From information processing to decisions: Formalizing and comparing psychologically plausible choice models

Daniel W. Heck a,*, Benjamin E. Hilbig b,c, Morten Moshagen d

a School of Social Sciences, Department of Psychology, University of Mannheim, Mannheim, Germany
b Department of Psychology, University of Koblenz-Landau, Landau, Germany
c Max-Planck Institute for Research on Collective Goods, Germany
d Department of Psychology, Ulm University, Germany

Article history:
Accepted 27 May 2017

Keywords:
Judgment and decision making
Take-the-best
Model selection
Minimum description length
Bayes factor

Abstract

Decision strategies explain how people integrate multiple sources of information to make probabilistic inferences. In the past decade, increasingly sophisticated methods have been developed to determine which strategy explains decision behavior best. We extend these efforts to test psychologically more plausible models (i.e., strategies), including a new, probabilistic version of the take-the-best (TTB) heuristic that implements a rank order of error probabilities based on sequential processing. Within a coherent statistical framework, deterministic and probabilistic versions of TTB and other strategies can directly be compared using model selection by minimum description length or the Bayes factor. In an experiment with inferences from given information, only three of 104 participants were best described by the psychologically plausible, probabilistic version of TTB. Similar as in previous studies, most participants were classified as users of weighted-additive, a strategy that integrates all available information and approximates rational decisions.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Every day, decision makers face situations in which judgments must be made based on uncertain information, that is, inferences must be drawn on the basis of observable information that may differ in terms of how well it actually predicts the to-be-judged criterion (Brunswik, 1955; Hammond, 1955). In such probabilistic inferences, psychological theories explain how information from multiple cues is integrated. For instance, when judging which of two cities is larger, one can rely on cues (e.g., whether it is a capital) that are probabilistically related to the criterion (i.e., population size) and thus indicate larger cities. Such judgments are made under uncertainty because the available cues typically only offer limited validity. Often, this is formalized by the cue validity $v_i$, the conditional probability that a binary cue $i$ leads to correct decisions given that it discriminates between options (Gigerenzer & Goldstein, 1996).

In such situations, Franklin's Rule or the weighted-additive strategy (WADD; Lee & Cummins, 2004) assume that decision makers integrate all available cues weighted by their validity. This strategy is in line with Simon’s (1976) notion of substantive rationality, because the decision maker maximizes her utility by combining all available information in an approximately optimal way (Lee & Cummins, 2004). In addition, heuristics were proposed as psychologically plausible shortcuts that aim at a reduction of cognitive effort (Shah & Oppenheimer, 2008). In line with Simon’s (1976) principle of procedural rationality,
these heuristics often use only a subset of the available information to draw inferences, thereby adhering to the assumption that the cognitive system may only have limited capacity and that certain heuristics are particularly useful in certain environments. Most prominently, Take-the-best (TTB; Gigerenzer & Goldstein, 1996) predicts that cues are considered lexicographically from most to least valid. A decision is made once a cue discriminates between the choice options, thereby eliminating the cost of further information search. This strategy produces high accuracy in non-compensatory environments (Hogarth & Karelaia, 2007) in which the dispersion of cue validities is large and the most valid cues are substantially more valid than others (see also Bröder, 2003; Glöckner, Hilbig, & Jekel, 2014). In contrast, TTB fares relatively poorly in compensatory environments with little dispersion in cue validities in which case equal-weight (EQW; Dawes, 1979) offers an alternative manner of simplification: EQW favors the option with the larger number of positive cue values without considering their validity, thereby reducing the costs for information integration.

As strategies are typically thought to differ in how well they approximate rational solutions but also in how effortful they are, research focuses on the moderating conditions that determine the trade-off between effort and accuracy (Payne, Bettman, & Johnson, 1988, 1993), that is, under which circumstances strategies like WADD, TTB, or EQW describe decisions best. Such factors include the environmental structure (e.g., Rieskamp & Otto, 2006), time pressure (Hilbig, Erdfelder, & Pohl, 2012; Rieskamp & Hoffrage, 2008), monetary costs (Newell & Shanks, 2003), working memory load (Bröder, 2003), and the task format (Platzer, Bröder, & Heck, 2014). For instance, decisions are better described by effort-reducing heuristics such as TTB whenever cues have to be retrieved from memory and are not readily available on display (Bröder & Schiffer, 2003b; but see Glöckner et al., 2014).

However, one recurring challenge in attempting to infer which decision strategies individuals use—or, more generally, which model may best account for the cognitive process of decision making—is that many models overlap in their choice predictions (Bröder & Newell, 2008; Glöckner, 2009; Jekel, Fiedler, & Glöckner, 2011). For instance, almost all decision strategies of substantive interest predict that an option is chosen if all cues favor it over alternative options, and therefore choices in such scenarios are not informative for testing competing explanations of the underlying judgment process. As a consequence of this overlap in choice predictions, decision strategies cannot be reliably teased apart based solely on counting how often individuals choose the option predicted by a strategy (Bröder & Schiffer, 2003a). Also, strategies may predict guessing (e.g. EQW would guess if the sum of cue values does not differ between options) and thus it cannot be determined whether any one choice was in line with this prediction. Taken together, as Newell (2005) emphasized, ‘it is essential to incorporate an error theory […] to account for the stochastic deviation from the deterministic rules of heuristics’ (p. 12).

To overcome these limitations, Bröder and Schiffer (2003a) proposed classifying participants as users of one of the decision strategies by means of statistical model selection. Instead of merely counting how often decisions coincide with strategy predictions, this method selects the most likely strategy given a specific response pattern or vector across different constellations of cue values (see details in Section 2). Importantly, these cue patterns are carefully selected to differentiate between the decision strategies under scrutiny. To derive the likelihood of observable response patterns, statistical strategy classification rests on the critical assumption that each strategy is applied with a constant, small error probability (reflecting, for instance, attention fluctuations; Bröder & Schiffer, 2003a). Put differently, it is assumed that, in some trials, individuals using a particular strategy choose a strategy-inconsistent option due to unsystematic factors such as inattention or strategy execution errors. Thereby, systematic deviations from the predicted response pattern are penalized more strongly than unsystematic deviations arising from application errors.

1.1. Psychologically plausible models of strategy application errors

The strategy-classification method of Bröder and Schiffer (2003a) builds on the assumption that strategies are applied with unsystematic execution errors formalized by a constant error probability. However, concerning WADD, it is implausible that the weighted integration of cue values and validities results in a constant error probability. Empirically, decisions are often faster and more consistent (i.e., the predicted option is chosen more often) when cues provide coherent evidence compared to scenarios where cues provide contradictory information (e.g., Birnbaum & Jou, 1990; Brown & Tan, 2011; Glöckner & Betsch, 2012; Heck & Erdfelder, in press; Pohl & Hilbig, 2012). Importantly, psychological theories that assume a weighted integration of cue values and validities predict that more evidence for an option leads to smaller error probabilities. Many axiomatic decision theories (e.g., Luce’s choice rule; Luce, 1959) imply graded instead of constant error probabilities, assuming that differences on a unidimensional latent variable (utility, evidence) are mapped to choice probabilities by a monotonic link function (e.g., by a logistic or probit link function; Luce, 1979; Thurstone, 1927). On the other hand, process models implementing or approximating WADD describe the mechanism of how cue information is integrated, for instance, by an automatic parallel search for a coherent representation of the available information (Glöckner & Betsch, 2008a) or by sequential evidence accumulation (Busemeyer & Townsend, 1993; Diederich, 1997; Trueblood, Brown, & Heathcote, 2014). Irrespective of specific underlying mechanisms, process models usually predict how strongly an option is preferred, which in turn implies graded instead of constant error probabilities.

Overall, psychological theory implies that error probabilities of WADD must not be constant across different cue constellations, but depend on the amount of evidence in favor of the preferred choice option (i.e., on the evidence difference). Importantly, this psychologically plausible assumption should be reflected by the statistical model used for strategy classification. Addressing this issue, Hilbig and Moshagen (2014) proposed a probabilistic version of WADD (WADDprob) reflecting the assumption that larger differences in the evidence for the preferred option result in smaller error probabilities. Put
differently, WADDprob assumes a specific rank order of the error probabilities that depends on the cue constellations, whereas WADD assumes that error probabilities are constant across all cue patterns. In an experiment with decisions from givens (Hilbig & Moshagen, 2014), a majority of 58% participants were classified as WADDprob users, compared to 29% for the originally proposed, deterministic version of WADD, which highlights the importance of one’s error theory for the plausibility and empirical success of models for probabilistic inferences. Note that these results are also in line with the fact that probabilistic models often outperform their deterministic counterparts in risky choices (e.g., Regenwetter, Dana, & Davis-Stober, 2011; Rieskamp, 2008).

Given the theoretical and empirical superiority of probabilistic over the deterministic representations, the question arises whether constant errors are psychologically plausible when modeling heuristic strategies such as TTB. This is a vital question given that prior experiments have consistently suggested that, in inferences from given information, virtually no participant uses TTB (despite its high accuracy and noteworthy effort reduction; Marewski, Gaissmaier, & Gigerenzer, 2009). Potentially, this is a premature conclusion driven by an inappropriate—that is, psychologically implausible—error model attached to TTB. According to the universally implemented TTB model, cues are processed by applying the corresponding search, stopping, and decision rules without making errors (Gigerenzer & Gaissmaier, 2011). Given the deterministic outcome of this lexicographic judgment process, strategy-inconsistent decisions can only be due to errors in response selection, inattention, and other unsystematic factors, resulting in a constant error probability.

However, a cognitive-process interpretation of TTB suggests that strategy-inconsistent choices might actually be due to error-prone elementary processing steps (Glöckner, 2009; Payne et al., 1988). By definition, TTB assumes a sequential process of lexicographic cue comparisons, which might be prone to errors with some probability. Since such stepwise errors accumulate in a serial process, a psychologically plausible version of TTB implies that the strategy-application error increases with the required number of processing steps. Put differently, TTB will result in few strategy-inconsistent responses when the cue considered first (i.e., the most valid one) discriminates between the options, but in more strategy-inconsistent responses when the decision-relevant, discriminating cue follows several non-discriminating, more valid cues. This psychologically plausible, probabilistic version of TTB (TTBprob) makes more explicit assumptions about the source of errors when applying TTB and immediately follows from the framework of elementary processing steps (Payne, Bettman, & Johnson, 1993).

More generally, a serial-process account for TTB follows directly from the processing assumptions made in the adaptive-toolbox framework (Gigerenzer & Goldstein, 1996), and merely requires a rank order of cue validities for its application (instead of a precise numerical representation of ecological validities; Gigerenzer, Hoffrage, & Goldstein, 2008). Specifically, the assumption that more processing steps require more effort (and are also more error-prone) is at the core of the theoretical foundation of fast and frugal heuristics (Gigerenzer & Gaissmaier, 2011) and empirical tests. For instance, Glöckner (2009) used the number of elementary processing steps of TTB to derive predictions for response times while retaining the assumption of constant error probabilities for all cue constellations. Moreover, the serial-processing assumption is in line with findings that individuals most likely using TTB need more time for decisions that require a larger number of steps, that is, decision times increase as the position of the first discriminating cue in the cue rank order increases (Bröder & Gaissmaier, 2007). Note that other heuristics of the adaptive toolbox also assume that the building blocks of heuristics can be prone to errors instead of assuming a constant error probability on the outcome of the sequential process. For instance, in the domain of risky choices, the priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006) predicts a rank order of error probabilities due to error-prone, attribute-wise comparisons.

To test the psychologically plausible TTBprob strategy, we present new item types for which the competing decision strategies make distinct predictions. Importantly, TTBprob can directly be compared against a deterministic TTB version with constant errors, other heuristics (such as EQW or guessing), and different rational models that integrate all information (WADD and WADDprob) within the same modeling framework. For this purpose, we implement two model-selection methods (minimum description length and the Bayes factor; Hilbig & Moshagen, 2014; Lee, 2016) and rely on simulations to determine the number of choices per participant required for high recovery rates. We also address a methodological issue by comparing minimum description length and the BF and show that both methods resulted in similar classifications for the present set of strategies, items, and number of observations.

2. Generalized outcome-based strategy classification
2.1. Deterministic decision strategies

Outcome-based strategy classification uses the observed choice pattern across item types to select the decision strategy that most likely generated the data (Bröder & Schiffer, 2003a). In a typical experiment, participants repeatedly choose between two options that differ on several probabilistic cues. The approach relies on so-called item types such as those shown in Table 1 for which the decision strategies under scrutiny predict distinct choice patterns. Note that, in empirical studies, each item type is represented by multiple cue constellations, thereby increasing the perceived variability of choice scenarios. Usually, the cues differ in validity, defined as the probability that a positive cue value correctly predicts a higher criterion value.

For a set of item types, each strategy predicts a pattern of choices as shown in the lower part of Table 1. TTB assumes that cues are compared from most to least valid. Once a cue discriminates between the options, the option with the positive value
is chosen without considering less valid cues (Gigerenzer & Goldstein, 1996). For instance, TTB chooses Option A for Item Type 4 since it is preferred by the most valid cue, whereas less valid, positive cues for Option B are not considered. In contrast, WADD integrates all cues weighted by validity. According to rational decision models (Lee & Cummins, 2004), the option with the higher odds of having a large criterion value is chosen. More precisely, given the cue validities vi, the relative evidence for Option A over B is defined by the log-odds

\[
L_{AB} = \sum_{i \in FA} \log \frac{v_i}{1 - v_i} - \sum_{i \in FB} \log \frac{v_i}{1 - v_i},
\]

where the two sums are only over those cues that favor Option A and B, respectively (as indexed by the sets FA and FB; Lee & Cummins, 2004). For Item Type 4, this results in choosing Option B since the sum of log-odds of the cues favoring this option exceeds that of Option A (2.23 vs. 2.19 resulting in the difference \(L_{AB} = -0.04\)).\(^1\) EQW merely chooses the option with the larger number of positive cues, which also results in choosing B since it is supported by two cues compared to a single cue for Option A. Finally, GUESS models responses at chance level (i.e., 50% for any binary choice).

Based on a pattern of deterministic predictions, Bröder and Schiffer (2003a) assumed that each strategy is applied with a constant, small error probability across item types. In other words, decision makers are assumed to follow a strategies' prediction only imperfectly. In some trials, the strategy-inconsistent option is chosen merely due to decision-irrelevant factors such as inattention, misreading, or motor errors. This explicit error theory transforms the vector of deterministic strategy predictions into a statistical model. More precisely, the observed response frequencies per item type follow a binomial likelihood in which the error probability serves as a free parameter. For instance, TTB predicts that Option A is chosen for all item types in Table 1 and thus constrains the probabilities bi of choosing Option B in item type i to be identical, \(b_1 = b_2 = b_3 = b_4 \leq 0.50\), where the upper bound of 0.50 ensures that the strategy actually predicts the preferred option above chance level (Hilbig & Moshagen, 2014). Similarly, WADD implies the constraint \(b_1 = b_2 = b_3 = 1 - b_4 \leq 0.50\), and EQW the constraint \(b_1 = b_2 = 1 - b_4 \leq b_3 = b_4 = 0.50\). Note that, due to the constraint of \(\leq 0.50\), bi denotes that Option B is the non-preferred option (the strategy predicts choice of Option A) and \(1 - b_i\) denotes that Option B is the preferred option (the one predicted by the strategy).

The transformation of deterministic choice predictions into statistical models allows for using model-selection methods to classify each participant as a user of one of the strategies. For instance, the BIC can be used to select the model that makes the best trade-off between fit and the number of free parameters (e.g., GUESS does not require a free parameter and therefore is more parsimonious; Bröder & Schiffer, 2003a).

### 2.2. Probabilistic versions of WADD and TTB

As derived in Section 1.1, axiomatic decision theories and process models such as evidence-accumulation (e.g., Diederich, 1997) or parallel-constraint-satisfaction models (Glöckner & Betsch, 2008a) predict that WADD will not be applied with a constant error rate, but that responding will be less error-prone if the evidence for the superior option increases. To incorporate this psychologically plausible assumption in a statistical model, Hilbig and Moshagen (2014) proposed a probabilistic version of WADD that imposes order constraints on the error probabilities instead of constraining the errors to be identical across item types. For instance, Table 1 shows the relative evidence in favor of Option A for each item type as the difference in the sum of the log-odds for the two options (\(L_{AB}\)). This relative evidence in favor of the preferred alternative decreases from

---

\(^1\) When implementing WADD based on weighted sums of cue values, validities have to be corrected for chance by subtracting 0.50 (Jekel & Glöckner, in press).
Item Type 1 to 4, implying that the error of strategy application increases. Formally, this error theory for WADDprob is represented by the order constraint $b_1 < b_2 < b_3 < 1 - b_4 < 0.50$. Similar as for the deterministic strategies, the upper bound of 0.50 ensures that the preferred option is predicted above chance level (Hilbig & Moshagen, 2014).

Whereas the standard version of TTB assumes constant error probabilities, a serial-process interpretation of TTB implies that more processing steps result in a higher total error probability. However, for all three item types used by Bröder and Schiffer (2003a) and Hilbig and Moshagen (2014), TTB requires only a single step. Hence, the predictions of TTBprob do not differ from those of a deterministic version of TTB and the models cannot be teased apart. To allow for an empirical test of TTBprob, we therefore developed the new item types shown in Table 1. For these items, the probabilistic version of TTB assumes that the error is smallest for Item Types 2 and 4, which only require the evaluation of the most valid cue. In contrast, Item Types 1 and 3 require two and four processing steps, respectively, which leads to the following order constraints on the error probabilities: $b_2 = b_4 < b_1 < b_3 < 0.50$. Importantly, the new item types are also diagnostic for WADDprob and all deterministic strategies under consideration.

3. Model-selection methods

Fig. 1 gives an overview of the competing decision strategies and their constraints on the probabilities of choosing Option B for the new item types in Table 1. By specifying all strategies and error models within the same framework, the competing theories can directly be tested against each other. However, the absolute fit of some decision strategies can no longer directly be assessed as recommended by Moshagen and Hilbig (2011). In the case of WADDprob, the standard goodness-of-fit test cannot be applied because the model requires four free parameters $b_1, b_2, b_3, \text{ and } 1 - b_4$ to explain four independent choice frequencies for Item Types 1 to 4 (Moshagen, 2010). A related issue concerns TTBprob, which also places order constraints on the binomial error probabilities and therefore requires advanced methods to derive the correct degrees of freedom for the goodness-of-fit test (Davis-Stober, 2009). Alternatively, observed choice frequencies can be compared to posterior-predictive frequencies in a Bayesian framework (cf. Section 4.2). However, the corresponding posterior-predictive $p$-values (Meng, 1994) provide only a heuristic measure of goodness of fit, because their asymptotic distribution under the data-generating model is not known in general (Meng, 1994).

As a remedy, Hilbig and Moshagen (2014) proposed a more principled approach, namely to include a maximally flexible baseline model in the set of competing strategies. This baseline model does not constrain the probabilities $b_1, b_2, b_3, \text{ and } b_4$ of choosing Option B for the four item types, and thereby captures response patterns that are not predicted by any substantive strategy under consideration (e.g., choosing Option B with high probability). Statistically, the baseline model is saturated, since it matches any observed response frequencies perfectly, and represents the reference model in classical goodness-of-fit tests (Read & Cressie, 1988). Moreover, it is often referred to as the 'encompassing model' in Bayesian tests of order-constraints (with the additional assumption that the prior is proportional to that of the order-constrained models; Hoijtink, 2011; Davis-Stober, Brown, & Cavagnaro, 2015). The inclusion of this baseline in model selection protects against selecting substantive strategies that do not fit the data (Hilbig & Moshagen, 2014).

Fig. 1 shows that all substantive strategies are statistically nested in the baseline model, that is, they are special cases due to their constraints on the error probabilities. By implication, the baseline model will fit at least as good as all substantive strategies to any response pattern. Importantly, this flexibility is only partly reflected in the number of free parameters—WADDprob has four free parameters similar to the baseline model but will fit strategy-inconsistent data patterns worse. Thus, model selection should not rely solely on the number of free parameters as an indicator of complexity (which is the case for the BIC); instead, model selection also needs to consider functional complexity due to order constraints (Myung, 2000).

3.1. Normalized maximum likelihood

To test WADDprob, Hilbig and Moshagen (2014) relied on model selection based on the minimum-description-length principle (MDL; Rissanen, 1996). According to MDL, the model that provides the shortest description of the data should be selected because this model will predict new data best due to an optimal trade-off between fit and complexity. Importantly, MDL does not merely count the number of free parameters but also considers functional model complexity as that emerging from order constraints.

As a specific model-selection criterion, we compute the normalized maximum likelihood (NML; Rissanen, 2001), which provides an intuitive definition of model complexity in the MDL framework. For a given strategy, NML is defined as the sum of two terms that quantify model misfit and complexity:

\[
NML = -\log f(x; \hat{\theta}_x) + \log \sum_{y \in Y} f(y; \hat{\theta}_y),
\]

where $f(x; \theta)$ is the model’s likelihood function given the parameters $\theta$ (here, the product-binomial likelihood with the model-specific probability parameters shown in Fig. 1) and $\hat{\theta}_x$ the corresponding maximum-likelihood estimate given the

---

2 Strictly speaking, Eq. (2) is the negative logarithm of the NML distribution (Grünwald, 2007).
In the definition of NML in Eq. (2), the first term is the negative log maximum likelihood for the observed response pattern \( \mathbf{x} \), which becomes larger if a model fits worse, whereas the second is the complexity term (the logarithmized sum of maximum-likelihood values across all possible response patterns \( \mathbf{y} \) in the outcome space \( \Omega \)), which becomes larger if a model is more complex. Intuitively, complexity simply captures the ability of a model to fit arbitrary data \( \mathbf{y} \). Once NML values have been computed for all models under scrutiny, the decision strategy with the smallest criterion value is selected (i.e., the model that provides the best compromise between misfit and complexity).

Note that Hilbig and Moshagen (2014) originally used the Fisher information approximation (FIA; Rissanen, 1996; Wu, Myung, & Batchelder, 2010), which approximates NML for large sample sizes. However, FIA can lead to biased model selection in small samples and may assign a larger complexity term to a nested model than to a more general model (Navarro, 2004). According to NML, however, a nested model is always less complex by definition because the general model will fit all possible response patterns as least as well as the nested model. To avoid such a bias for the seven models in Fig. 1, FIA requires a lower-bound sample size of \( N_0 = 326 \) per participant, that is, \( n = 82 \) trials for each of the four item types (Heck, Moshagen, & Erdfelder, 2014). This large number of trials is empirically unfeasible and likely to result in lack of motivation, fatigue, and other unwanted side effects.

To compute the NML complexity term, each strategy is fitted to all possible observable choice patterns and the resulting maximum-likelihood values are summed up. For strategies for which analytical estimators are not available (e.g., TTBprob), it is important to perform multiple optimization runs with different starting values to reduce the chance of stopping at a local maximum and thereby underestimating model complexity. Fig. 1 shows the resulting NML complexity terms for all strategies and \( n = 40 \) choices per item type. Despite having the same number of free parameters, WADDprob is much less complex than the baseline model (4.37 vs. 8.61 on the log-likelihood scale, respectively). Note that this difference is even more pronounced than that between WADDprob and the deterministic version of WADD (4.37 vs. 2.17, respectively), although the latter actually differ in the number of free parameters whereas the former do not.

Note that the NML complexity terms in Fig. 1 depend on the overall number of choices, the particular item types in Table 1, and the proportion of choices per item type. Whereas the overall number of choices only affects the absolute complexity values, the latter factors may change the rank order of nonnested models. By definition, according to NML, a nested model can never be more complex than its more general counterpart. Since the statistical relation of decision strategies is identical for all possible cue patterns (e.g., TTB is always nested in TTBprob), the arrows in Fig. 1 imply a weak partial order of NML complexities for any subset of item types. However, for pairs of nonnested strategies that are not connected by arrows (e.g., WADDprob and TTBprob), the order of complexities can in principle switch depending on the item types. For instance, when selecting cue patterns with constant log-odds, WADDprob predicts constant error probabilities and thus reduces to WADD. If the first discriminating cue has a different position in the validity hierarchy across these cue patterns, TTBprob requires at least two error probabilities and will thus be more complex than WADDprob. In sum, this shows that the complexity of decision strategies depends on the stimulus material, a fact that not only needs to be considered when determining the number of choices per participant for empirical studies (see Section 3.3), but that is also of interest theoretically: Whether or not a model (reflecting a judgment strategy) is comparatively complex cannot be determined per se but also depends on the structure of the environment.

### 3.2. The Bayes factor

The Bayes factor (BF) is an alternative model-selection method that also accounts for functional model flexibility and is thereby appropriate to test order constraints (Jeffreys, 1961; Kass & Raftery, 1995). The BF centers around the marginal probability of a response pattern \( \mathbf{x} \) given a model \( \mathcal{M}_i \) with likelihood function \( f_i(\mathbf{x} | \theta) \),
\[ p(\mathbf{x} | \mathcal{A}_1) = \int_{\Theta_1} f_1(\mathbf{x} | \theta) \pi_1(\theta) \, d\theta. \]

where \( \pi_1(\theta) \) is the model-specific prior distribution for the probability parameters \( \theta \) over the parameter space \( \Theta_1 \). Intuitively, this integral averages the likelihood for a response pattern \( \mathbf{x} \) weighted by the prior \( \pi_1 \) (Lee & Wagenmakers, 2013, ; Ch. 7). Hence, the marginal probability becomes larger if the prior puts more weight on ‘appropriate’ parameters (i.e., those with a high likelihood). Similarly, it becomes smaller if the prior puts most weight on ‘inappropriate’ parameters (i.e., those with a small likelihood). The BF is defined as the ratio of the marginal probabilities of model \( \mathcal{A}_1 \) over model \( \mathcal{A}_0 \) (Kass & Raftery, 1995). Essentially, this is the factor by which the prior odds for model \( \mathcal{A}_1 \) over \( \mathcal{A}_0 \) need to be updated to obtain posterior odds. In the present case, the BF simply quantifies the evidence that an observed response pattern \( \mathbf{x} \) emerged from strategy A vs. strategy B.

Lee (2016) proposed relying on the BF instead of MDL for strategy classification, arguing that the Bayesian framework, which requires the precise specification of a prior distribution \( \pi_1 \) for the error probabilities, bears several advantages. For instance, prior predictive distributions allow illustrating probabilistic strategy predictions graphically, and posterior probabilities provide the amount of evidence in favor of the selected strategy per participant. Based on the theoretical view that all admissible combinations of error probabilities are equally likely, Lee (2016) assumed uniform priors for all decision strategies. For instance, in the case of WADDprob, this means that the prior density function \( \pi_{\text{WADDprob}} \) is constant but that the parameter space \( \Theta_1 \) is truncated to allow only those probabilities \( \theta = (b_1, b_2, b_3, b_4) \) of choosing Option B that satisfy the order constraint \( b_1 \leq b_2 \leq b_3 \leq 1 - b_4 \leq 0.50 \). In a similar way, the remaining decision strategies in Fig. 1 alter the admissible parameter space \( \Theta_1 \), while treating the prior density \( \pi_1 \) as a constant.

Lee (2016) computed the BF by means of the software JAGS (Plummer, 2003) based on Markov chain Monte Carlo sampling. Essentially, this approach uses a discrete indicator variable that switches between the competing models (Lodewyckx et al., 2011). The method is conceptually clear, simple to implement, and can directly be extended to further decision strategies such as TTBprob. However, depending on the models, the indicator variable may slowly converge (as indicated by infrequent switches between models; Lodewyckx et al., 2011), thus requiring substantial computing time. Whereas this was not an issue for the models analyzed by Lee (2016), the convergence of JAGS was rather slow in the present scenario for the seven decision strategies in Fig. 1, resulting in a substantial variability of the estimated marginal probabilities (Heck, Overstall, Gronau, & Wagenmakers, submitted for publication). To increase precision and speed, we integrated the marginal probability in Eq. (3) directly using numerical methods (Heck & Wagenmakers, 2016). Technical details and an implementation in R are provided in the supplementary material available at the Open Science Framework (https://osf.io/jcd2c).

### 3.3. Determining the number of trials

Both NML and the BF take the reduced complexity of order-constrained models into account. In fact, the two methods are also related theoretically, since model selection with NML is asymptotically identical to model selection with the BF if a specific, noninformative prior is used (i.e., Jeffreys’ prior; Grünwald, 2007). Moreover, when sample size increases towards infinity, both methods will select the data-generating model with probability one if it is included in the set of competing models (Myung, Balasubramanian, & Pitt, 2000). However, in finite samples, the two methods can lead to different model-selection results (Heck, Wagenmakers, & Morey, 2015). In case of the decision strategies compared by Hilbig and Mosshagen (2014), the BF resulted in more WADD than WADDprob classifications compared to the original analysis with MDL (Lee, 2016). From a substantive perspective, such a divergence—although limited to the same theoretical ‘group’ of decision strategies—is not entirely satisfactory and raises the question which of the two methods recovers the data-generating strategies better.

As a remedy, simulations can be used to estimate the extent to which NML and the BF produce equivalent and unbiased classifications in finite samples. Moreover, simulations are also useful to assess the informativeness of the specific experimental design (Lee, in press), that is, how well the competing decision strategies in Fig. 1 can be teased apart in actual data as would be realistically produced in an experiment. Specifically, if the number of individual responses is very small, both NML and the BF tend to prefer parsimonious models over more complex models, because the data cannot provide sufficient evidence that warrants the additional complexity.

To address these issues, we relied on a simulation to determine the number of choices that is required for reliable model selection with either NML or the BF using the new item types in Table 1. For a given number of choices \( n \) per item type, we generated response patterns for each decision strategy and then used NML and the BF to select the best of the seven strategies in each replication. Beyond the number of trials, the accuracy of both model-selection methods depends on the numerical values for the data-generating parameters. For instance, if TTBprob is operationalized with only small differences in the error probabilities, this data-generating setting might be too close to the deterministic version of TTB and thus lead to misclassification in small samples. In the simulation, we therefore used parameter values that resembled substantively relevant instances of the data-generating strategies (e.g., differences in order-constrained error probabilities ranged between 0.10 and 0.20; see supplementary material for details).

This simulation showed that \( n = 40 \) trials per item type are sufficient to ensure recovery rates above 90% for all strategies and both NML and the BF. Moreover, at this high level of precision, both model-selection methods resulted in identical classifications in 97.6% of the replications. With respect to the remaining discrepancies, misclassifications between any variant of
TTB and any variant of WADD were almost nonexistent, showing that the method allows for a reliable discrimination between noncompensatory and compensatory decision strategies. Moreover, concerning the different (probabilistic versus deterministic) variants of each strategy, misclassifications between TTBprob and TTB on the one hand, and WADDprob and WADD on the other hand were very rare. This shows that the proposed modeling framework allows teasing apart deterministic and probabilistic versions of the same decision strategy using a realistic number of trials. In sum, \( n = 40 \) trials per item type are sufficient to ensure high levels of precision and very similar classification results by NML and the BF.

4. Experiment

4.1. Method

One-hundred and four participants (82% female, aged 18 to 43 years, \( M = 22.8, \text{SD} = 4.0 \)) were recruited from a local subject pool at the University of Koblenz-Landau. Similar to the procedure used by Hilbig and Moshagen (2014), participants were asked to repeatedly judge the superior of two fictitious products (choice options) based on positive and negative ratings of four testers (cues). The instructions explicitly stated that the four testers gave correct judgments with probabilities of 90%, 80%, 70%, and 60% (cue validities), respectively, and that a completely uninformative tester would perform at 50% (chance level). Participants were asked to make as many correct judgments as possible and provided with feedback after completion of the 160 trials (40 for each item type in randomized order). Note that each of the four item types was presented to the participants in the form of several distinct cue patterns that match the strategy predictions in Table 1 (the supplementary material provides the complete table of cue patterns used). Feedback was based on the Bayesian solution (Lee & Cummins, 2004).

4.2. Strategy-classification results

Table 2 shows that a majority of participants were classified as using a variant of WADD (77.9% based on both NML and the BF), whereas a minority was classified as using a variant of TTB (8.7%), which replicates previous results (Hilbig & Moshagen, 2014). Within the subset of 81 users of either variant of WADD, the probabilistic version WADDprob was selected for 48 and 33 participants according to NML and the BF, respectively. This shows that both the probabilistic and the deterministic version of WADD explain choice behavior of the vast majority of participants. However, the results do not allow for drawing strong conclusions on whether a the probabilistic or deterministic model variant of WADD is superior: NML classified a majority of WADD users as WADDprob, whereas the BF classified a majority as WADD. Within the subset of TTB users, more participants were classified by the deterministic than the probabilistic variant (i.e., for six and seven of nine TTB users based on NML and the BF, respectively). Overall, only few participants were classified as TTBprob users (2.9% and 1.9%, respectively).

To assess decision behavior in more detail, Fig. 2 shows the individual relative frequencies of choosing Option B for each of the four item types, with separate panels for subsets of participants according to strategy use as classified by NML. Note that 82.7% participants were consistently classified by NML and the BF (shown by red points), whereas a few participants were classified differently (blue triangles). Before discussing this discrepancy in the next section, the patterns of choice frequencies in Fig. 2 are assessed descriptively.

Overall, the individual patterns of choice frequencies closely matched the predictions of the decision strategies. All participants classified as WADDprob users by NML preferred Option A for the first three item types and, with the exception of 9 participants, preferred Option B or were indecisive for Item Type 4. This high choice variability in Item Type 4 can most plausibly be attributed to the small difference in log-odds \( L_{\text{A vs. B}} = -0.04 \) both relative to the other item types and in absolute terms, resulting in near-chance responding. In contrast, participants classified as users of the deterministic version of WADD showed much less choice variability and consistently chose the predicted option.

Concerning the noncompensatory strategies, participants classified as users of any TTB version almost always preferred Option A. The only exception was observed for one participant classified as TTBprob user and Item Type 3, for which only the least valid cue differed between the two choice options. Moreover, choices were highly consistent for the deterministic TTB but showed larger variability for the probabilistic TTBprob. The choice frequencies of the only participant classified as EQW closely matched the deterministic predictions for Item Type 1, 2, and 4 but showed substantial choice variability for Item Type 2, in which EQW needs to guess as both options yield the same number of positive cue values. Finally, the 13 participants left unclassified by NML (i.e., were classified as baseline) had very diverse and inconsistent choice patterns. Even though some features of the substantively predicted patterns can be recognized visually (e.g., preferences for A1 and A3 but choice variability for Item Type 2 as predicted by EQW), the choice frequencies across all four item types were still best explained by the baseline model according to NML and the BF. This highlights the necessity to constrain error probabilities across item types a priori instead of merely counting the number of choices that adhere to a strategy's predictions (Bröder & Schiffer, 2003a; Moshagen & Hilbig, 2011; Regenwetter & Robinson, in press).

Concerning the slight deviations observed in Fig. 2, the inclusion of the saturated baseline model ensured that models of substantive interest were only selected when they fitted the data sufficiently well in absolute terms (Hilbig & Moshagen, 2014). Indeed, despite the large complexity of the baseline model, Table 2 shows that several participants (12.5%) were not classified as users of any of the substantive decision strategies. To illustrate that the inclusion of the baseline model leads...
to similar results as goodness-of-fit tests, we also computed posterior-predictive $p$-values based on the $T_1$ statistic (Klauer, 2010), which quantifies the discrepancy between observed and posterior-predicted choice frequencies (similar to Pearson's $X^2$). According to the criterion $p > 0.05$, WADDprob and WADD had a good fit for 84 and 49 of 104 participants, respectively, compared to only 9 and 7 participants for TTBprob and TTB, respectively. Moreover, when including only those substantive strategies in model selection that met the $p > 0.05$ criterion, the classification by NML was almost identical as in Table 2 (except for one WADD user with $p = 0.038$, then classified as baseline). For the BF, the choice patterns of six participants classified as WADD indicated misfit with posterior-predictive $p$-values ranging from 0.009 to 0.042 and were thus clas-

<table>
<thead>
<tr>
<th>NML</th>
<th>WADDprob</th>
<th>WADD</th>
<th>TTBprob</th>
<th>TTB</th>
<th>EQW</th>
<th>baseline</th>
<th>(Total NML)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WADDprob</td>
<td>33</td>
<td>13</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>WADD</td>
<td>–</td>
<td>33</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>33</td>
</tr>
<tr>
<td>TTBprob</td>
<td>–</td>
<td>–</td>
<td>2</td>
<td>1</td>
<td>–</td>
<td>–</td>
<td>3</td>
</tr>
<tr>
<td>TTB</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>6</td>
<td>–</td>
<td>–</td>
<td>6</td>
</tr>
<tr>
<td>EQW</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>1</td>
<td>1</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>Baseline</td>
<td>–</td>
<td>2</td>
<td>–</td>
<td>11</td>
<td>–</td>
<td>–</td>
<td>13</td>
</tr>
<tr>
<td>(Total BF)</td>
<td>33</td>
<td>48</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>–</td>
<td>104</td>
</tr>
</tbody>
</table>

Note. No participant was classified as GUESS. Classification rates of zero are replaced by dashes for better readability.

<table>
<thead>
<tr>
<th>Item Type</th>
<th>Relative frequency: Option B</th>
</tr>
</thead>
<tbody>
<tr>
<td>WADDprob</td>
<td>$b_1 \leq b_2 \leq b_3 \leq 1 - b_4 \leq 0.5$</td>
</tr>
<tr>
<td>TTBprob</td>
<td>$b_2 = b_4 \leq b_1 \leq b_3 \leq 0.5$</td>
</tr>
<tr>
<td>EQW</td>
<td>$b_1 = b_3 = 1 - b_4 \leq b_2 = 0.5$</td>
</tr>
<tr>
<td>baseline</td>
<td>$b_1, b_2, b_3, b_4 \leq 1$</td>
</tr>
</tbody>
</table>

Fig. 2. Relative frequencies of choosing Option B for the four item types in Table 1. Across the six panels, participants are classified as users of a decision strategy based on NML. Red points and blue triangles indicate individual choice frequencies for which the BF resulted in identical or different strategy classifications, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
sified differently (three as baseline and three as WADDprob). Overall, these results show that the inclusion of the saturated baseline model can replace additional (classical or Bayesian) goodness-of-fit tests, even though the exact threshold for a ‘good model fit’ might slightly differ (especially given an arbitrary cutoff such as \( p > 0.05 \)). Importantly, however, the proposed model-selection approach has the advantage that a coherent, statistically principled method is applied (i.e., model selection by NML or the BF) instead of a two-step procedure.

4.3. Comparison of NML and the BF

Table 2 shows that the classification by NML and the BF lead to similar substantive conclusions—a majority of participants relied on the compensatory strategies WADDprob or WADD and only a minority on the noncompensatory strategies TTB or TTBprob. To assess the remaining discrepancies in more detail, Fig. 2 uses blue triangles to highlight the choice frequencies of participants classified differently by the two model-selection methods. The largest discrepancy concerns 13 participants (12.5% of the total sample) that were classified as WADDprob users by NML but as WADD users by the BF. Note that Lee (2016) found a similar trend in the classification frequencies of WADDprob/WADD for the data of Hilbig and Moshagen (2014) (32/27 for NML compared to 44/24 for the BF, respectively).

Concerning the 13 inconsistent classifications between WADDprob and WADD, the choice patterns \( x_i \) of choosing Option B fell into two qualitative classes. Four of these participants exhibited choice patterns of either \( x_B = (0.0.0.38) \) or \( x_B = (0.0.1.38) \) and thus satisfied the order constraint of WADDprob descriptively. Nevertheless, according to the BF, these patterns were 46.1 and 24.0 times more likely under WADD, respectively (whereas differences in NML in favor of WADDprob were small: \( \Delta \text{NML} = 0.61 \) and 0.08, respectively). Qualitatively, these patterns include more extreme frequencies than those classified consistently as WADDprob. However, the other nine participants classified as WADDprob by NML but as WADD by the BF preferred Option B2 more often than Options B3, resulting in a descriptive violation of the WADDprob prediction \( b_2 < b_3 \). Despite this descriptive mismatch, WADDprob was still selected by NML with 0.40 < \( \Delta \text{NML} \leq 2.13 \), whereas the BF had a tendency to prefer the deterministic WADD version with 1.29 < \( \Delta \text{BF} \leq 4.87 \) (except BF = 21.15 for one participant). Moreover, as visible by the peak at Item Type 2 in Fig. 2, these inconsistently classified patterns such as \( x_B = (0.5.0.36) \) differed from the remaining ones in including very few Option B1 or B3 choices. Overall, these qualitative comparisons suggest that NML and the BF apply a different weight when choice frequencies violate the order constraints of WADDprob, which may result in diverging classifications especially when frequencies are at the boundary of the data space (e.g., if an option is never or always chosen).

To test the robustness of the classification results, the relative evidence in favor of the competing strategies can be inspected. For this purpose, we computed NML model weights on the one hand \( \omega_i^{\text{NML}} = \exp(-\text{NML}_i)/\sum_i \exp(-\text{NML}_i) \) and model posterior probabilities based on the BF assuming equal prior odds on the other hand \( p(\mathcal{M}_i|x) = p(x|\mathcal{M}_i)/\sum_j p(x|\mathcal{M}_j); \) Heck et al., 2015). Both measures are computed for each participant and strategy and facilitate a direct comparison of the two model-selection methods. When requiring a lower threshold of 75% evidence in favor of the selected strategy (compared to the priori probability of 1/7 ≈ 14.3%), NML and the BF still classified 62 and 84 of the participants, respectively. Importantly, the pattern of strategy classifications remained robust for this subset of participants and resulted in 56% WADDprob, 24% WADD, 2% TTBprob, 6%, TTB, 2% EQW, and 10% unclassified users based on NML (37%, 46%, 0%, 7%, 1%, and 8% based on the BF, respectively).

For a more detailed comparison of the two model-selection methods, Fig. 3 shows NML weights and posterior probabilities for all participants. Across rows, each panel shows the relative evidence in favor of one of the competing strategies (GUESS is omitted since all values were close to zero). Moreover, strategies selected by NML are highlighted with gray background. In line with the result that NML and the BF classify most response patterns identically, NML weights and posterior probabilities were often similar or almost identical. Importantly, for participants classified as users of a compensatory strategy (WADDprob, WADD), neither method provided evidence in favor of TTB or TTBprob. Similarly, for participants classified as TTB or TTBprob, both methods indicated no evidence in favor any compensatory strategy.

In contrast to these similarities, the BF indicates a stronger preference in favor of the deterministic strategies WADD, TTB, and EQW than NML for almost all participants, whereas NML often assigns higher weights to probabilistic strategies. Thus, in the trade off between model fit and complexity, the BF weighs the parsimony of the strategies in Fig. 1 more strongly than NML. This difference in penalizing flexible models is in line with the observation of Lee (2016, p. 40) that ‘the Bayesian classification remains with the simpler model that the data are most consistent with.’ Moreover, the higher penalty of the BF for model complexity drives the inconsistencies in the discrete NML and BF classifications in Table 2, and also explains why choice patterns classified as WADD by the BF more often indicated a misfit according to the posterior-predictive \( p \)-values. However, despite these discrepancies for a minority of choice patterns, Fig. 3 shows that NML and the BF lead to very similar estimates of the relative evidence in favor of the competing strategies, even though the two methods are not guaranteed to result in identical classifications in small samples.

5. Discussion

Previous applications of outcome-based strategy classification assumed that TTB is applied with a constant error probability. Based on a cognitive-process interpretation of heuristics, we proposed a psychologically plausible version of TTB
assuming that more processing steps lead to higher error probabilities. To the best of our knowledge, this is the first variant of TTB that does not assume a constant probability of strategy application errors. Thereby, we follow the crucial principle that statistical models need to reflect psychological theory closely to allow for a fair empirical test (Jekel & Glöckner, in press). Importantly, the new model can directly be tested against competing strategies (both deterministic or probabilistic) within the same framework. In principle, our modeling approach for TTBprob can be generalized to other heuristics that predict serial processing steps. Also note that in scenarios where cue validities are unknown to participants, TTB could be modeled by assuming probabilistic cue orders, whereas the strategy TTBprob assumes probabilistic cue comparisons (Mistry, Lee, & Newell, 2016). However, in the present experiment, where cue orders are directly accessible due to the presentation of exact validities, it is more plausible to model TTB by a fixed order of serial, probabilistic cue comparisons.

Replicating a previous study (Hilbig & Moshagen, 2014), most participants were classified as users of WADD, which integrates all cues weighted by validity and thereby approximates (substantively) rational decisions (Lee & Cummins, 2004; Simon, 1976). Approximately half of the WADD users were better described by a deterministic version with a constant error probability, whereas the other half was better described by a probabilistic, psychologically plausible WADDprob version. The latter model assumes that strategy-inconsistent errors are more likely if the outcome of the cue integration process is more ambiguous, that is, if the cues do not provide clear evidence for one option. This error theory is implied by axiomatic decision theories (Luce, 1959) and process models (Glöckner & Betsch, 2008a), whereas the deterministic version of WADD assumes that errors are unsystematic and thus independent of relative preferences. However, the conclusion that a majority of participants is best described as WADD users is limited to inferences from givens with cues being available during decision making. Other paradigms such as memory-based decisions are likely to impair information search, integration, or both (Bröder & Schiffer, 2003b). Given that the effort-reduction capabilities of heuristics such as TTB or EQW become more relevant in such scenarios (Glöckner & Betsch, 2008b; Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015), our result cannot be generalized to other paradigms and domains.

Within the subset of 9 participants classified as TTB users, only three were classified by NML as users of the new, probabilistic version (only two according to the BF). Given its derivation based on psychological theory, this inferior performance of TTBprob is surprising. There are several reasons that could explain the low proportion of TTBprob users. With only one to four cue comparisons, the item types in Table 1 might not have required a sufficient number of processing steps to affect the error probabilities to a meaningful degree. Similarly, processing steps might simply not be sufficiently error-prone in inferences from givens (a neatly pre-structured, simplified and fully available information board; cf. Söllner, Bröder, & Hilbig, 2013) to produce any measurable increase in error probabilities. Therefore, despite the low proportion of TTBprob users in our study, the probabilistic version of TTB should still be included in future applications in which heuristic decisions are expected to be more likely.

A general limitation of the TTBprob model might be its increased complexity relative to the deterministic TTB version, which assumes a constant error probability for all item types (cf. Fig. 1 for the NML complexity terms). However, even

![Fig. 3. Posterior model probabilities based on the BF (assuming equal prior odds) and NML model weights for each of the strategies. Gray background color shows the classification by NML.](image-url)
though TTBprob is indeed penalized for its higher complexity, it can also fit many more choice patterns that are psychologically plausible and predicted by the adaptive-toolbox theory (which defines TTB as a strictly serial process). Given that both NML and the BF balance model fit and flexibility based on statistical principles (Myung, 2000), TTBprob cannot be discarded merely on grounds of its increased complexity. Moreover, the choice patterns of participants classified by the deterministic TTB version did not follow the order of error probabilities predicted by the sequential-cue-search hypothesis (i.e., only very few participants showed higher frequencies of choosing Option B for Item Types 1 and 3 than for Item Types 2 and 4). Given the absence of any trend in favor of the predicted pattern, a nested, more constrained version of TTBprob is unlikely to outperform the TTB version with constant errors. Put differently, the TTBprob strategy performed poorly because TTB errors did not increase whenever the most valid cue was not the first to discriminate between options. Thus, our results provide evidence against a specific serial processing mechanism, thereby narrowing down psychologically plausible conceptualizations of TTB within the adaptive-toolbox theory—much like the empirically successful extension of the deterministic WADD to a probabilistic version implies that compensatory cue-integration should be viewed through the lens of process models such as evidence accumulation. Whereas the deterministic WADD version often implemented in prior research is not sufficient to model error probabilities determined by the underlying process, the deterministic TTB version is a good approximation compared to a specific, psychologically plausible TTB version.

In sum, we proposed a psychologically plausible version of TTB that can directly be tested against a deterministic model of TTB and other strategies such as WADD (both deterministic and probabilistic) within a coherent modeling framework. Building on our empirical result that only a few participants used TTBprob in inferences from givens, the new modeling approach allows for testing under which conditions the hypothesized serial processing steps underlying TTB have a measurable impact on observable decisions.

5.1. Methodological considerations

Model-selection methods such as NML or the BF ensure that models are appropriately penalized for their increased statistical complexity. This is especially important for models with the same number of parameters that nevertheless differ in complexity (e.g., WADDprob and the baseline model). Despite the conceptual and philosophical differences between NML and the BF, both methods lead to similar classifications for the new item types. A noteworthy difference concerns the fact that the BF with uniform priors on the error probabilities penalizes model complexity of probabilistic strategies more strongly than NML as indicated by smaller posterior probabilities compared to NML model weights, an observation in line with the results of Lee (2016). Overall, however, for the item types in Table 1 and the corresponding strategies in Fig. 1, NML results are likely to be similar to those obtained by the BF for realistic samples sizes of $n = 40$ choices per item type. This similarity is important for substantive applications of generalized strategy classification that implement only a single model-selection method, since it ensures that conclusions do not hinge on a specific model-selection criterion.

Concerning the question which method should be preferred in future applications, both NML and the BF come with unique advantages and limitations. Derived from the MDL principle, NML defines an explicit complexity term that measures a model’s fit to all possible data. As shown in Fig. 1, this intuitively plausible measure of complexity facilitates understanding of the competing decision strategies. For instance, from a philosophy-of-science perspective (Glöckner & Betsch, 2011), models should ideally make precise predictions. Within the MDL framework, the NML complexity term provides an intuitive measure to quantify this important property of theoretical decision strategies for specific experimental paradigms. Moreover, once the complexity terms have been computed, strategy classification reduces to the simple task of determining the maximum-likelihood value for each strategy. This computational advantage could be utilized for adaptive design optimization (Myung, Cavagnaro, & Pitt, 2013), which selects the most informative items during the experiment to increase the precision of strategy classification. Finally, NML does not require the specification of exact and possibly difficult-to-derive prior distributions for the error probabilities.

In contrast to NML, the BF has the advantage that it provides an intuitive metric in terms of the relative likelihood of two competing models given the data (even though equivalent quantities can be computed for NML). Moreover, the BF allows researchers to explicitly assume priors on the error probabilities and thereby draw inferences that better reflect psychological theory (Vanpaemel, 2010). For instance, more prior weight can be assigned to smaller as compared to larger error probabilities to constrain decision strategies more strongly; alternatively, assumptions about the dependence of choice probabilities across different sets of alternatives can be encoded (McCausland & Marley, 2013). In NML, the relative weight of error probabilities is determined implicitly by information-theoretic principles and therefore more difficult to assess for plausibility. However, current theories of judgment and decision making usually do not predict exact prior distributions for the error probabilities, thus Lee (2016) assumed noninformative uniform priors. The importance of choosing theoretically appropriate priors becomes clear when considering that some strategies can approximate other strategies when using specific, theoretically implausible priors. For instance, if WADDprob is equipped with a precise prior that heavily outweighs indistinguishable error probabilities $b_1 \approx b_2 \approx b_3 \approx 1 - b_4 \leq 0.50$, WADDprob becomes highly similar to the deterministic version of WADD. Similarly, all substantive decision strategies approximate GUESS if almost all prior probability is assigned to error probabilities close to 0.50. Note that such implausible priors will affect model selection even in arbitrarily large samples. Thus, although it is cer-

---

3 If desirable, the ‘luckiness NML’ criterion allows for explicitly including prior knowledge about parameters into NML (Grunwald, 2007).
tainly an advantage that one can specify prior distributions for the error probabilities, this also comes with the burden of doing so in a theory-consistent and reasonable way. This profound issue is not simply resolved by the fact that one can test different priors empirically by including the same strategy (e.g., WADDprob) with multiple prior specifications (Lee, 2016)—high theoretical precision will be necessary to select a promising and reasonable subset of competing prior distributions for such a test (in addition, recovery rates might diminish considerably). Unfortunately, since judgment and decision theories are insufficiently specified in this regard, choice of exact prior distributions is left to idiosyncrasies of the researcher.

A drawback of both methods concerns their applicability to new and especially more item types. As the number of item types and error probabilities increases, Monte Carlo integration is necessary to compute both NML (Klauer & Kellen, 2015) and the BF (Heck & Wagenmakers, 2016; Lee, 2016). However, the increase in required computing time may pay off by classification rates clearly above 90% as in our simulation—especially since performance is likely to improve as the number of informative item types increases. For the item types in Table 1, the supplementary material provides R code to facilitate the application of generalized strategy classification using NML or the BF.

5.2. Conclusions

Generalized outcome-based strategy classification provides an encompassing modeling framework in which competing decision strategies can directly be tested against each other. The original method by Bröder and Schiffer (2003a) included heuristics that reduce costs for information search (TTB) or information integration (EQW), a strategy approximating rational decisions (WADD), and pure guessing. Hilbig and Moshagen (2014) proposed including a psychologically plausible version of WADD assuming that error probabilities decrease when the evidence for the preferred option increases. We further extended the scope of the method to include psychologically plausible versions of serial-process strategies such as TTB assuming that error probabilities increase when more elementary processing steps are required. Two conceptually different model-selection techniques (NML and the BF) allow for reliably classifying participants as users of the competing strategies using a realistic number of trials. Overall, this shows that generalized strategy classification has evolved as a powerful and general method for investigating decision making under uncertainty.

Author note

The first author was supported by the Grant Er 224/2-2 from the Deutsche Forschungsgemeinschaft (DFG) and the University of Mannheim’s Graduate School of Economic and Social Sciences, also funded by the DFG.

Data and R code for the analysis are available at the Open Science Framework at https://osf.io/jcd2c

References


