Behav Res (2016) 48:29–41
DOI 10.3758/s13428-014-0557-9

Bayesian outcome-based strategy classification

Michael D. Lee

Published online: 20 February 2015

Abstract Hilbig and Moshagen (*Psychonomic Bulletin & Review, 21*, 1431–1443, 2014) recently developed a method for making inferences about the decision processes people use in multi-attribute forced choice tasks. Their paper makes a number of worthwhile theoretical and methodological contributions. Theoretically, they provide an insightful psychological motivation for a probabilistic extension of the widely-used "weighted additive" (WADD) model, and show how this model, as well as other important models like "take-the-best" (TTB), can and should be expressed in terms of meaningful priors. Methodologically, they develop an inference approach based on the Minimum Description Length (MDL) principles that balances both the goodness-of-fit and complexity of the decision models they consider. This paper aims to preserve these useful contributions, but provide a complementary Bayesian approach with some theoretical and methodological advantages. We develop a simple graphical model, implemented in JAGS, that allows for fully Bayesian inferences about which models people use to make decisions. To demonstrate the Bayesian approach, we apply it to the models and data considered by Hilbig and Moshagen (*Psychonomic Bulletin & Review, 21*, 1431–1443, 2014), showing how a prior predictive analysis of the models, and posterior inferences about which models people use and the parameter settings at which they use them, can contribute to our understanding of human decision making.

Code and data associated with this paper have been lodged with the Open Science Framework at https://osf.io/8ftug/.

M. D. Lee (✉)
Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100, USA
e-mail: mdlee@uci.edu

Keywords Decision making · Strategy classification · Bayesian inference · Graphical models · Individual differences

Introduction

Hilbig and Moshagen (2014) recently developed and demonstrated a method for making inferences about the processes used in multi-attribute decision making. The research challenge they tackle is to infer how people choose between two alternatives presented as a set of binary cues, where each cue has a known validity, which is defined as the probability that the cue indicates the correct answer when it distinguishes between two alternatives. The six decision models they consider include a number of well-studied heuristics, such as the one-reason heuristic take-the-best as well as tallying strategies. A new model that extends tallying approaches in a theoretically interesting and plausible way is also developed and considered.

Hilbig and Moshagen (2014) develop an approach for inferring the use of these models—as well as simple guessing and saturated models to provide upper and lower bounds—from data, based on the idea of "outcome-based classification" (e.g., Bröder and Schiffer, 2003; Bröder, 2010; Glöckner, 2009). This approach involves designing sets of problems that serve of critical tests, for which different models make different patterns of predictions as to how people will answer. The approach is logical, simple to understand, and experimentally feasible. Most of the development of the method—and much of the focus of Hilbig and Moshagen (2014)—revolves around developing a methodology that is able to express the models in a common formalism, and make inferences that are sensitive to both the goodness-of-fit and complexity of the models. Hilbig

and Moshagen (2014) use a multinomial processing tree (MPT: Batchelder & Riefer, 1999) formalism to implement the models of interest, and rely on an approximation to the Minimum Description Length (MDL: Grünwald, 2007, Wu et al., 2010.)

In this paper, we approach the same challenge tackled by Hilbig and Moshagen (2014), but make different choices regarding the formalism for expressing models and method for making inferences. Our approach is a fully Bayesian one (Lee & Newell, 2011; Scheibehenne et al., 2013). Using the Bayesian approach to formalize the models has the advantage of providing priors as a vehicle to express theoretical assumptions. We argue that the models considered by Hilbig and Moshagen (2014) have key theoretical assumptions that are naturally expressed by prior distributions, including joint prior distributions that are constrained in theoretically meaningful ways. The Bayesian approach to making inferences about the use of the models from data also has a number of statistical advantages. Because Bayesian inference is complete, coherent, and principled, it avoids approximations and limitations inherent in some other approaches, and provides detailed information about which decision models people use, and how they use them.

The structure of this paper is as follows. First, we describe the experimental data reported by Hilbig and Moshagen (2014). Secondly, we implement the six models considered by Hilbig and Moshagen (2014) in a Bayesian way, developing the appropriate priors, and examining the prior predictive distributions. Thirdly, we develop a graphical model for making inferences about the models in relation to data, and report the results of Bayesian inferences about model and parameter use from this graphical model for the experimental data. Finally, we discuss the advantages of our Bayesian approach to outcome-based strategy classification.

Hilbig and Moshagen's (2014) experiment

The experimental data collected by Hilbig and Moshagen (2014) involve a series of forced-choice decisions between two alternatives. The alternatives were always represented in terms the presence or absence of the same four cues, with all of the cue information presented on each trial. The validities of the cues were provided to participants, and cues were arranged in order of decreasing validity when the alternatives were presented. The different problems were designed to fall into three types, so that each model made the same prediction for each problem within a type, but different models make different predictions across the types.

The experimental design is summarized in Table 1. The definitions of "alternative A" and "alternative B" in terms of the four cues—"C1" through "C4"—are shown, and the partitioning of the 16 unique problems into the three types is also shown. The cue validities were 0.9, 0.8, 0.7, and 0.6 for the first, second, third, and fourth cues. Given the cue patterns and validities, each model predicts either that alternative A will be chosen, that alternative B will be chosen, or that a choice will be made at random. Table 1 shows these predictions for four models—guessing, take-the-best, equal weighting, and weighed additive—that were the basis for the experimental design. These four models, and the two additional models considered by Hilbig and Moshagen (2014), are formalized in the next section.

A total of 79 participants completed the experiment reported by Hilbig and Moshagen (2014). Each participant made 32 decisions for problems in each of the three types, for a total of 96 problems. The relevant data are simply counts, for each participant, of how many times alternative A was chosen for all of the problems within each type. Because of the structure of the model predictions for the problems types, these counts provide evidence for and against each participant's use of the various models.

Bayesian model implementation

Implementing the six competing decision models as probabilistic cognitive models requires specifying, for each model, the probability that alternative A will be chosen for each of the three types. For all of the models considered by Hilbig and Moshagen (2014), this is naturally done by specifying appropriate priors on these probabilities.

Priors for the six models

Guessing The guessing model simply assumes a random choice for each problem. Thus, the probability of choosing alternative A is one-half for all three types:

$$\theta_1^{\mathrm{g}} = \frac{1}{2},\ \theta_2^{\mathrm{g}} = \frac{1}{2},\ \theta_3^{\mathrm{g}} = \frac{1}{2}.$$

Take-the-best The Hilbig and Moshagen (2014) problem types are designed such that the alternative A always has the highest-validity discriminating cue, and thus corresponds to the decision predicted by take-the-best (Gigerenzer et al., 1999). Using the same error-bound of 0.5 as Hilbig and Moshagen (2014) for probabilistic adherence to this deterministic prediction, and assuming all values from perfect adherence to this upper bound are equally likely, gives the prior on an error rate $\epsilon^{\mathrm{t}} \sim \mathrm{Uniform}\,(0, 0.5)$. Using this error rate, the probability of take-the-best choosing alternative A for each of the problem types:

$$\theta_1^{\mathrm{t}} = 1 - \epsilon^{\mathrm{t}},\ \theta_2^{\mathrm{t}} = 1 - \epsilon^{\mathrm{t}},\ \theta_3^{\mathrm{t}} = 1 - \epsilon^{\mathrm{t}}.$$

Table 1 Design of Hilbig and Moshagen's (2014) experiment.

	Alternative A				Alternative B				Model Predictions			
Type	C1	C2	C3	C4	C1	C2	C3	C4	Guess	TTB	EQW	WADD
1	+	−	+	+	−	−	−	−	?	A	A	A
1	+	+	−	−	−	−	−	−	?	A	A	A
1	+	+	−	+	−	−	−	+	?	A	A	A
1	+	+	+	−	−	−	+	−	?	A	A	A
1	+	+	+	+	−	−	+	+	?	A	A	A
1	+	+	+	+	−	+	−	−	?	A	A	A
2	+	−	−	−	−	+	+	−	?	A	B	B
2	+	−	−	+	−	+	+	+	?	A	B	B
3	+	−	−	−	−	−	−	+	?	A	?	A
3	+	−	+	−	−	−	+	+	?	A	?	A
3	+	+	−	−	−	+	−	+	?	A	?	A
3	+	+	+	−	−	+	−	+	?	A	?	A

Each row corresponds to a problem, with alternative A and alternative B defined in terms of the presence or absence of four cues, C1–C4. The problems are partitioned into three types, so that models make the same prediction within a type, but generally different predictions between types. The predictions of the guess, take-the-best (TTB), equal weighting (EQW), and weighted additive (WADD) models are shown, with "A" indicating alternative A, "B" indicating alternative B, and "?" indicating random choice.

Intuitively, the error rate measures how probable it is that decisions will deviate from the deterministic prediction of take-the-best, and the same error rate is assumed to apply to all three types of problems.

Equal weighting The equal weighting model simply counts or tallies the number of cues each alternative has. Alternative A has the most cues for type 1 problems, alternative B has the most cues for type 2 problems, and both alternatives have the same number of cues for type 3 problems. Thus, given the error rate $\epsilon^{\mathrm{q}} \sim \mathrm{Uniform}\,(0, 0.5)$, the probabilities of choosing alternative A using the equal weighting model for the three problem types are

$$\theta_1^{\mathrm{q}} = 1 - \epsilon^{\mathrm{q}},\ \theta_2^{\mathrm{q}} = \epsilon^{\mathrm{q}},\ \theta_3^{\mathrm{q}} = \frac{1}{2}.$$

Weighted additive The weighted additive model combines cues not simply by counting them, but by weighting each according to their cue validities. A commonly-used version of this model uses the validities themselves as the weights, so that if v_k is the validity of the kth cue, $\sum_{k\in A} v_k$ is the total evidence for alternative A (e.g., Gigerenzer et al., 1999 Rieskamp and Otto, 2006). Hilbig and Moshagen (2014) note, however, that a cue with validity $\frac{1}{2}$ is completely uninformative, which leads them consider the modified weighted additive model with total $\sum_{k\in A}\left(v_k - \frac{1}{2}\right)$.

The justification given for this approach is that it is appropriate "to control for chance" (Hilbig & Moshagen, 2014). The issue is not, however, one of controlling for chance, but rather the need to change the scale on which the evidence provided by cue validities is expressed. Subtracting $\frac{1}{2}$ is not the appropriate transformation to change the scale. Two cues with validity 0.7 provide much less evidence for an alternative than a single cue with validity 0.9, yet these are rendered equivalent in the approach used by Hilbig and Moshagen (2014). If the goal of a weighted additive model is the independent combination of cue validity information, using the information or evidence in each cue in a rational or normative way, the weights should be expressed and summed on the log-odds scale, giving the total $\sum_{k\in A} \log \frac{v_k}{1-v_k}$ (Katsikopoulos & Martignon, 2006; Lee & Cummins, 2004; Lee & Zhang, 2012).[1] The only justification for considering the weighted additive model used by Hilbig and Moshagen (2014) is as a process account of a non-normative decision process people might use. Of course, there are infinitely many possible such models, but the idea that people might sum cue validities presented to them to assess an alternative is a reasonable one. Whether people routinely undertake some analysis that leads them to subtract $\frac{1}{2}$ from these validities seems more debatable.

As it turns out, both the subtraction of $\frac{1}{2}$ and the use of the log-odds scale produce weighted evidence tallies that favor the same alternative within a problem type for all of the specific problems used by Hilbig and Moshagen (2014). Under

[1] It is especially strange that Hilbig and Moshagen (2014) do not use this formulation, because in other parts of their paper, when generating simulated data from the probabilistic weighted additive model, they use logistic functions that invert the log-odds transformation.

both approaches, the alternative B has the greater weighted sum for type 1 and type 3 problems, but alternative A has the greater weighted sum for type 2 problems.[2] Thus, given the error rate $\epsilon^{w} \sim$ Uniform (0, 0.5), the probabilities of choosing the alternative A using the equal weighting model are

$$\theta_1^{w} = 1 - \epsilon^{w},\ \theta_2^{w} = \epsilon^{w},\ \theta_3^{w} = 1 - \epsilon^{w}.$$

Probabilistic weighted additive A theoretically innovative model introduced by Hilbig and Moshagen (2014) is a probabilistic extension of the weighted additive model. The basic idea is that people should be more likely to choose the alternative consistent with a weighted sum if that total provides more decisive evidence. Thus, not only which alternative has the greater weighted evidence, but how much greater the evidence is, are used in modeling choice behavior. Rather than committing to a particular relationship between weighted evidence and choice probabilities, Hilbig and Moshagen (2014) observe that the basic assumptions of the model can be expressed by order constraints on probabilities. This is a clever insight. It allows the idea of decisions depending on the magnitude of evidence to be included in the model, without having to make specific assumptions that are not central to the theoretical motivations for the model.

As noted in defining the weighted additive model, the evidence favors alternative B for type 1 and type 3 problems, and alternative A for type 2 problems. It turns out that the *magnitude* of this difference is largest for type 1, then type 3, and then type 2 problems, regardless of whether cue validities with $\frac{1}{2}$ subtracted or the log-odds approach is used.[3] There are now three error-rate parameters—ϵ_1^{p}, ϵ_2^{p}, and ϵ_3^{p}—for the three problem types. These error rates are order constrained such that $\epsilon_1^{p} \leq \epsilon_3^{p} \leq \epsilon_2^{p}$, and $\epsilon_2^{p} \sim$ Uniform (0, 0.5).[4] The probabilities of choosing alternative A are

$$\theta_1^{p} = 1 - \epsilon_1^{p},\ \theta_2^{p} = \epsilon_2^{p},\ \theta_3^{p} = 1 - \epsilon_3^{p}.$$

Saturated The final model we consider is a saturated model, which allows each choice probability to be its own unconstrained rate. This makes the model a "catch all" account of the data, since it has the flexibility to accommodate any pattern of behavior. In this sense, the saturated model fulfills the role of the "unclassified" classification in Hilbig and Moshagen (2014). The error-rate parameters for each problem type now allow for any possible probability of choosing alternative A, so that $\epsilon_1^{s} \sim$ Uniform (0, 1), $\epsilon_2^{s} \sim$ Uniform (0, 1), and $\epsilon_3^{s} \sim$ Uniform (0, 1), with simply

$$\theta_1^{s} = \epsilon_1^{s},\ \theta_2^{s} = \epsilon_2^{s},\ \theta_3^{s} = \epsilon_3^{s}.$$

One way of thinking about the guessing and saturated models is that they provide deliberately extreme accounts of decision making that "bookend" the substantive models being considered. The guessing model provides an extremely simple account of the data, while the saturated model provides a very complicated account that indexes all possible observed data. The take-the-best, equal weighting, weighted additive, and probabilistic weighted additive models lie between these extremes. The role of the guessing and saturated models is to insure that evidence for the substantive models is not inferred simply because they are the most or least complicated models among the set considered. Thus, for example, if evidence is found for the probabilistic weighted additive model, it is not because it is the most complicated, but because it provided a better account than both the simpler substantive models and the more complicated saturated model.

Prior predictive distributions

A good way to understand the similarities and differences between the six decision models is to examine their prior predictive distributions. This is a basic Bayesian construct, and gives the distribution of data expected under a model. It is found by considering the data a model would predict at every possible parameterization, and then weighting those predictions by the density of the prior at each parameterization.

Figure 1 shows the prior predictive distributions for the six decision models. The distribution is in the data space, and each point in the space corresponds to a possible outcome of the experiment. Each axis corresponds to one of the problem types, counting in how many of the 32 trials alternative A is chosen. The area of the cubes at each point in the space corresponds to the prior probability that a model gives to that experimental outcome.

The prior predictive distributions in Fig. 1 provide a number of intuitions about the models. As expected, the guessing model expects each alternative to be chosen roughly equally often for all problem types, and the saturated model gives all experimental outcomes significant prior probability. The take-the-best model expects alternative A to be chosen most often for all problem types, while the equal weighting, weighted additive, and probabilistic weighted additive models make different predictions, especially with respect to

[2] Although the weighted evidence is extremely close for both alternatives in the log-odds formulation for type 2 problems.

[3] This is true for all relevant individual problem comparisons, despite the fact that the log-odds formulation leads to differences in weighted tallies that are not equivalent for all of the problems within a type.

[4] Thus, conceptually, the set of error rate parameters $(\epsilon_1^{p}, \epsilon_2^{p}, \epsilon_3^{p})$ have a joint prior distribution that is uniform on the subspace of the half-unit cube that satisfies the order constraints.

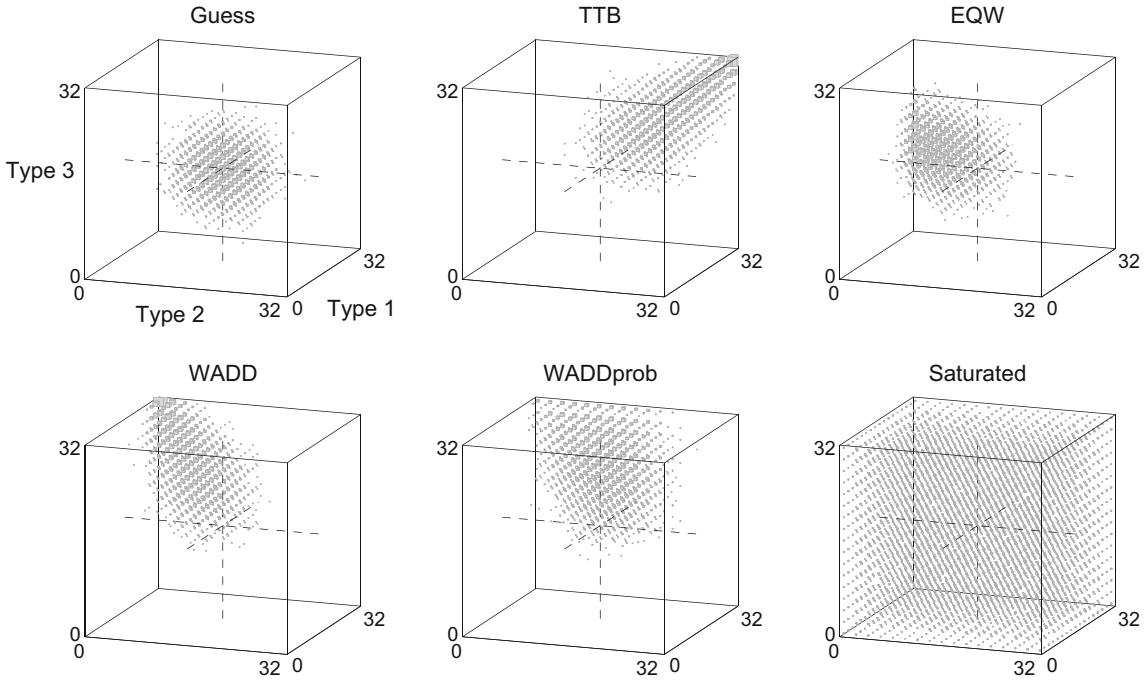


Fig. 1 Representations of the prior predictive distributions for each of the six decision models. Each panel corresponds to a model, representing the data space in terms of the number of times alternative A is chosen in the 32 total trials for problem types 1, 2, and 3. In this data space, the area of the cubes represent the prior probability the model assigns to each possible experimental outcome. (TTB = "take-the-best", EQW = "equal weighting", WADD = "weighted additive", WADDprob = "weighted additive probabilistic")

type 2 problems. A sensible interpretation of the volume of a model's prior predictions is as a measure of model complexity (Myung et al., 2000). Thus, it is clear that the saturated model, as it is designed to be, is the most complicated model. It is also clear, as should be expected, that the take-the-best and weighted additive models are equally complicated. It is interesting to observe the modest increase in complexity that arises from the probabilistic extension of the weighted additive model.

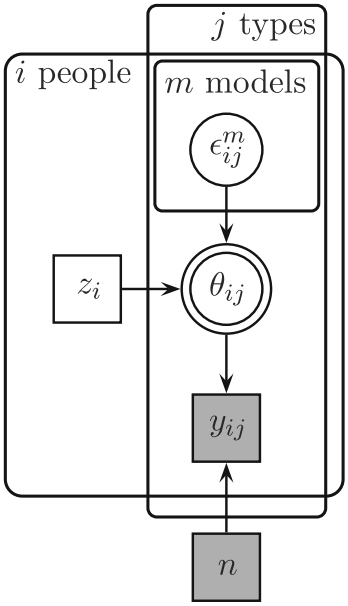
Bayesian inference

Graphical model

Figure 2 shows a graphical model that implements the six decision models, and can be used to infer which model individual participants used, and which parameterization of the model they used. In graphical models, nodes in a graph represent model parameters and data, and the graph structure indicates how the model assumes the parameters generate behavioral data, with encompassing plates used to show repetitions in the graph structure. Graphical models are especially convenient for conducting Bayesian analyses, because they are easy implemented in standard software such as WinBUGS (Lunn et al., 2000) and JAGS (Plummer, 2003), which apply standard computational methods to provide samples from the joint posterior distribution of the model conditional on the data. Lee and Wagenmakers (2013) provide an introduction to graphical modeling aimed at cognitive scientists.

In Fig. 2, the data are the counts y_{ij} of how often the ith participant chose alternative A for problems from the jth problem type. These counts are shown by a shaded and square node, because they are observed and discrete. The model assumes these data are generated, for the ith participant, by exactly one of the six decision models, and that they are all *a priori* equally likely. The model parameter z_i takes the values $1, 2, \ldots, 6$ to indicate which model the ith participant uses. It is shown as an unshaded and square node because it is unobserved and discrete. Each model, in our Bayesian formulation, corresponds to prior distribution over the probability of choosing alternative A for the different problem types, as formalized in the previous section.

The probability the ith participant will choose alternative A for a problem of the jth type is denoted θ_{ij} in Fig. 2. This is a deterministic node, as indicated by the double border. Basically, θ_{ij} take the values for the jth problem type that correspond to those of the decision model indicated by the z_i parameter. Thus, if $z_i = 1$ so that the ith participant is guessing, then $\theta_{i1} = \theta_{i2} = \theta_{i3} = \frac{1}{2}$, whereas if $z_i = 2$ so that they are using take-the-best $\theta_{i1} = \theta_{i2} = \theta_{i3} = 1 - \epsilon_i^t$, and so on. The observed data then follow a binomial distribution in terms of the probability θ_{ij}, so that for the



$$
\begin{aligned}
\epsilon_i^{\mathrm{t}}, \epsilon_i^{\mathrm{q}}, \epsilon_i^{\mathrm{w}} &\sim \mathrm{Uniform}(0, 0.5) \\
\epsilon_{i1}^{\mathrm{p}}, \epsilon_{i2}^{\mathrm{p}}, \epsilon_{i3}^{\mathrm{p}} &\sim \mathrm{Uniform}(0, 0.5) : \epsilon_{i1}^{\mathrm{p}} \leq \epsilon_{i3}^{\mathrm{p}} \leq \epsilon_{i2}^{\mathrm{p}} \\
\epsilon_{i1}^{\mathrm{s}}, \epsilon_{i2}^{\mathrm{s}}, \epsilon_{i3}^{\mathrm{s}} &\sim \mathrm{Uniform}(0, 1) \\
(\theta_{i1}, \theta_{i2}, \theta_{i3}) &= \begin{cases}
(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}) & \text{if } z_i = 1 \\
(1 - \epsilon_i^{\mathrm{t}}, 1 - \epsilon_i^{\mathrm{t}}, 1 - \epsilon_i^{\mathrm{t}}) & \text{if } z_i = 2 \\
(1 - \epsilon_i^{\mathrm{q}}, \epsilon_i^{\mathrm{q}}, \frac{1}{2}) & \text{if } z_i = 3 \\
(1 - \epsilon_i^{\mathrm{w}}, \epsilon_i^{\mathrm{w}}, 1 - \epsilon_i^{\mathrm{w}}) & \text{if } z_i = 4 \\
(1 - \epsilon_{i1}^{\mathrm{p}}, \epsilon_{i2}^{\mathrm{p}}, 1 - \epsilon_{i3}^{\mathrm{p}}) & \text{if } z_i = 5 \\
(\epsilon_{i1}^{\mathrm{s}}, \epsilon_{i2}^{\mathrm{s}}, \epsilon_{i3}^{\mathrm{s}}) & \text{if } z_i = 6
\end{cases} \\
z_i &\sim \mathrm{Categorical}(\tfrac{1}{6}, \ldots, \tfrac{1}{6}) \\
y_{ij} &\sim \mathrm{Binomial}\,(\theta_{ij}, n)
\end{aligned}
$$

Fig. 2 A graphical model for inferring which of six decision-making models each of a set of people uses, based on their decisions in a set of problems of different types

n problems completed, $y_{ij} \sim \mathrm{Binomial}\,(\theta_{ij}, n)$. In statistical terms, the graphical model in Fig. 2 is a simple latent mixture model, where the six decision models comprise the mixture components, and the latent categorical indicator variable z_i indicates which decision model is used by the ith participant.

Results

The JAGS code implementing the graphical model in Fig. 2 is presented in the Appendix. We applied the model to the data from Hilbig and Moshagen (2014), using 3 chains, and collecting 5000 samples from each chain after 5000 discarded burn-in samples, and thinning by a factor of 5, so that only each fifth sample was collected after burn-in. Convergence was checked using the $\hat{R}$ statistic and visual inspection of the chains (Brooks & Gelman, 1997).

Model use The key results involve the posterior distributions of the z_i classification parameters. These distributions gives the posterior probabilities, for each participant, that each of the six decision models was the one the participant used. They have a natural interpretation in terms of Bayes factors (Kass & Raftery, 1995) at the level of individual participants, and share with this standard Bayesian model selection method the advantage of automatically balancing goodness-of-fit with all forms of model complexity. Intuitively, the posterior probabilities inferred from the z_i parameters are measures of how well, on average, the prior predictions of each model shown in Fig. 1, correspond to the observed behavior of a participant, which is a single point in the data space.

The posterior probabilities of model use are summarized in Fig. 3. The individual participants are organized by the decision model for which they had the greatest posterior mass. There are 3 participants assigned in this way to the guessing model, 4 to the take-the-best model, 44 to the weighed additive model, 24 to probabilistic weighted additive model, and 4 to the saturated model. No participant was classified as most likely using the equal weighting model.

These results are largely consistent—except for one interesting difference—with those reported by (Hilbig and Moshagen, 2014, Table 2), who assigned 4 participants to the guessing model, 3 to take-the-best, none to equal weighting, 23 to weighted additive, 46 to probabilistic weighted additive, and left 3 unclassified. The noticeable difference is in the classification to the weighted additive model and its probabilistic extension, with the Bayesian analysis assigning more to the former, and the approximate MDL analysis assigning more to the latter. The appendix investigates this difference, first by calculating an exact version of MDL known as Normalized Maximum Likelihood (NML: Rissanen, 2001) and then comparing the NML and Bayesian classifications. These comparisons allow the impact of both the approximation used by Hilbig and Moshagen (2014) and the theoretical differences between the MDL and Bayesian approaches to be examined. The overall finding is that the classification differences are more superficial than fundamental. Because the classification of participants is

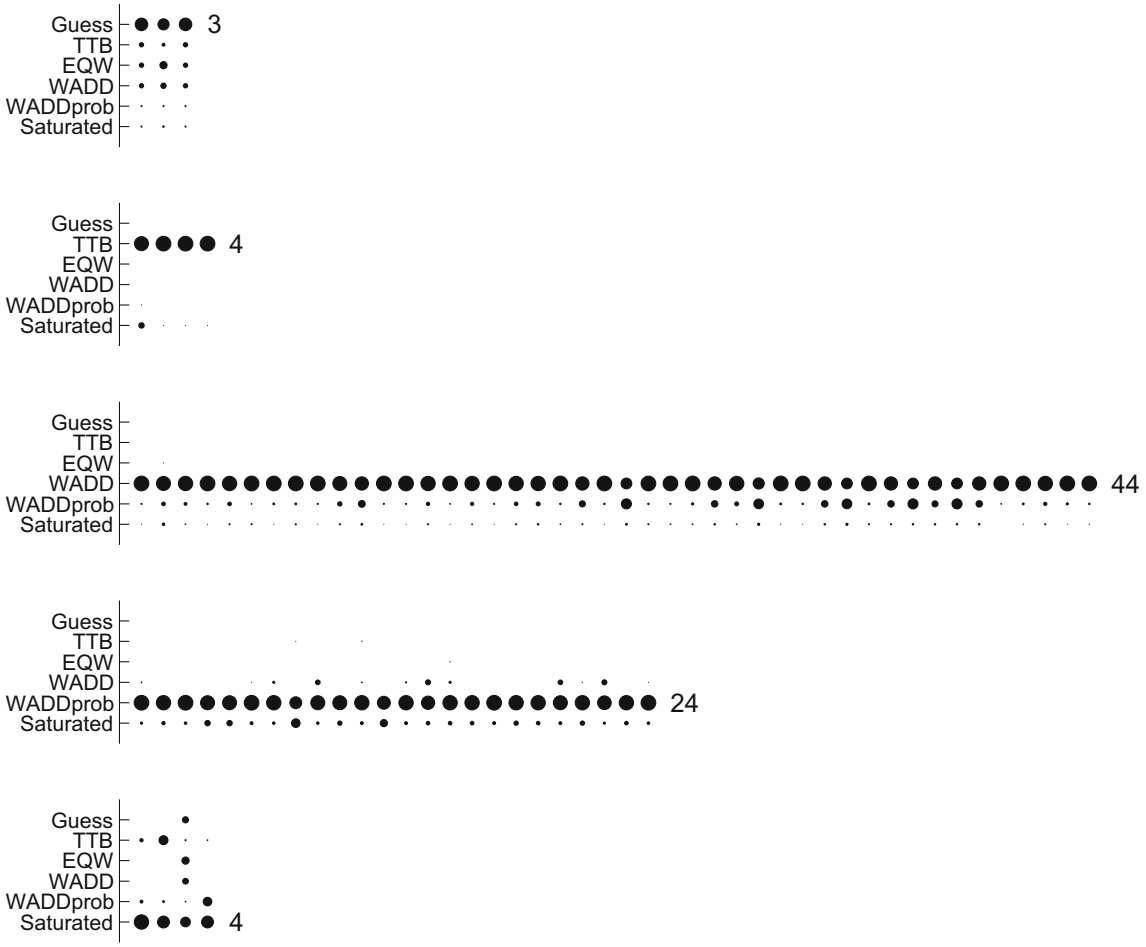


Fig. 3 Posterior inferences for model use for all participants, organized according to the most likely model for each participant. The area of each circle corresponds to the posterior probability that a participant used a model

based on which value is minimal over the set of models, even a small change in the values can change the minimum and hence the classification. The appendix shows that most of the differences in classification arise from such small differences in the underlying MDL, NML, and Bayesian values. In addition, using the exact NML criterion results in 9 probabilistic weighted additive participants being reclassified as weighted additive participants, giving total counts of 32 and 27 that more closely align with the Bayesian classification. A reasonable conclusion is that both the Hilbig and Moshagen (2014) and current Bayesian classifications tell a consistent story with very few take-the-best participants, and a large number of weighted additive participants who are roughly even divided between the original and the probabilistic extension of that strategy.

The results in Fig. 3 show additional information about the uncertainty of the classification that is automatically provided by the posterior probabilities from the Bayesian analysis. Organized by the most likely model, Fig. 3 shows the posterior probabilities each participant has for all of the six models. These probabilities are shown by the area of the circles. For example, the 3 participants classified as most likely using the guessing model corresponds to the three columns in the "guessing" panel. The areas of the circles shows that the inference is that, while all of these participants most likely guessed, there is some posterior probability they used each of the other models. For one of the participants, there is a noticeably larger posterior probability they used the equal weighting model.

With this interpretation in mind, it is clear from Fig. 3 that most the classifications are quite certain. The uncertainty largely involves participants classified as using the weighted additive model possibly having used the probabilistic extension, and participants classified as using the probabilistic weighted additive model possibly being better described by the saturated model. There are also interesting patterns of posterior uncertainty for those participants classified as using the saturated model. For many of these participants, one of the other decision models has appreciable posterior probability. One possible interpretation is that these participants followed one of the substantive models for some significant set of trials (or problems of a certain

type) but did not adhere closely enough overall to make this a probable account of all of their data, and so the "catch all" saturated model is inferred.

Parameter use Beyond the model indicator parameters z_i, the Bayesian analysis permits inferences about the error rate parameters within the various models. The guessing model has no parameters. As discussed earlier, the saturated model is less a substantive cognitive model than a statistical model that provides an upper bound on complexity. And no participant was inferred to use the equal weighting model. Thus, the interesting results about parameters involve the error rate parameters for the take-the-best, weighted additive, and probabilistic weighted additive models.

Figure 4 summarizes this information, showing the posterior distributions for the error rates for each participant, according to the model they were inferred as most likely to have used. For the take-the-best and weighted additive models, there is a single error rate parameter. It is clear most of the participants in these classifications followed the deterministic predictions of the models very closely, with very low error rates. For the probabilistic weighted additive model, richer information is available from the three order-constrained error rates for the three problem types. These marginal posteriors are shown in Fig. 4, with the three error rates for each participant displayed adjacent to one another, and visually grouped by the alternating shaded background. It is clear that, for essentially all of these participants, the error rates for the type 1 and type 3 problems are both generally low. What distinguishes these participants from those classified to the weighted additive model is their error rates for type 2 problems, which indicate they often chose alternative B, rather than alternative A as predicted by the model. The probabilistic extension of the weighted additive model affords this flexibility.

Discussion

Hilbig and Moshagen (2014) chose to use MPTs to implement the cognitive models of interest, and an approximation to MDL to evaluate the evidence data provide for the use of these models. These choices have a number of desirable properties. MPTs are a well developed and easily interpreted formalism for implementing processing

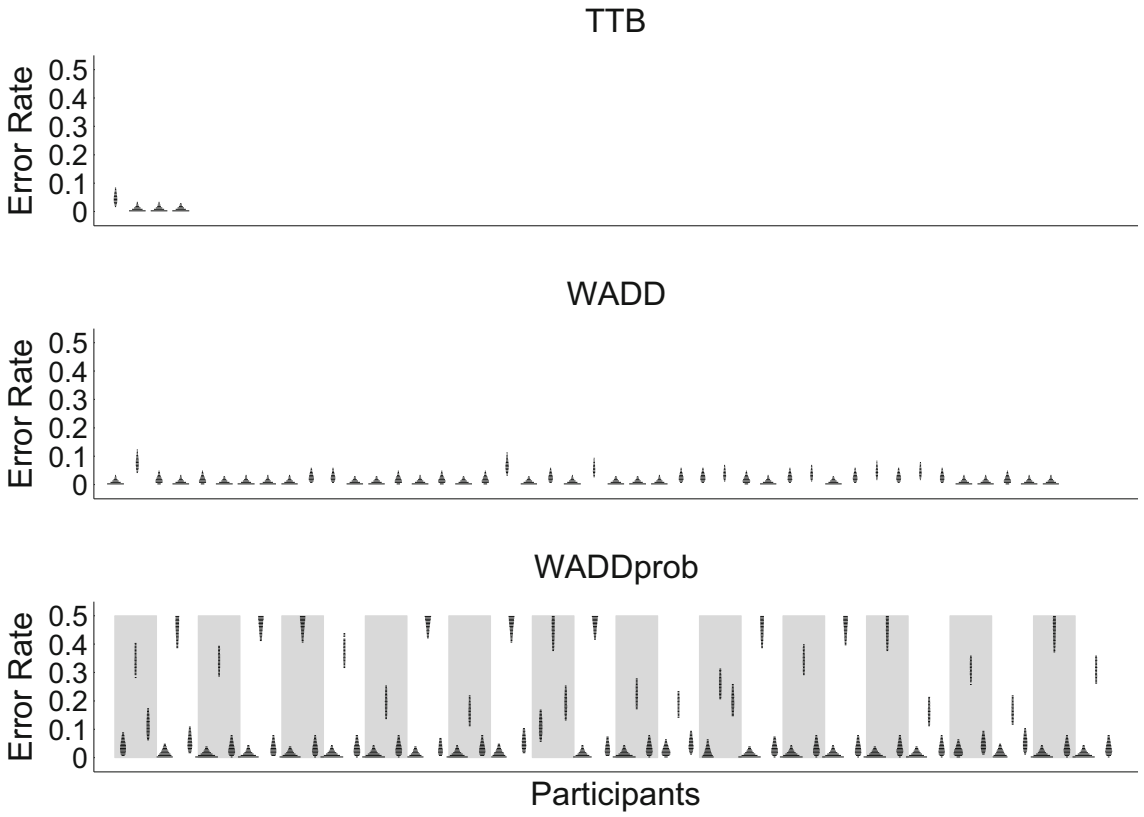


Fig. 4 Posterior inference for parameters for all participants classified as most likely using the take-the-best *(top panel)*, weighted additive *(middle panel)*, and probabilistic weighted additive *(bottom panel)* models. For the take-the-best and weighted additive models, the posterior distribution over the single error rate parameter is shown for each participant. For the probabilistic weighted additive model, the three error rate parameters for each participant are shown adjacent to one another

models of decision making, and Hilbig and Moshagen (2014) correctly note a number of ways the MDL approach improves on previously used model selection approaches. In particular, they note that their MDL approach addresses the key limitation of information criteria like the AIC and BIC, and frequentist approaches based on χ^2-style statistics, which are not sensitive to the component of model complexity that arises from the functional form of parametric interaction (see, for example Pitt et al., 2002, Shiffrin et al., 2008).

In this paper, we took a different approach, and applied Bayesian methods to implement and evaluate the decision models. On the methodological front, our approach is not new. Lee and Newell (2011, see also Lee & Wagenmakers 2013, Ch. 18) applied hierarchical Bayesian graphical modeling to the problem of inferring which decision processes people use, including making inferences about searching and stopping processes. Scheibehenne et al. (2013) recast the same methodological approach as one of testing "adaptive toolboxes", and demonstrated its advantages in a variety of detailed case studies. The same advantages apply to the problem tackled by Hilbig and Moshagen (2014). Bayesian inference provides a complete, coherent, and principled approach to making inferences about models and data. From a Bayesian perspective, everything that can be inferred about the models and their parameters from the data is expressed in the joint posterior distribution over the model indicator and error rate parameters, and the graphical model developed here allows that joint posterior distribution to be approximated as accurately as desired by sampling.

The practical advantages of the Bayesian approach are evident in our re-analysis of the Hilbig and Moshagen (2014) models and data. The prior predictive distributions in Fig. 1 provide intuitions about how the models are the same and different in the behavioral patterns they predict, and how they vary in complexity. The inferences about model use summarized in Fig. 3 provide expressions of uncertainty about the posterior probability that each participant used each model. The inferences about parameter use in Fig. 4 provide information about how participants used the various models. For the newly-developed probabilistic weighted additive model this information is especially valuable. The posterior distributions of the error rate parameters suggest that type 2 problems played a key role in distinguishing whether people were sensitive to the difference in the weighted evidence tallies for the two alternatives, as assumed by the probabilistic extension of the weighted additive model.

Our use of Bayesian methods is innovative, since Lee and Newell (2011) and Scheibehenne et al. (2013) largely use uninformative priors that are not based on theory.[5] Our use of priors to express relevant theory that distinguishes the models of interest follows a recent push to use priors as mechanisms for expressing psychological theory (Vanpaemel, 2010; 2011; Vanpaemel & Lee, 2012). The decision models considered by Hilbig and Moshagen (2014) involve constraints on possible values of parameters, or required relationships between parameters in the form of order constraints. For example, part of their definition of take-the-best is that it is followed with an error rate that is equally likely to be between 0 and 0.5, which is naturally expressed as a prior on the relevant parameter. Similarly, the core of the new probabilistic weighted additive model is a set of order constraints on error rates, which is again naturally expressed as joint prior information over the relevant parameters.

The MPTs used by Hilbig and Moshagen (2014) do not employ priors, and so they are forced to express this information as part of the data-generating process. This leads to the use of "dummy" variables that are not psychological parameters, but act to constrain the values of other variables that are psychological parameters. This is technically accurate, but we think the direct use of priors afforded by the Bayesian approach affords an elegant and useful complementary approach. More pointedly, we think the Bayesian approach has the advantage of forcing theoretical commitments to be precise, and highlighting deficiencies in existing theory, in ways that non-Bayesian approaches do not. For example, the prior distribution on an error-rate parameter must specify not just the range of possible values—Hilbig and Moshagen (2014) argue for 0.5, although smaller values seem theoretically defensible as required rates of adherence to the predictions of a deterministic model—but must also specify the prior distribution of possible error rates. Inference, especially in terms of model use, will be sensitive to the form of this prior, which is to be expected and is desirable. The error rate captures part of the theory about the decision processes being modeled, and changes to the theory and models should lead to changes in conclusions for the same data. Our setting of simple uniform priors on the error-rate parameters reflects a lack of sufficiently mature theorizing regarding the error rates that might be associated with the various decision processes. It seems likely that a more complete theory might predict that error rates closer to zero, so that the deterministic model is closely followed, are more likely than error rates near one-half, consistent with guessing, and these predictions are naturally formalized by

[5] The use of a Uniform (0.5, 1) prior on an "accuracy of execution" parameter, analogous to an error-rate parameter, by Lee and Newell (2011) is a possible exception.

a prior distribution. But, the theory must first be developed. In this way, the Bayesian approach acts as a spur to develop more complete accounts of the decision-making processes involved.

Which decision-making processes people use, and how they use them, are basic questions for decision-making research. Presenting people with problems for which different models make different predictions is an important research strategy in our attempts to understand these decision processes. We think Bayesian methods are extremely useful both for implementing decision models, for making inferences about those models in relation to the behavioral data, and for encouraging the development of more complete theories and models.

Acknowledgments I thank Ben Hilbig, E.-J. Wagenmakers, and Wolf Vanpaemel for helpful discussions.

Appendix A: Comparing MDL and Bayesian Classifications

Hilbig and Moshagen (2014) use a Minimum Description Length (MDL) approach to model selection, based on its application to Multinomial Processing Trees (MPTs) using the Fisher Information approximation (FIA) developed by Wu et al. (2010). One way to examine the impact of the FIA is to calculate the Normalized Maximum Likelihood (NML: Grünwald, 2007, Rissanen 2001). The NML is a form of MDL measurement that can be calculated exactly for the current models and data. Intuitively, as discussed by (Hilbig and Moshagen, 2014, p. 1437) the NML measure involves normalizing the maximum likelihood of the data for a given model by the sum of the maximum likelihood for that model over all possible data sets. Formally,

$$\mathrm{NML} = \frac{f\left(\theta^{\star}\left(D\right) \mid D\right)}{\sum_{D'} f\left(\theta^{\star}\left(D'\right) \mid D'\right)},$$

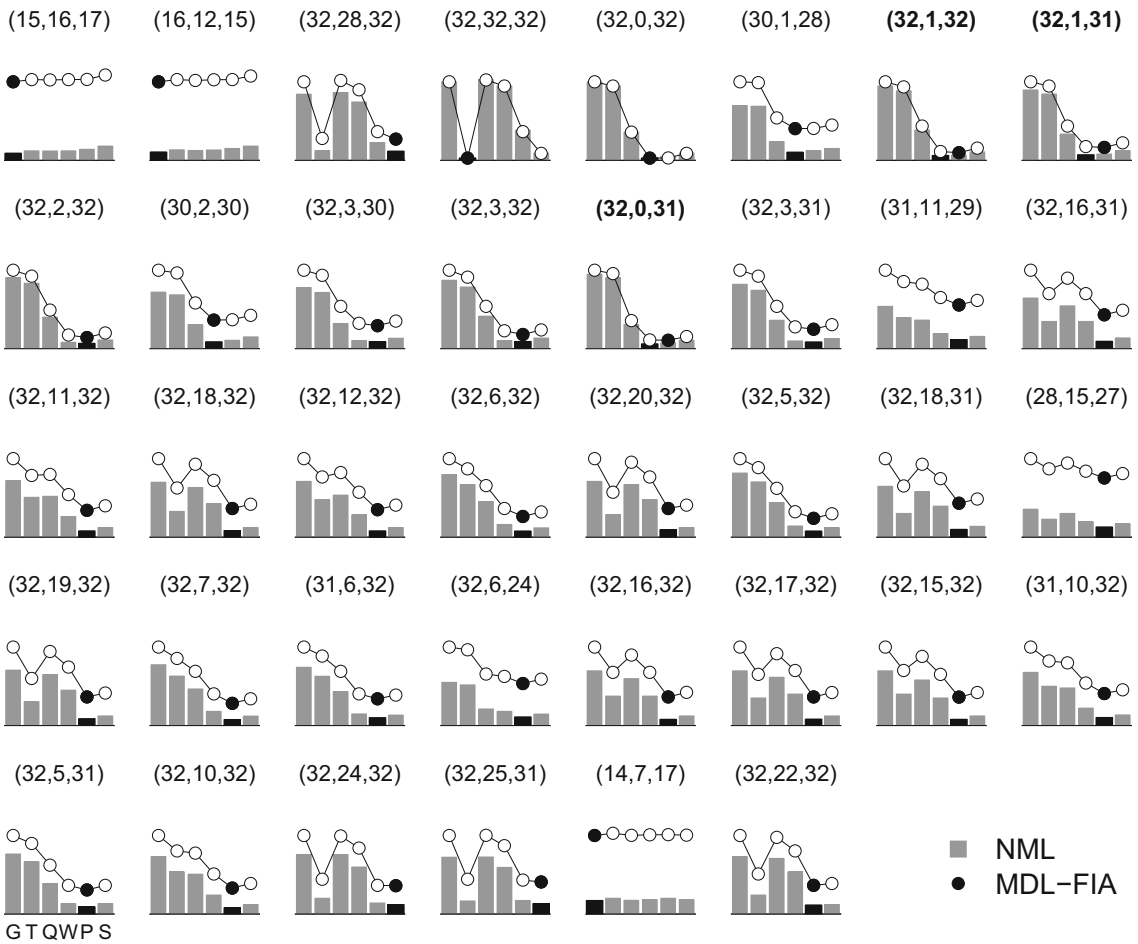


Fig. 5 Comparison of NML and MDL-FIA criteria. Each panel corresponds to the data pattern, shown by the counts of how often the first alternative was chosen for the three problem types. The bars show the NML values for each model and data pattern, and the circles show the MDL-FIA values. The minimum values corresponding to the participant classifications are indicated in black. Bold data pattern labels indicate data patterns for which the classifications disagree. (Model abbreviations: G = "guessing", T = "take-the-best", Q = "equal weighting", W = "weighted additive", P = "probabilistic weighted additive", S = "saturated")

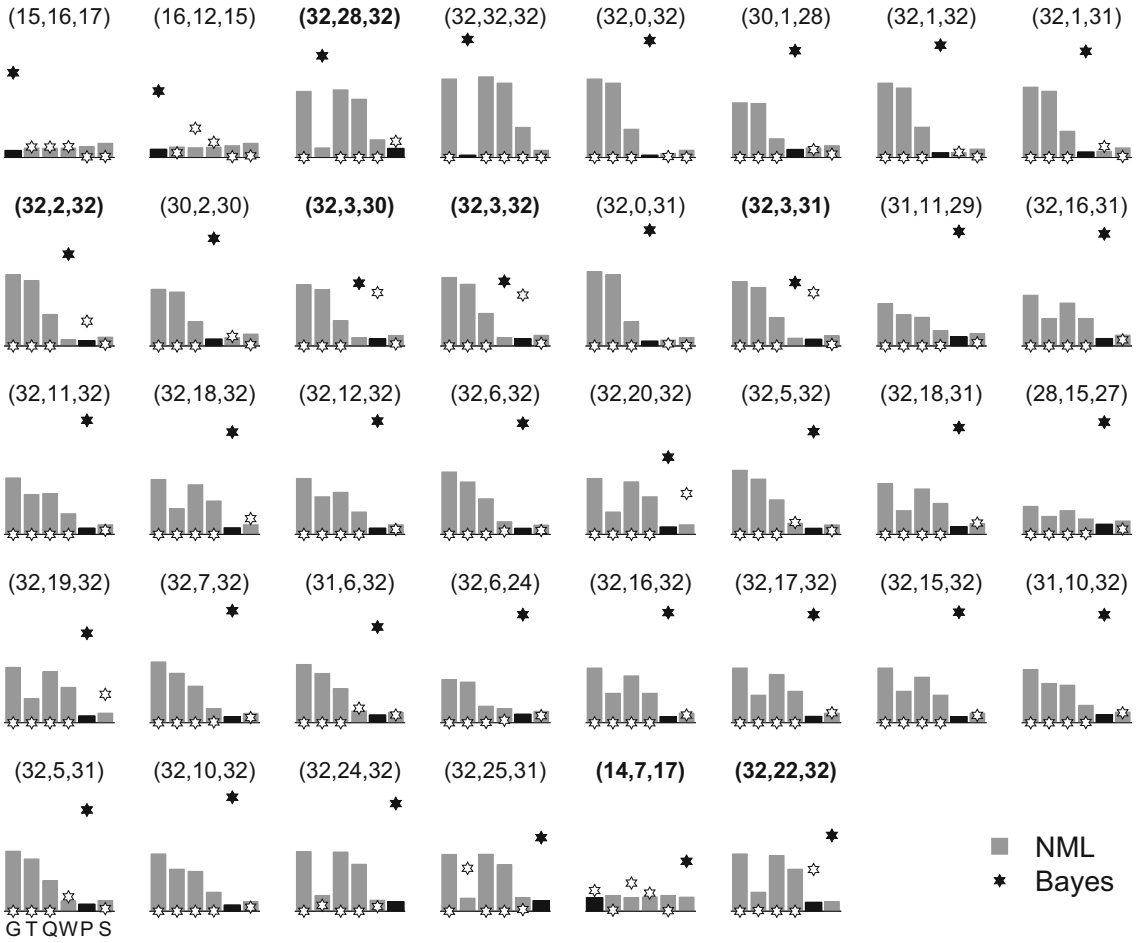


Fig. 6 Comparison of NML and Bayesian criteria. Each panel corresponds to the data pattern, shown by the counts of how often the first alternative was chosen for the three problem types. The bars show the NML values for each model and data pattern, and the stars show the Bayesian posterior probabilities. The minimum values corresponding to the participant classifications are indicated in black. Bold data pattern labels indicate data patterns for which the classifications disagree. (Model abbreviations: G = "guessing", T = "take-the-best", Q = "equal weighting", W = "weighted additive", P = "probabilistic weighted additive", S = "saturated")

where $f(\cdot)$ is the likelihood function, $\theta^\star(\cdot)$ is the parameterization of a model that maximizes the likelihood, D are the observed data, and D' are possible data. We calculated the NML for every model and every participant. This is computationally feasible, because the data space is discrete, and contains only $33^3 \approx 36,000$ possibilities. We found maximum likelihood values for the parameterized models using the MATLAB optimization functions `fminbnd` and `fmincon`.

Figure 5 compares the NML values to the MDL-FIA values used by Hilbig and Moshagen (2014).[6] Over the 79 participants, there are 38 different data patterns, comprising the counts of how often the first alternative was chosen for the three problem types. Each panel in Fig. 5 corresponds to one of these data patterns, shown by the counts at the top. The bars show the NML values for each model, and the circles show the MDL-FIA values. The bar and circle corresponding to the minimum value are shown in black, highlighting the classification of participants with that data pattern. The sub-panels are organized to start with those where the NML classification is for the guessing model, followed by take-the-best, weighted additive, probabilistic weighted additive, and the saturated model.

Figure 5 shows that the NML and MDL-FIA are extremely similar—up to a constant difference, which does not affect classification—for all of the data patterns. The 3 data patterns for which the classifications are different are the data patterns (32, 1, 32), (32, 1, 31), and (32, 0, 31). In all three cases the NML criterion favors the simpler single parameter weighted additive model whereas the MDL-FIA criterion favors the order-constrained three-parameter probabilistic weighted additive model. Intuitively, the deviations in the three data patterns from the pattern (32, 0, 32)

[6] I am very grateful to Ben Hilbig for supplying these values.

predicted by the deterministic application of the weighted additive model seem small, and it is clear from Fig. 5 that the disagreements in classification arise from small differences in the NML and MDL-FIA values. One possible reason for this difference is that the approximation assumptions for the MDL-FIA measure are least well satisfied at the boundary of the parameter space, and these data patterns correspond to parameterizations at the boundaries (Myung et al., 2000, p. 172).

Because a total of 9 participants produced one of the three data patterns, the classification counts change significantly moving from the MDL-FIA to the NML measure, but the results in Fig. 5 suggest this is a property of the all-or-none classification being sensitive to very small differences in the underlying values, rather than any serious deficiency in the MDL-FIA approximation.[7]

In a similar way, Fig. 6 compares the NML values to the Bayesian posterior probabilities. The bars again show the NML values for each model and each observed data pattern, while the stars show the Bayesian posterior probabilities. The minimum NML and maximum Bayesian posterior probabilities are highlighted in black, and correspond to the classification of participants with that data pattern. The 7 data patterns that cause disagreement between the NML and Bayesian classifications are highlighted.

Figure 6 makes it clear that in some cases—the data patterns (32, 3, 30), (32, 3, 32), and (32, 3, 31)—the posterior probabilities are very close for the two classifications. For the data patterns (32, 28, 32) and (32, 2, 32), the Bayesian classification remains with the simpler model that the data are most consistent with (take-the-best and weighted additive, respectively, whereas the NML classification is for the more complicated saturated and probabilistic weighted additive models. Finally, for the data patterns (14, 7, 17) and (32, 22, 32),which are very different from the predictions of the substantive models, the Bayesian classification is for the saturated model, whereas the NML classification is for guessing and the probabilistic weighted additive models, respectively.

Overall, for the three of the disagreeing data patterns, the relevant posterior probabilities are very close. For four of the disagreeing data patterns, the differences in the NML values for the two possible classifications are small. Thus, all of the disagreements could again be regarded as more a property of all-or-none classification rather than fundamental differences between the measures.

[7] Although we note that the MDL-FIA value for the guessing model is the same for all data patterns, which is strange, since the normalizing factor is constant, but the likelihoods of the observed data under the guessing model vary.

Appendix B: Code

The JAGS script implementing the graphical model in Fig. 2 is shown below.

```
model{
  for (i in 1:nSubj){
    # Model used by individual
    z[i] ~ dcat(base[1:6])
    for (j in 1:nItem){
      # Data
      y[i,j] ~ dbin(theta[i,z[i],j],nTrial)
    }
  }
  # Decision models
  for (i in 1:nSubj){
    # Guess
    theta[i,1,1] <- 0.5
    theta[i,1,2] <- 0.5
    theta[i,1,3] <- 0.5
    # TTB
    theta[i,2,1] <- 1-epsilon[i,1]
    theta[i,2,2] <- 1-epsilon[i,1]
    theta[i,2,3] <- 1-epsilon[i,1]
    # EQW
    theta[i,3,1] <- 1-epsilon[i,2]
    theta[i,3,2] <- epsilon[i,2]
    theta[i,3,3] <- 0.5
    # WADD
    theta[i,4,1] <- 1-epsilon[i,3]
    theta[i,4,2] <- epsilon[i,3]
    theta[i,4,3] <- 1-epsilon[i,3]
    # WADDprob
    theta[i,5,1] <- 1-epsilon[i,4]
    theta[i,5,2] <- epsilon[i,6]
    theta[i,5,3] <- 1-epsilon[i,5]
    # Saturated
    theta[i,6,1] <- epsilon[i,7]
    theta[i,6,2] <- epsilon[i,8]
    theta[i,6,3] <- epsilon[i,9]
    # Parameter priors
    # Error rate bounded for TTB, EQW and WADD
    for (j in 1:3){
      epsilon[i,j] ~ dunif(0,0.5)
    }
    # Order constrained and bounded for WADDprob
    for (j in 1:3){
      epsilonTmp[i,j] ~ dunif(0,0.5)
    }
    epsilon[i,4:6] <- sort(epsilonTmp[i,1:3])
    # Unconstrained for Saturated
    for (j in 7:9){
      epsilon[i,j] ~ dunif(0,1)
    }
  }
  # Model selection prior
  for (i in 1:6){
    base[i] <- 1/6
  }
}
```

References

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

Bröder, A. (2010). Outcome-based strategy classification. In Glöckner, A., & Witteman, C. (Eds.) *Foundations for tracing intuition: Challenges and methods, (pp. 61–82)*: Psychology Press.

Bröder, A., & Schiffer, S. (2003). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making*, *16*, 193–213.

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

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

Glöckner, A. (2009). Investigating intuitive and deliberate processes statistically: The multiple-measure maximum likelihood strategy classification method. *Judgment and Decision Making*, *4*, 186–199.

Grünwald, P.D. (2007). *The Minimum Description Length Principle*. Cambridge: MA: MIT Press.

Hilbig, B.E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, *21*, 1431–1443.

Kass, R.E., & Raftery, A.E. (1995). Journal of the American Statistical Association, *90*, 377–395.

Katsikopoulos, K.V., & Martignon, L. (2006). Naive heuristics for paired comparisons: Some results on their relative accuracy. *Journal of Mathematical Psychology*, *50*, 488–494.

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

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

Lee, M.D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*: Cambridge University Press.

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

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

Myung, I.J., Balasubramanian, V., Pitt, M.A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175.

Pitt, M.A., Myung, I.J., Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

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

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

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and informationin data. *IEEE Transactions on Information Theory*, *47*, 1712–1717.

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

Shiffrin, R.M., Lee, M.D., Kim, W.-J., Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 248–1284.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology*, *55*, 106–117.

Vanpaemel, W., & Lee, M.D. (2012). Using priors to formalize theory: Optimal attention and the Generalized Context Model. *Psychonomic Bulletin & Review*, *19*, 1047–1056.

Wu, H., Myung, J.I., Batchelder, W.H. (2010). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, *54*, 291–303.