Figure 4 demonstrates this, by showing the same analysis for an individual participant. This participant's search orders are best predicted by biased learning, and by the validity, discriminability, or additive combination of these two measures. It is not possible to display the same analysis for all conditions and participants, but the one in Figure 4 was chosen as prototypical. Results for all conditions (including Italian cities) and participants, are available as supplementary material at `www.osf.io/uqf5p`.

Effectiveness of cue learning

The prediction of cue search orders and evaluation of measures of usefulness depends on the learned v and d. Our models predict these for each individual cue at each trial. Figure 5 shows the predicted learning for 3 of the 8 different cues in the USA condition for a single participant. Cue 2 is whether the city has a sports team, cue 3 is whether the city has an airport, and cue 7 is whether the city is a national capital. The large circles show the mean learned values, that is, calculations based on counts of successes and failures, without taking into account uncertainty. The gray dots represent the set of 100 mental samples drawn from learned distribution, and make clear the associated uncertainty, which reduces across trials. For standard learning, the uncertainty reduces earlier for d than v, as expected. This results in greater difficulty in learning v for cues with empirically lower d (e.g., cue 7). The crosses at the top of each plot show the trials on which the cue was searched. As expected, cues accessed frequently show a greater reduction in uncertainty as well as higher accuracy (e.g. compare cue 3 to cue 7). Contingencies for cue 3 are learned accurately, as the cue is repeatedly selected.

A series of paired sample Bayesian t-tests with default priors to test the one-sided alternative hypothesis that the standard deviation of generated samples, across all cues, for the last trial for both the USA and the Italy environments, were lower for discriminability compared to validity revealed log Bayes factors in the range of 6 to 34; and lower for the biased compared to the standard model revealed log Bayes factors in the range of 15 to 26. This provides evidence that the uncertainty over discriminability is significantly lower than that about validity, and the uncertainty for both are significantly lower for the biased compared to the standard model. While the biased model shows lower uncertainty, cues not frequently accessed under this model are not learned accurately. For example, accuracy of learned v for cue 7 is less than cue 2 which in turn, is less than cue 3. Most cue validities and discriminabilities are learned quite effectively and quickly. For example, cue 3 is accurately assessed in the biased model by about the 10th trial.

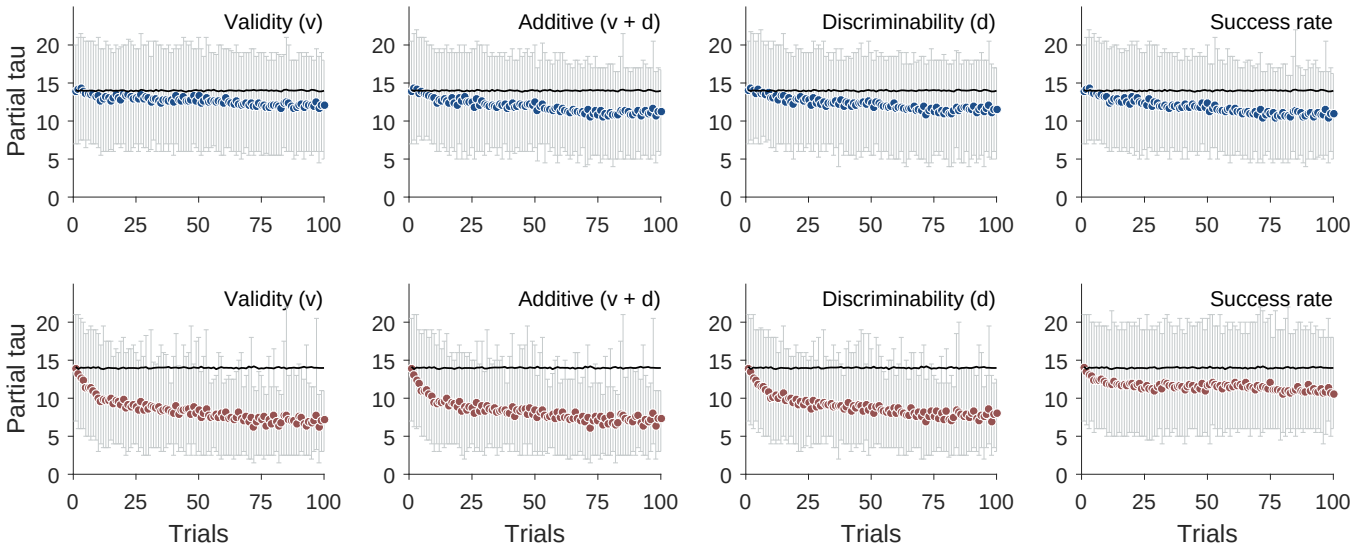

Figure 3: Each panel shows the mean and 95% interval for τ over all participants (USA sub-condition) for different cue measures. Top row: standard learning; Bottom row: biased learning. The black line is mean τ for random orders.

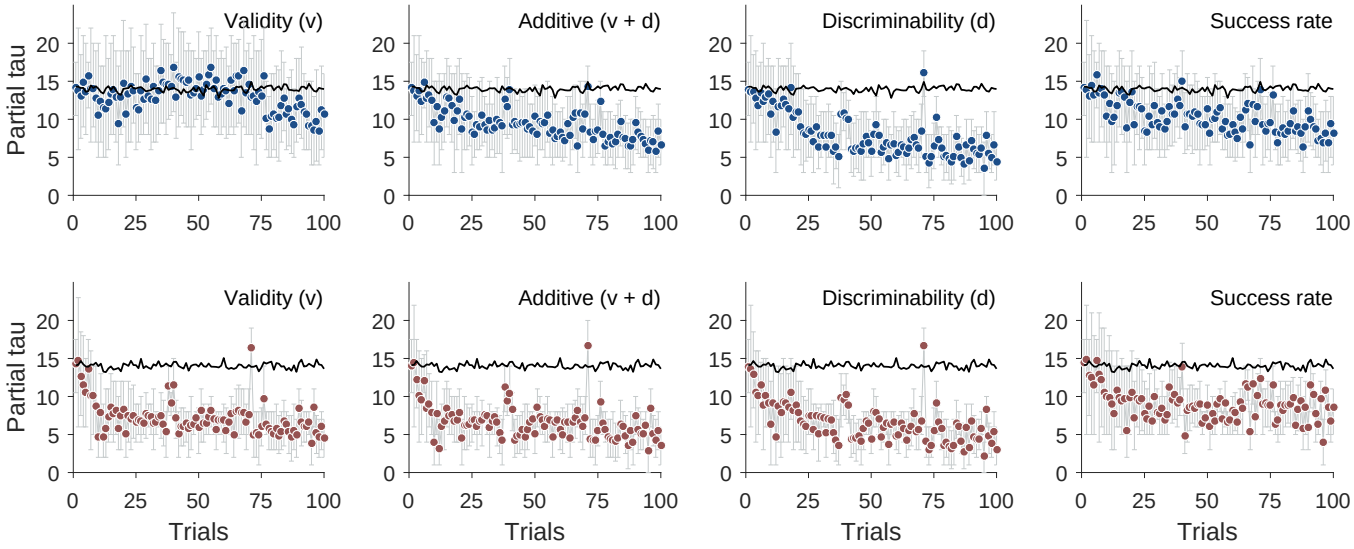

Figure 4: Each panel shows the mean and 95% interval for τ for a single participant (USA sub-condition) for different cue measures. Top row: standard learning; Bottom row: biased learning. The black line is mean τ for random orders.

Discussion

We have shown that simple parameter-free learning mechanisms make reasonable predictions about people's cue search orders. Our two key results are evidence for biased learning, and the demonstration that simple validity and discriminability (or additive combinations of them), make better predictors of cue search orders than more sophisticated measures of cue usefulness. We did, however, find that there were individual differences in use of the various measures. In future work, we propose examining parameterized models—such as a generalization of the additive model into a linear weighted model $wv + (1-w)d$, with w as a free parameter—to capture some of these differences.

Differences may also arise from memory, discounting phenomena, or sensitivity to cost and effort. All of these can be incorporated into extended learning mechanisms. For example, recency effects can be incorporated by using a decay rate for the counts, and cost sensitivity by appropriately weighting the process by which counts are updated. Finally, the biased learning model suggests under-exploration, but this could be on account of strong causal priors that people may have regarding the various cue attributes. Prior causal beliefs generated outside experimental settings can be difficult to measure, although appropriate parameterization of models could be used to infer such prior beliefs and improve the quality of predictions.

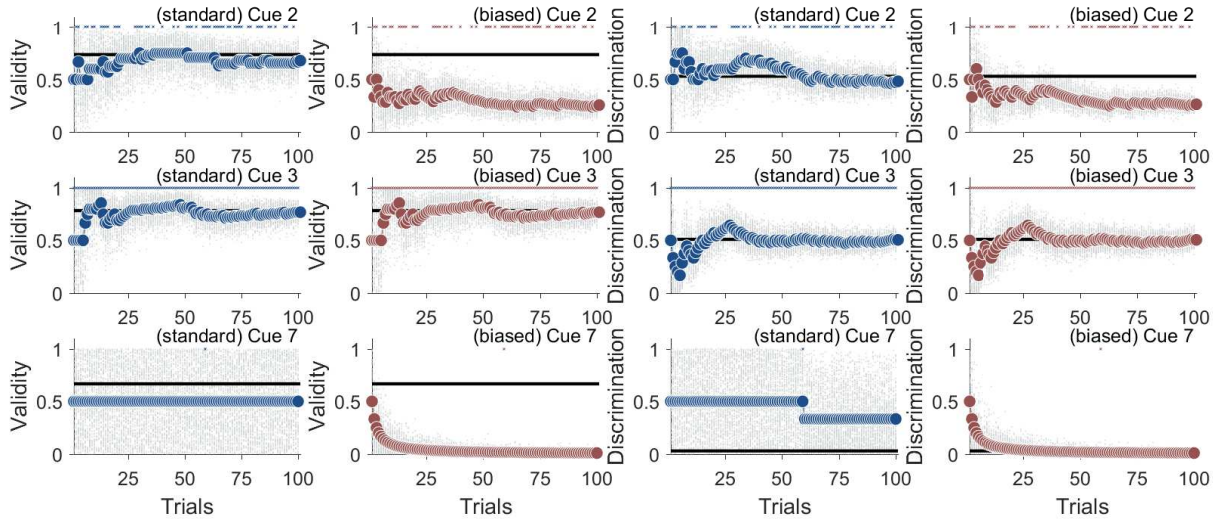

Figure 5: Learning predictions for a single participant, for 3 of the 8 cues in the USA environment. The first two columns show validities, and the last two columns show discriminabilities. Blue circles show standard learning; Red circles show biased learning. The thick black lines are the true values. Gray dots show the mental samples drawn from the learned distributions. The crosses at the top show the trials on which the cue was searched.

References

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *The Quarterly Journal of Experimental Psychology*, 1–25.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2006). Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, *20*, 628–648.

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

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart* (pp. 75–95). Oxford University Press.

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

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211.

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*, 162.

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., Newell, B. R., & Vandekerckhove, J. (2014). Modeling the adaptation of search termination in human decision making. *Decision*, *1*, 223.

Lee, M. D., & Zhang, S. (2012). Evaluating the process coherence of take-the-best in structured environments. *Judgment and Decision Making*, *7*, 360–372.

Le Pelley, M., Beesley, T., & Griffiths, O. (2011). Overt attention and predictiveness in human contingency learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*, 220.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A., & Wagenmakers, E. (2015). *Jasp (version 0.7)[computer software]*. Amsterdam, The Netherlands: JASP Project. Retrieved from https:// jasp-stats. org.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*, 29–71.

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979.

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.

Ravenzwaaij, D., Moore, C. P., Lee, M. D., & Newell, B. R. (2014). A hierarchical Bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cognitive Science*, *38*, 1384–1405.

Todd, P. M., & Dieckmann, A. (2004). Heuristics for ordering cue search in decision making. In *Advances in Neural Information Processing Systems* (pp. 1393–1400).

van Ravenzwaaij, D., Newell, B. R., Moore, C. P., & Lee, M. D. (2014). Using recognition in multi–attribute decision environments. *Proceedings of the 35th Annual Conference of the Cognitive Science Society Austin, TX: Cognitive Science Society*, *17*, 3627–3632.

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in Human Neuroscience*, *5*.