Psychological Review
2013, Vol. 120, No. 1, 39–64

0033-295X/13/$12.00 DOI: 10.1037/a0030777

Testing Adaptive Toolbox Models: A Bayesian Hierarchical Approach

Benjamin Scheibehenne and Jörg Rieskamp
University of Basel

Eric-Jan Wagenmakers
University of Amsterdam

Many theories of human cognition postulate that people are equipped with a repertoire of strategies to solve the tasks they face. This theoretical framework of a cognitive toolbox provides a plausible account of intra- and interindividual differences in human behavior. Unfortunately, it is often unclear how to rigorously test the toolbox framework. How can a toolbox model be quantitatively specified? How can the number of toolbox strategies be limited to prevent uncontrolled strategy sprawl? How can a toolbox model be formally tested against alternative theories? The authors show how these challenges can be met by using Bayesian inference techniques. By means of parameter recovery simulations and the analysis of empirical data across a variety of domains (i.e., judgment and decision making, children's cognitive development, function learning, and perceptual categorization), the authors illustrate how Bayesian inference techniques allow toolbox models to be quantitatively specified, strategy sprawl to be contained, and toolbox models to be rigorously tested against competing theories. The authors demonstrate that their approach applies at the individual level but can also be generalized to the group level with hierarchical Bayesian procedures. The suggested Bayesian inference techniques represent a theoretical and methodological advancement for toolbox theories of cognition and behavior.

Keywords: cognitive strategies, mixture models, Bayesian statistics, Bayes factor

Supplemental materials: http://dx.doi.org/10.1037/a0030777.supp

In line with the classic proverb that many roads lead to Rome, people may use various strategies or cognitive tools to deal with the challenges they face in daily life. The idea that there is usually more than one strategy to reach a single goal was described by the late psychologist Egon Brunswik as *vicarious functioning,* meaning that "there is a variety of 'means' to each end" (Brunswik, 1952, p. 18). The notion that people can choose among different strategies within a cognitive toolbox allows one to explore why different people may approach the same task in different ways and provides a fruitful basis for understanding variations in behavior across time and situations (Einhorn, 1970).

Toolbox Models Are Widely Used

The concept of a cognitive toolbox can be traced across various theories in psychology and related fields. For instance, in linguistics, it has been argued that people use different strategies for recognizing words based on semantic context (Eisenberg & Becker, 1982). Developmental psychologists have found that children use multiple strategies when solving mathematical exercises, balance-scale problems, or memory tasks (Coyle, Read, Gaultney, & Bjorklund, 1998; Lemaire & Siegler, 1995). In the social domains, researchers have argued that people rely on different strategies for social interactions (Erev & Roth, 2001; Fiske, 1992; Milinski & Wedekind, 1998), mating choices (Buss & Schmitt, 1993), and predicting other people's behavior (Costa-Gomes & Crawford, 2006). Furthermore, researchers have argued that people use different strategies for categorization (Busemeyer & Myung, 1992; Patalano, Smith, Jonides, & Koeppe, 2001; Sewell & Lewandowsky, 2011), resource allocation (Ball, Langholtz, Auble, & Sopchak, 1998), estimation and frequency judgments (Brown, 1995; Brown, Cui, & Gordon, 2002; von Helversen & Rieskamp, 2008), skill acquisition (Anderson & Lebiere, 1998; Lovett, 1988), function learning (Lewandowsky, Kalish, & Ngang, 2002), and learning processes (Erev & Barron, 2005; Gigerenzer & Gaissmaier, 2011). The idea that processes such as information search or choice are guided by qualitatively different strategies also figures prominently in research on judgment and decision making. For example, Payne, Bettman, and Johnson (1988) argued that "a decision maker possesses a repertoire of well-defined strategies and selects among them when faced with a decision" (p. 550; see also Payne, Bettman, & Johnson, 1993). Likewise, the heuristics-and-biases program (Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1981) follows the assumption that people are equipped with a set of heuristics or simple rules of thumb.

This article was published Online First December 3, 2012.

Benjamin Scheibehenne and Jörg Rieskamp, Department of Psychology, University of Basel, Basel, Switzerland; Eric-Jan Wagenmakers, Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands.

We would like to thank Brenda Jansen, Stephan Lewandowsky, Lee-Xieng Yang, and Han van der Maas for providing us with the data of their experiments; Jet Tang for his help with the programming; and Anita Todd for editing the manuscript. This work was supported by SNSF Research Grant 100014_130149 to Benjamin Scheibehenne and Jörg Rieskamp.

Correspondence concerning this article should be addressed to Benjamin Scheibehenne, Department of Psychology, University of Basel, Missionsstrasse 62a, 4055 Basel, Switzerland. E-mail: benjamin.scheibehenne@unibas.ch

Alternative Models of Cognition

In contrast to the toolbox approach, theory building in psychology may also follow the idea of a single comprehensive model to describe human cognition. This single-model approach does not assume qualitatively different cognitive processes or changes in strategies. Instead, behavioral variations and individual differences within a given task are captured through free parameters, thus retaining the basic notion of a single model that can be broadly applied. Examples of such models are plentiful. In the realm of judgment and decision making, alternatives to toolbox models include sequential sampling models (Busemeyer & Townsend, 1993; Lee & Cummins, 2004; Newell, 2005; Newell & Lee, 2011; Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2001, 2004), exemplar models (Juslin & Persson, 2002; Nosofsky, 1991), and neural network models (Glöckner, Betsch, & Schindler, 2010; Gluck & Bower, 1988), to name only a few.

Difficulties When Testing Toolbox Models

The idea of strategy toolboxes has had its share of criticism. An important point of contention is that toolbox models as a whole are difficult to falsify and it is not always clear how they can be tested against alternative models of cognition (e.g., Bröder, 2000; Dougherty, Franco-Watkins, & Thomas, 2008; Hilbig, 2010; Newell, 2005; Todd & Gigerenzer, 2001).

One of the reasons for this difficulty with falsification is that a criterion is needed to decide on the number of strategies that belong to a particular toolbox. Including more strategies increases the toolbox's flexibility and consequently yields a better fit to the data. This may create a loophole for researchers who wish to immunize their toolbox model against falsification because one can always add another tool to capture the observed behavior. This problem was also noted by Glöckner et al. (2010), who criticized toolbox models for being flexible storage devices that provide unlimited space for additional strategies. In general, the more strategies are included, the higher the risk that one of the strategies provides a good description of the observed data merely by chance, so that the improved fit stems from fitting random noise (Domingos, 1999; Myung, 2000). Thus, even if each single strategy in the toolbox is rather simple, including many such strategies still results in a highly complex and flexible model that is difficult to test and falsify empirically—this is the *strategy sprawl problem.* In the extreme case, a sprawl of simple toolbox strategies always provides a superior description of the data by assuming a specific strategy for each person and situation. In contrast, an increase in complexity is worthwhile if it leads to new insights. Therefore, restricting the repertoire a priori to only a few strategies would defeat the whole purpose of describing intra- and interindividual differences through qualitatively different processes.

To formalize and guide this balancing act, a methodological procedure is needed that can quantify the trade-off between a toolbox's flexibility and its descriptive adequacy (Pitt & Myung, 2002). Here, we show how the Bayesian formalism allows toolbox approaches to be rigorously tested. This rigorous evaluation procedure is crucial to overcome the strategy sprawl problem and thereby advance the toolbox approach as a testable competitor to alternative accounts of cognition. In the following, we lay out this Bayesian approach and illustrate its advantages by means of concrete applications in four different research domains.

The remainder of this article is structured as follows: In Part I, we outline and specify the theoretical basis of a Bayesian toolbox approach. Using simulated data and model recovery studies, we further illustrate how this approach can be fruitfully applied to individual-level data that resemble the general structure of many empirical studies conducted to test toolbox models. In Part II, we apply the framework to various empirical data from four research domains: judgment and decision making, children's cognitive development, function learning, and perceptual categorization. In these seemingly unrelated areas, we compare toolboxes (i.e., mixtures of strategies) against single strategies and alternative cognitive models on the individual level. Following this, we provide an example of how to tackle the problem of strategy sprawl based on an enlarged toolbox. Finally, in Part III, we extend the approach to the group level by applying hierarchical Bayesian techniques that coalesce data from multiple individuals. Here, we start again from simulations before progressing to the analysis of empirical data.

Part I: A Bayesian Approach to Testing Toolboxes

Preventing strategy sprawl when testing and comparing a toolbox requires a trade-off between the increased complexity introduced through additional strategies and the expected increase in explanatory power for the observed data. Here, the Bayesian method provides a unifying comparison metric known as the Bayes factor (BF; Jeffreys, 1961; Kass & Raftery, 1995) that quantifies the extent to which the data support one model over another, taking model complexity into account. Its interpretation has intuitive appeal as it indicates how much one should shift one's beliefs in each model based on the observed data. For example, when comparing any two models M_1 and M_2 (i.e., a toolbox and a single-strategy model), a $BF_{1,2}$ of 10 indicates that the observed data are 10 times more likely to have occurred under M_1 than under M_2.

The BF can be obtained by comparing the marginal likelihoods of each of the models under consideration. Conceptually, the marginal likelihood measures the average quality of the predictions that a model M_k has made for the observed data D. The better the predictions, the greater the evidence in favor of M_k. To determine how well a model predicted the observed data, we need to take into account *all* predictions that the model made and weight these by their prior probability. Statistically, this is accomplished by averaging the likelihood of the observed data D across all possible parameter values θ_k of a model M_k, weighted by the prior probability of θ_k (e.g., Myung & Pitt, 1997; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010):

$$p(D \mid M_k) = \int p\left(D \mid \theta_k, M_k\right) \times p(\theta_k \mid M_k)\, d\,\theta_k, \qquad (1)$$

where $p(\theta_k \mid M_k)$ represents the prior probability distribution of θ_k, and $p(D \mid \theta_k, M_k)$ is a likelihood function that represents the probability of the observed data D given θ_k.

The foregoing argument also shows that to obtain strong support from the data, a model needs to make a high proportion of good predictions. This is precisely the problem with models that are overly complex; such models have a relatively large parameter space, and although this enables them to make many predictions, a high proportion of these predictions will turn out to be false. Complex models need to distribute the prior probability across

their entire parameter space, and ultimately, a model that predicts almost everything has its prior probability spread so thin that the occurrence of any particular event will not greatly add to that model's credibility. This is the Bayesian justification for the adage "a model that predicts everything predicts nothing."

As described above, the marginal likelihood for a model M_k is calculated by averaging the likelihood $p(D \mid \theta_k, M_k)$ over the prior $p(\theta_k \mid M_k)$. When the prior is very spread out, it occupies a relatively large part of the parameter space in which the likelihood is almost zero (i.e., where the predictions are false), and this greatly decreases the average or marginal likelihood. Hence, the computation of marginal likelihood embodies a reward for parsimony, or an automatic Occam's razor (cf. Jefferys & Berger, 1992; Myung & Pitt, 1997).

It is well known that there is more to model complexity than the size of the parameter space. For instance, another important factor that influences model complexity is the functional form of the model parameters. Consider, for instance, two laws of psychophysics that relate the objective intensity I of a stimulus (e.g., a sound, a flash of light) to its subjective experience $\Psi(I)$. The first, Fechner's law, states that $\Psi(I) = k \times \ln(I) + \beta$: Experienced intensity is a negatively accelerating function of stimulus intensity. The second, Stevens's law, states that $\Psi(I) = k \times I^{\beta}$: Experienced intensity can be a negatively or positively accelerating function of stimulus intensity. Fechner's law and Stevens's law each have two parameters, k and β, but nonetheless, Stevens's law is more complex: It can capture more data patterns and is therefore more difficult to falsify than Fechner's law (cf. Myung & Pitt, 1997; Townsend, 1975).

In sum, a model is complex when it makes many predictions. This occurs when a model has many free parameters, when the prior distributions of those parameters are relatively broad, or when the parameters have a complicated functional form. Such complexity may or may not be warranted by the data. By assessing the average quality of a model's predictions, the marginal likelihood takes all of these considerations automatically into account.

Dividing the marginal likelihoods for models M_1 and M_2 yields $BF_{1,2}$:

$$BF_{1,2} = \frac{p(D \mid M_1)}{p(D \mid M_2)}. \quad (2)$$

Hence, the BF compares the predictive performance of one model to that of another, for the data at hand. Here, $BF_{1,2}$ indicates the extent to which the data support M_1 over M_2, and as such, it represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648). To obtain the marginal likelihood in Equation 1, the prior distribution of the models' parameter and likelihood functions must be specified. With respect to testing and comparing toolbox models of different sizes, this requires specifying the strategies and how decision makers select among them, illustrated next.

Formal Specification of a Cognitive Toolbox

In general, a toolbox model TB can be conceptualized as a set of different psychological processes or strategies f and the different parameters θ_f that may be associated with each of them. Each strategy f predicts a specific behavior contingent on these parameters depending on internal and external influences.

Example of a toolbox. The details of each strategy vary greatly across different research areas. In the categorization domain, for example, researchers may be interested in which strategies people use to categorize objects contingent on the objects' features. In decision making, people's strategies for making probabilistic inferences are often examined. Here, a common experimental task is to infer which of two options has a higher criterion value (e.g., Gigerenzer & Goldstein, 1996). Imagine a person who predicts which of two used cars will last longer by using cues that are probabilistically related to the criterion, such as the cars' mileage, the sound of the engines, or the accident histories. This person could apply a simple noncompensatory decision strategy, such as take the best (TTB), that focuses on only the most important or valid cue (Gigerenzer & Goldstein, 1996). If that cue does not discriminate, the second-most-important cue is considered, and so on until a decision can be made. Alternatively, this person could use a compensatory weighted-additive (WADD) strategy (e.g., Payne et al., 1988, 1993) that computes an overall score for each option by summing up its cue values multiplied by their respective importance weights or validities. The decision maker selects the option with the highest score. Because the compensatory strategy WADD takes all available information into account, it is commonly assumed to be cognitively more demanding than TTB (Czerlinski, Gigerenzer, & Goldstein, 1999; Payne et al., 1993). Yet another alternative is to make use of both strategies. In this case, no one strategy will always be best for predicting a person's choices. Instead, a decision maker's behavior will be better described by a toolbox consisting of both strategies.

Strategy selection. If a toolbox consists of more than one strategy, the question of how tools are selected from the toolbox must be addressed. This is an area of active research, as the cognitive processes that determine this selection might depend on various factors (Lee, 2011; Payne et al., 1988, 1993), such as environmental and situational influences (e.g., Marewski & Schooler, 2011; Newell, 2005), specific context cues (Lewandowsky et al., 2002), previous learning experience (e.g., Rieskamp, 2006; Rieskamp & Otto, 2006), or a person's cognitive abilities (e.g., Mata, von Helversen, & Rieskamp, 2011) or cognitive development (Siegler, 1994).

To develop an ecological theory of strategy selection, scholars have explored in which environments particular strategies work well and to what extent individuals can adaptively choose particular strategies depending on the requirements of the situation they face (Kruglanski & Gigerenzer, 2011; Marewski & Schooler, 2011; Simon, 1990; Todd, Gigerenzer, & the ABC Research Group, 2012). In the car-decision example outlined above, for instance, people might select TTB under high time pressure and WADD when time is not an issue (e.g., Rieskamp & Hoffrage, 2008).

On a general level, the outcome of this selection process can be expressed as a mixture proportion parameter β that indicates the probability of choosing each strategy in the toolbox. For a toolbox TB consisting of J strategies, each strategy f_j will be selected with a probability β_j, and because one strategy must be used, the βs must sum to 1 ($\sum_{j=1}^{J} \beta_j = 1$). Given this specification, the likelihood function for a toolbox can be specified based on the sum of the likelihoods of each included f_j, weighted by β_j:

$$p(D \mid TB) = \sum_{j=1}^{J} [\beta_j \times p(D \mid f_j)]. \quad (3)$$

If the cognitive process underlying the selection of strategies is not specified, the specific strategy mix requires empirical validation, and the value of β is estimated from the data. In this case, for a toolbox consisting of J strategies, a total of $J - 1$ β parameters are estimated beyond the free parameters of each single strategy. However, note that the β parameters are not independent of each other, as a high value for one implies a low parameter value for the others.

Priors. Implementing a toolbox within a Bayesian framework further requires specifying prior probabilities or probability distributions for all free parameters. These priors form an integral part of the model, and they are informed by theoretical considerations and possibly also by available prior knowledge. Selecting appropriate prior distributions is of ongoing concern to Bayesian statisticians (e.g., Kass & Wasserman, 1995; Liang, Paulo, Molina, Clyde, & Berger, 2008). In some cases, for example, if the parameter space is bounded, the absence of prior knowledge can be expressed though uniform distributions, indicating that all values within the predefined range are equally likely a priori.

Simulating Individual Data

In many research areas associated with the idea of cognitive toolboxes, experiments rely on repeated choices between two options or actions. In decision making, for example, people often have to decide between two consumer products, cities, or job candidates. Similarly, in categorization, people often have to classify objects into two potential categories. In the developmental literature, experimenters have asked children to predict whether a balance scale will tip either left or right. In all these domains, people can apply different strategies that vary in the way information is processed and in what actions are finally taken. In the examples above, these actions are commonly repeated choices between two options, categories, or functions. Comparing these choices to the predictions of a set of predefined rules or strategies provides the basis to decide which strategy best describes the observed data. To enhance discrimination between the candidate strategies, stimuli are commonly designed so that the strategies' predictions differ. Following this general layout, we simulated data where participants repeatedly chose among pairs of options that were carefully designed to discriminate between two generic strategies, labeled A and B. In the car example above, for instance, A and B would represent the noncompensatory strategy TTB and the compensatory strategy WADD, respectively.

Likelihood function. Individuals sometimes make implementation errors when using a particular strategy. For example, a child solving a physical problem may use Rule A but may sometimes make an error when using that rule. In this case, an answer that is not predicted by Rule A would be given because of that erroneous application and not the application of an alternative strategy.

To allow for the possibility of inconsistent choices or application errors, each strategy in the simulation contains an explicit error term, such that parameter ε indicates the probability that a decision is made at random. The probability that an error occurs is assumed to be constant across situations. This simple error theory is sometimes called tremble error (cf. Loomes, Moffat, & Sugden, 2002). Hence, each strategy makes a probabilistic prediction contingent on an unknown parameter value ε that has to be estimated from the data. In the simulation at hand, $\varepsilon = 0$ indicates that no inconsistencies exist and all decisions are in line with the strategy's deterministic prediction. An $\varepsilon = 1$ implies random choice or pure guessing, such that the probability that a pairwise choice is in line with the deterministic prediction is .50. If a single choice matches a strategy's deterministic prediction, then the predicted probability of that choice equals $1 - \varepsilon/2$; the probability equals $\varepsilon/2$.

For instance, in the car example above, TTB might predict the choice of one car (labeled X) and WADD the other (labeled Y). If a decision maker applies a TTB strategy with error probability $\varepsilon = .2$, the probability of observing the choice of X should be 0.9, and the probability of observing the choice of Y should be 0.1. In general, the likelihood of observing data (D) in line with a choice of X can be expressed as a Bernoulli distribution:

$$p(D \mid \mathrm{A}, \varepsilon_{\mathrm{A}}) = \left(1 - \frac{\varepsilon_{\mathrm{A}}}{2}\right)^{k} \times \left(\frac{\varepsilon_{\mathrm{A}}}{2}\right)^{N-k}, \tag{4}$$

where k indicates the number of choices that are consistent with the deterministic prediction made by Strategy A and N represents the total number of choices.

Based on this implementation of a single strategy, a toolbox $\mathrm{TB}_{\mathrm{A,B}}$ consisting of two strategies A_{ε} and B_{ε} can be set up similar to Equation 3 where β indicates the probability of applying A_{ε} over B_{ε}; $\beta = 1$ indicates that an individual will always apply A_{ε}, whereas $\beta = 0.8$, for example, indicates A_{ε} is selected in 80% of the cases and B_{ε} otherwise. For simplicity, we assume a common implementation error for all strategies in the toolbox. Specified this way, in the simulation at hand, the parameter space of $\mathrm{TB}_{\mathrm{A,B}}$ consists of $\varepsilon_{\mathrm{TB}}$ and β. In other contexts and if alternative strategies are considered, different parameter sets may determine the models' predictions.

Priors. To retain the general scope of the simulation, we assume that there is no strong prior knowledge such that all possible parameter values are equally likely a priori (i.e., uniform prior probability distributions are assigned to all parameters). As the parameters correspond to rates or probabilities, this appears to be a reasonable choice. If theoretical reasons or prior knowledge suggest further restrictions or skewed distributions, alternative specifications may be justified. We consider more complicated scenarios when analyzing empirical data sets later.

Data. In a first step, we generated a set of stimuli that allows one to distinguish between two strategies A_{ε} and B_{ε}. In particular, stimuli are generated such that the error-free predictions (i.e., $\varepsilon = 0$) of both strategies are independent. In other words, knowing that A_{ε} predicts a specific answer does not change the probability that B_{ε} also predicts that answer, and vice versa. Furthermore, care was taken to ensure that an erroneous application of any one of the strategies does not increase the probability that the other strategy is selected.[1] Note that the Bayesian method does not require uncorrelated model predictions or independent error terms, as we outline when analyzing actual empirical data later on. Controlling for these dependencies in the simulation merely served as a means to reduce uninteresting noise and to carve out the dependencies between the BF and the model parameters, which were of primary interest.

We generated a total of 80 stimuli as a basis for the simulation study; this number is in the ballpark of many experiments in psychology, and at the same time, it allowed us to generate many combina-

[1] The online supplemental materials provide the programming code used to generate the data and the code used for the simulation and the subsequent analyses.

tions that satisfied the specified requirements. The generated stimuli provided the basis for comparing a toolbox $TB_{A,B}$ against a single strategy (either A_ε or B_ε) for single synthetic individuals with known strategy use and implementation error.

Simulated data. We varied the mixture proportion β from 0 (exclusive use of just A_ε) to 1 (exclusive use of just B_ε) in steps of 0.05 between the simulated participants. Values between 0 and 1 resembled the genuine use of a toolbox; that is, for $\beta = 0.5$, a simulated participant probabilistically used A_ε half the time and B_ε otherwise. As a second factor, the application error ε was set to 0 (deterministic choice), 1 (random choice), or 0.5. An error of 0.5 indicated a medium amount of error such that for 50% of the time a strategy is used, the overlap between its error-free prediction and the observed choice is at chance level.

Predictions. The specified simulation resembled the design of many experiments that aim to compare two probabilistic models with predictions contingent on the available options and a set of unknown parameters varying between participants. As a feasible way to solve the problem of strategy sprawl and to test and compare toolbox models, the Bayesian technique should recover the data-generating process so that it has a higher posterior probability than the alternative models. For example, if an individual always uses a strategy A_ε, the evidence for that model should be higher than that of a more elaborate toolbox $TB_{A,B}$, even though the latter contains A_ε as a special case. If so, the Bayesian method allows deciding how many strategies to include in a toolbox, thereby preventing strategy sprawl.

Estimation using BUGS. For the model specification at hand, it is not obvious how the marginal likelihood in Equation 1 could be derived analytically. Fortunately, the BF can be estimated using numerical integration techniques such as the Markov chain Monte Carlo method (MCMC; e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996). These methods are readily available in the various software packages, such as WinBUGS or JAGS, that utilize the BUGS (Bayesian analysis using Gibbs sampling) programming language (Lunn, Spiegelhalter, Thomas, & Best, 2009; Lunn, Thomas, Best, & Spiegelhalter, 2000). Using WinBUGS, the BF of a single model A_ε over a toolbox $TB_{A,B}$ was estimated by means of the transdimensional product space method that contains the competing models as well as a binary model indicator (Carlin & Chib, 1995; Han & Carlin, 2001). We drew 100,000 representative samples from the (joint) posterior distributions, split into four independent sampling chains with an initial burn-in of 6,000 steps. Estimation efficiency was facilitated through the use of pseudopriors (Carlin & Chib, 1995; Kruschke, 2011). For all reported parameters, convergence to stationary sample distributions was confirmed. Similar estimation and inspection procedures based on BUGS also apply to the other analyses outlined below.

Graphical model representation. Figure 1 provides a representation of the model comparison procedure in graphical model notation (see Koller, Friedman, Getoor, & Taskar, 2007; Lee & Wagenmakers, 2010; Shiffrin, Lee, Kim, & Wagenmakers, 2008, for details of this notation). In this notation, nodes correspond to variables, edges capture dependencies between variables, and encompassing plates are used to denote independent replications of model structures to indicate repeated choices or multiple decision makers. Observed data are displayed in shaded nodes, unobserved parameters (estimated from data) are displayed in unshaded nodes, continuous variables are indicated by round nodes and discrete variables by square nodes, borders of stochastic variables have a single line, and borders of deterministic variables have two lines.

In Figure 1, the dark square nodes labeled A_i and B_i depict the deterministic predictions of the two probabilistic strategies A_ε and B_ε for each choice i out of N pairwise choices. The round nodes ε_A and ε_{TB} indicate the application errors for A_ε and $TB_{A,B}$, respectively, and the round β node indicates the mixture proportion of probabilistic strategy A_ε over probabilistic strategy B_ε within the toolbox. The model indicator variable z determines whether the probabilistic predictions of A_ε (depicted π_A) or $TB_{A,B}$ (depicted π_{TB}) determine the observed choice data c. For comparing B_ε and $TB_{A,B}$, the model looks similar except that the deterministic predictions of B_ε and A_ε were swapped.

Results. The upper panel of Figure 2 shows the probability of each of the three models under consideration across different combinations of parameter values. The lower panel shows the same data expressed as the BF of A_ε and B_ε over $TB_{A,B}$. The left column in Figure 2 shows that a synthetic participant who made 80 pairwise choices by applying only A_ε with no application error (i.e., $\beta = 1$, $\varepsilon = 0$) yields a BF of 40 in favor of A_ε over $TB_{A,B}$, which translates into a relative probability of $40/(40 + 1) = .98$. Thus, the observed data increased the odds of the simple A_ε model over the more complex toolbox model by a factor of 40, which depicts the upper limit of the evidence that can be obtained in favor of A_ε for the simulated data at hand. In sum, when assuming equal prior probabilities of the two models, the results correctly point to model A_ε as the data-generating model.

For a hypothetical participant who was confronted with the same stimuli set but used A_ε for 64 of the 80 choices and B_ε otherwise (i.e., $\beta = 0.8$, $\varepsilon = 0$), the BF was 19 in favor of $TB_{A,B}$, which translates into a relative probability of .95. Thus, the Bayesian estimation revealed that the participant who used a mix of A_ε and B_ε was better described by the toolbox model than by the single A_ε model even if the majority of the choices were in line with a single strategy. If a participant used A_ε for half of the choices (i.e., $\beta = 0.50$), the BF in favor of the toolbox exceeded 10^6, and if A_ε was used for fewer than 28 choices ($\beta = 0.35$), the BF eventually exceeded 10^6, the limit of our estimation routine. These results indicate that the BF clearly points to the toolbox as soon as some mixing of strategies with no implementation error occurs.

Once an application error was introduced in the choice simulation, the general pattern remained the same, but it became more difficult to recover the data-generating model. For example, as indicated by the middle column of Figure 2, if a participant always selected A_ε to make a choice (i.e., $\beta = 1$) but often made an application error (i.e., $\varepsilon = 0.5$), the BF in favor of A_ε over the toolbox dropped from the initial 40 (for $\varepsilon = 0$) to 4. If A_ε was used for half of the choices (i.e., $\beta = 0.5$, $\varepsilon = 0.5$), the BF in favor of the toolbox dropped to 8. To some extent, this decrease can be compensated for by increasing the number of observations. For example, when doubling the number of simulated choices to 160, the previous parameter combination of $\beta = 0.5$ and $\varepsilon = 0.5$ yield a BF of 84 in favor of the toolbox model. For purely random choices (i.e., $\varepsilon = 1$, left column in Figure 2), the models could no longer be differentiated.

Posterior predictive check: How well did the models describe the data? The BF indicates which of two models provides a better account of the data in relative terms, but it is mute on the absolute quality of model fit. A model with a large BF could still

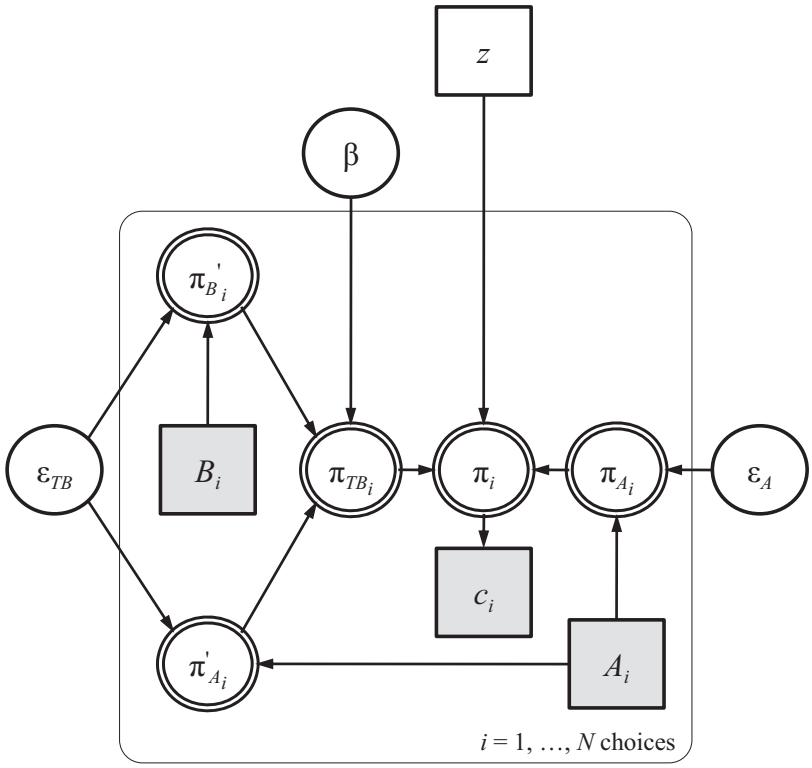

Figure 1. Graphical representation of the model comparison between a toolbox model $TB_{A,B}$ and a single model A_ε. The dark square nodes A_i and B_i depict the deterministic predictions of two strategies A and B for each choice i out of N choices. The εs (round nodes) indicate the application errors, and the round β node indicates the mixture proportion of A_ε over B_ε in the toolbox. The model indicator variable z determines whether the probabilistic prediction of A_ε (depicted π_A) or the probabilistic predictions of $TB_{A,B}$ (depicted π_{TB}) determine the observed choice data c.

provide a terrible fit to the observed data. To test whether the models actually predict the observed data, one can conduct a posterior predictive check by drawing representative samples from the joint posterior distribution of the respective model parameters and using these samples to generate new data (Kruschke, 2012). A count of how often the resulting model predictions match the observed data provides a rough measure of model fit (Gelman, Meng, & Stern, 1996; Meng, 1994). For the simulated data at hand, Figure 3 shows that when choices were simulated as a mix of A_ε and B_ε simulated sets of data, the toolbox yielded more accurate predictions for most of the parameter combinations used when simulating the choice data. The single A_ε and B_ε models were only marginally superior if almost all choices were made in line with either just A_ε or just B_ε. This seems plausible, as the toolbox was the more flexible model. Figure 3 further confirms that the prediction accuracy of all models increased as the application error ε decreased.

Discussion. The results of the model recovery study show that a participant who consistently used a single strategy was clearly better described by this single strategy than a more complex toolbox with two strategies. Once a participant started to systematically use a second strategy for at least some of the choices, the relative probability of a more complex toolbox that entailed both strategies quickly increased. In the context at hand, relatively few choices consistent with the second strategy were sufficient to provide strong support in favor of the more complex toolbox model. In contrast, even when all choices were in line with the single model, the BF did not exceed 40. This asymmetry is plausible as it reflects the fact that the single model can be falsified more easily than the toolbox model that includes the single model as a special case. When the strategies were applied with some degree of error, it became more difficult to identify the data-generating model, but the general pattern remained the same.

Part II: Applying the Toolbox Framework to Empirical Data

The idea of a cognitive toolbox readily applies to various areas of research in psychology and can be tested across seemingly different sets of empirical data. For this test, the same theoretical framework and the same Bayesian techniques outlined above can be successfully applied. In the following, we showcase this in four different research areas: judgment and decision making, children's cognitive development, function learning, and perceptual categorization. In these four domains, we compare toolbox models that differ in size and complexity on the level of a single individual. Finally, we show how the same theoretical and methodological principles can be used to test toolboxes against qualitatively different cognitive models,

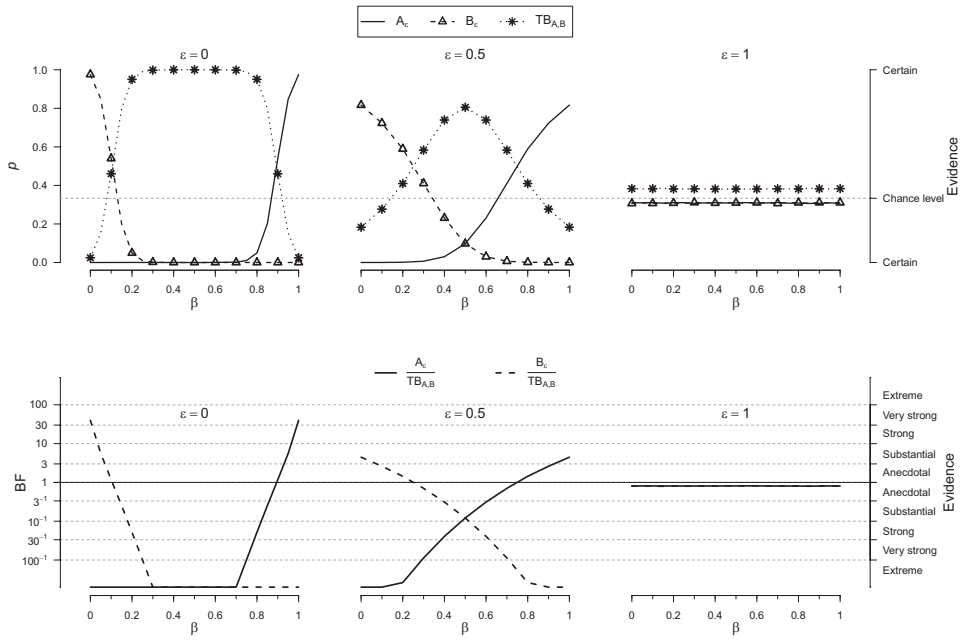

Figure 2. Evidence in favor of the single-strategy model (A_ε or B_ε) over the more complex toolbox model $TB_{A,B}$ depending on the level of noise (ε, across panels) and the proportion of choices according to A_ε relative to B_ε (β, on the *x*-axis). Results are based on independent simulations of 80 pairwise choices each. The upper panels indicate the relative probabilities of all three models for various parameter combinations (i.e., dots within one column add up to 1). The lower panel indicates the same data in terms of the Bayes factor (BF) on a logarithmic scale. Here, values above 1 indicate evidence in favor of the respective single-strategy model, whereas values below 1 indicate evidence in favor of $TB_{A,B}$. BFs are truncated at 500^{-1}.

and we provide a concrete example of how to avoid unwanted strategy sprawl.

Judgment and Decision Making

Directly following up on the first simulation study, we now apply the Bayesian toolbox framework to the domain of judgment and decision making. Here, as described above, researchers have often compared simple noncompensatory decision strategies, such as TTB, to more complex strategies, such as WADD. Moreover, researchers might argue that people could use both strategies depending on the decision situation to make their inferences (cf. Rieskamp & Otto, 2006).

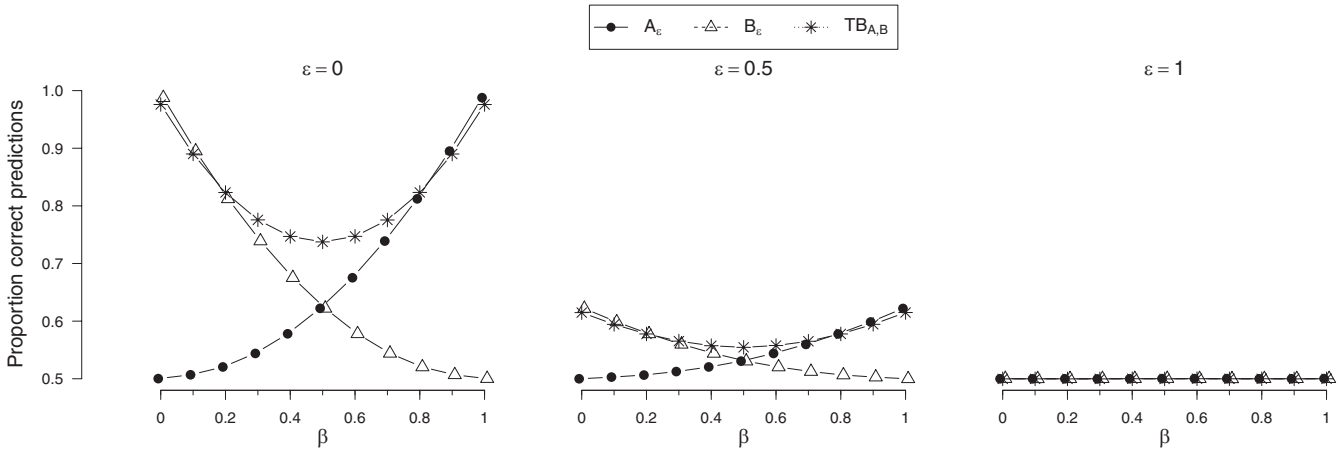

Figure 3. Mean proportion of correct predictions based on posterior estimates of single-strategy models A_ε and B_ε and toolbox model $TB_{A,B}$ for different β and ε parameter values used in the simulation. A mean proportion of .5 indicates chance level; 1 indicates that all pairwise choices are correctly predicted. For most parameter combinations, the toolbox achieves a higher proportion of correct predictions than the respective single models. β = proportion of choices according to A_ε; ε = level of noise.

Data. To explore the use of WADD and TTB, participants in a choice experiment by Rieskamp and Otto (2006) repeatedly chose between options presented as 168 pairs described on six binary (+/−) cues. The cues' importance weights were given, so that the deterministic predictions of TTB and WADD could be determined for each option pair. After each decision, participants received feedback on whether they had successfully chosen the superior option as defined by the experimenters. This feedback differed between two experimental conditions. Participants in the compensatory condition ($N = 20$) received feedback that reinforced WADD, whereas participants in the noncompensatory condition ($N = 20$) received feedback that reinforced TTB.

If participants in the noncompensatory condition learned to use the reinforced strategy exclusively, then a simple TTB_ε model should be better than a toolbox $TB_{TTB,WADD}$ that contains TTB_ε as one of its tools in predicting participants' inferences. Likewise, if the feedback in the compensatory condition led participants to adopt WADD, $WADD_\varepsilon$ should predict their choices better than $TB_{TTB,WADD}$. As an alternative prediction, decision makers may have continued to use $TB_{TTB,WADD}$ and the feedback just influenced the probabilities of selecting TTB_ε and $WADD_\varepsilon$ (i.e., the β parameter) from the toolbox. In contrast to the previous simulation study, the experiment was designed so that the strategies' predictions would not be completely independent and adopting either strategy would result in above-chance performance. These design characteristics make it harder to identify the data-generating strategy.

Method. To test these predictions, we examined the last two blocks in the experiment (48 choices) because analyses by Rieskamp and Otto (2006) indicated that participants' choices did not change much after the first five blocks. We used the same Bayesian estimation techniques as in the previous model recovery simulation to estimate the probabilities of both single-strategy models over the more complex toolbox separately for each participant. In particular, we replaced the single strategies A and B in the simulation with TTB_ε and $WADD_\varepsilon$, as defined by the original researchers, and used similar priors for the respective model parameters as in the simulation: The prior probability of β was set to a uniform distribution ranging from 0 (always choose according to $WADD_\varepsilon$) to 1 (always choose according to TTB_ε); likewise, prior on ε was set to a uniform distribution ranging from 0 (deterministic choice) to 1 (random choice). Following this, the estimation procedure was implemented based on the same BUGS script used in the simulation study.

Results. For a hypothetical participant who always uses the reinforced strategy without an application error, the BF for the single-strategy model over the toolbox would be 25 (i.e., a probability of .96). This depicts the upper limit of the evidence for the experiment at hand.

The results show that in the noncompensatory condition, where participants were reinforced to use TTB, the estimated posterior probability of TTB_ε over $TB_{WADD,TTB}$ ranged from .96 (BF = 25 in favor of TTB_ε) to virtually 0 between individual participants. Figure 4 (upper left) shows that for 16 of the 20 participants, the BF of TTB_ε over $TB_{TTB,WADD}$ was smaller than 1, indicating that for most participants the last 48 choices were better described by $TB_{TTB,WADD}$ than by TTB_ε. Figure 4 (upper right) further shows that the BF of $WADD_\varepsilon$ over $TB_{TTB,WADD}$ was smaller than 1 for 19 of the 20 participants, indicating that $WADD_\varepsilon$ did not provide a good description of participants' choices in the noncompensatory condition. This is plausible, as participants were not incentivized to use WADD in this condition.

In the compensatory condition, the BF of the reinforced $WADD_\varepsilon$ over $TB_{TTB,WADD}$ ranged from 25 in favor of $WADD_\varepsilon$ to 776 in favor of $TB_{TTB,WADD}$ between participants. Figure 4 (lower right) shows that for 14 of the 20 participants, the BF of $WADD_\varepsilon$ over $TB_{TTB,WADD}$ was larger than 1, indicating that the last 48 choices of those individuals were better described by a single $WADD_\varepsilon$ model than by $TB_{TTB,WADD}$. Furthermore, TTB_ε did not provide a good description for any of the 20 participants (see Figure 4, lower left). This suggests that participants had fewer difficulties applying WADD in the compensatory condition than applying TTB in the noncompensatory condition or that they had a preference for using WADD from the beginning, so that less reinforcement was required for using $WADD_\varepsilon$.

Discussion. The results demonstrate the successful application of the Bayesian method to empirical choice data. In particular, assuming equal model probabilities a priori, the method quantifies the probability of the more complex toolbox model over the single strategies for each individual participant. Contrary to Rieskamp and Otto's (2006) findings, these results show that the evidence for a toolbox model differed substantially between the two experimental conditions, with relatively strong support in the noncompensatory condition and relatively weak support in the compensatory condition. This differentiated conclusion relies on quantified posterior model probabilities that Rieskamp and Otto did not derive.

Children's Cognitive Development

The concept of cognitive toolboxes is also common in the developmental literature. Here, children's cognitive advancement is often characterized as an invariant sequence of increasingly complex rules or strategies (Flavell, 1982; Piaget, 1952). Thus, an important research question is which and how many strategies are required to describe children's reasoning within a given domain and how children advance from one developmental stage to the next. The so-called staircase models predict that children transition from one developmental stage to the next in sudden, discrete steps (or stairs), suggesting that at any given point in time, children will exclusively use one single rule. The overlapping waves model, on the other hand, predicts that cognitive development evolves gradually such that the probability of using a more complex rule increases incrementally over developmental time and that transition periods are characterized by a mix of two consecutive rules (e.g., Jansen & van der Maas, 2002; Siegler, 1994). From the perspective of a cognitive toolbox, the overlapping waves model predicts that children at times may have two alternative tools at their disposal to solve a given task, whereas the staircase model predicts that children will use different rules depending on their age but that any given child will use only one particular tool.

Children's cognitive development has been extensively studied with the well-known balance-scale task, which requires predicting the movement of a balance scale depending on the number of weights and their distance from the fulcrum on each of the two arms. Siegler (1976) initially proposed that children may use at least four different rules (labeled Rules 1–4 by the original author) that are characterized by an increasing integration of the weight

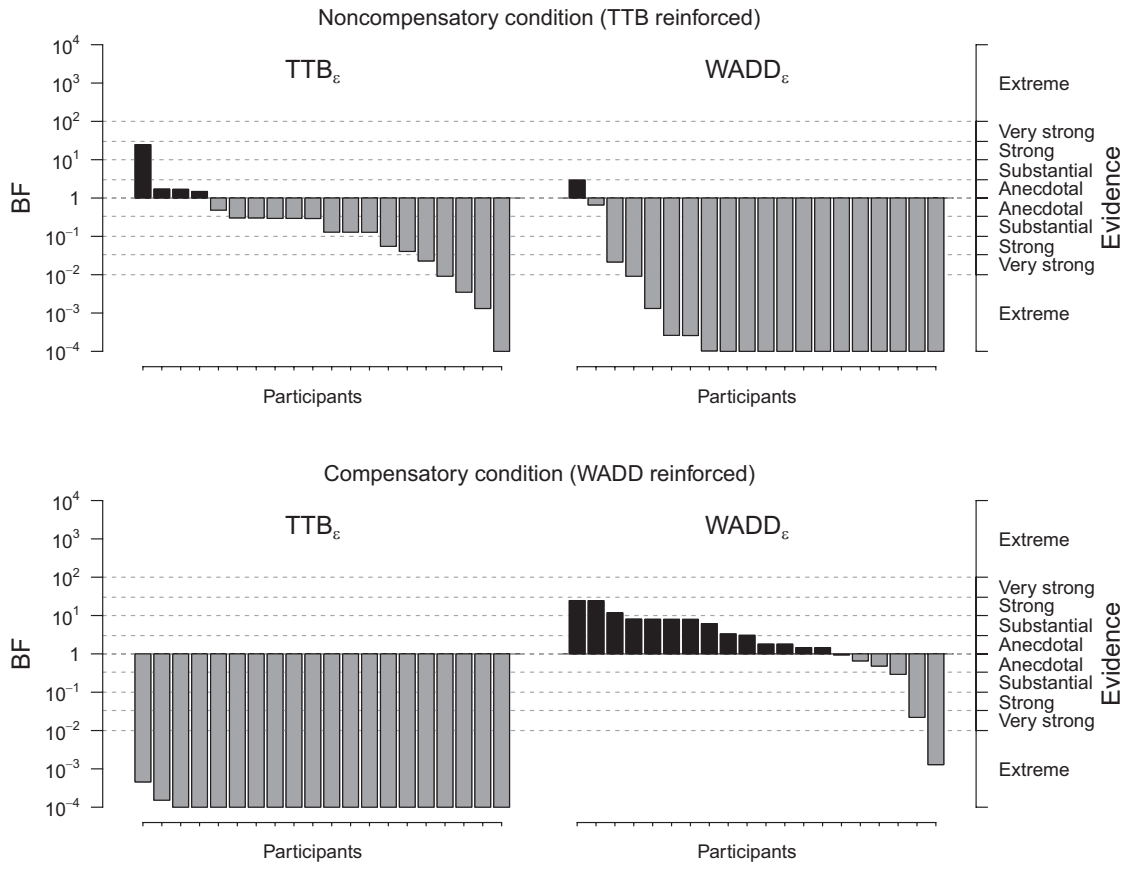

Figure 4. Bayes factor (BF) of the single strategies take the best (TTB_ε, left) and weighted additive ($WADD_\varepsilon$, right) over the toolbox strategy ($TB_{TTB,WADD}$). Bars represent estimates for each participant in the experiment by Rieskamp and Otto (2006), ordered by BF (logarithmic scale). In the noncompensatory condition (upper panel), TTB was reinforced; in the compensatory condition (lower panel), WADD was reinforced. Positive values (black bars) indicate stronger evidence in favor of the respective single model (TTB_ε or $WADD_\varepsilon$), whereas negative values (gray bars) indicate stronger evidence for $TB_{TTB,WADD}$. BFs are truncated at 10^{-4}.

and distance dimension.[2] For example, Rule 1, the simple weight rule, considers only the weights and disregards distance. Rule 2 takes distance into account only if the weights are equal. Rule 3 takes weights and distance into account but without knowing how to combine them, so it often guesses. Rule 4, the multiplication rule, derives the normative solution by multiplying the number of weights by their distances. The rules that particular children adhere to are commonly identified by counting how often their responses match the rules' predictions (Siegler, 1976).

Data. In one experiment, Jansen and van der Maas (2002) asked 805 children and adolescents, ages 5 to 19 years, to solve 25 different balance-scale items. The items varied in complexity such that the simplest ones could be solved by (almost) any rule, whereas other items could only be solved by more advanced rules requiring the integration of weight and distance. To test if children's strategies were better described by a single rule or by a toolbox consisting of a mix of two consecutive rules, we applied a similar Bayesian toolbox model as before. In particular, in a first step, we implemented each of the four models (i.e., rules) proposed by Siegler (1976) in BUGS based on their respective predictions for each of the items. Next, we employed a total of three different toolbox models consisting of two consecutive rules, respectively (i.e., Rules 1 and 2, Rules 2 and 3, and Rules 3 and 4). Thus, for each child, there were a total of seven candidate models—the four single rules and the three toolboxes. The staircase theory predicts that children's answers at any given age will usually be best described by one of the four single rules, whereas the wave model predicts that some children will be better described by a toolbox, that is, a mix of two consecutive strategies.

Method. To estimate which of the candidate models best described the individual answer patterns, we repeatedly conducted pairwise model comparisons using a similar implementation to that used before.[3] Each pairwise model comparison yielded a BF. A total of six pairwise comparisons were conducted in a tournament-

[2] For simplicity, we only consider these four rules; additional rules have been proposed in the literature and could also be tested through our approach (e.g., Jansen & van der Maas, 1997).

[3] When implementing the models in BUGS, the main difference from the previous cases that had to be accounted for was that some of Siegler's (1976) original rules occasionally predict guessing and not deterministic choices.

like fashion (i.e., Rule 1 against a toolbox $TB_{1,2}$, $TB_{1,2}$ against Rule 2, etc.). From these pairwise comparisons, the posterior probability of each single model relative to the set of all seven candidate models was obtained using simple algebra.

Results. Figure 5 shows that the proportion of children using a simple rule (e.g., Rule 1) declines with age, that the medium-complex Rule 3 peaks around the age of 12 years, and that Rule 4 is increasingly used by children above the age of 13 years.[4] These results are in line with the basic notion that cognitive development can be described as the progression through a sequence of increasingly complex rules.

The results further indicate that many children were best described by a toolbox that combined two consecutive rules. In particular, many teenagers seemed to use a toolbox consisting of Rules 3 and 4, and about 10% of the children ages 5 to 13 years were best described by a toolbox that combines Rules 1 and 2. This suggests that the cognitive development of many children in the balance-scale task, in particular, teenagers, seemed to progress in waves rather than in discrete steps. Note that even for the latest developmental stage (i.e., the oldest age group), the toolbox with Rules 3 and 4 was still prevalent, suggesting that many people continue to use the simpler Rule 3 even at an older age. Owing to the cross-sectional nature of the data, it is difficult to tell if the wave model applies to all children or whether some children switch to more complex rules in discrete steps.

Discussion. These results extend the findings reported by the original authors in important ways. Jansen and van der Maas (2002) analyzed the data with a latent class analysis (LCA). LCA is related to factor analysis as it detects response patterns in the data and the exact rules do not have to be known beforehand (van der Maas, Quinlan, & Jansen, 2007). LCA assumes that each individual belongs to one single latent class or rule. Therefore, to test for possible switches between rules, as predicted by the overlapping waves model, van der Maas et al. (2007) divided their data into blocks and then tested for consistency.

In line with the staircase model, the van der Maas LCA analysis indicated a considerable degree of consistency in rule use across the different blocks of the experiment. However, van der Maas et al. (2007) also observed some inconsistencies between the blocks, in line with the waves model. Our Bayesian analysis allows a more detailed articulation of these results by showing exactly how many individuals were better described by a single rule (i.e., staircase model) as compared to a toolbox model (i.e., waves model). Van der Maas et al. further hypothesized that Rules 1 and 2 will usually not overlap in development. Our Bayesian analysis suggests that these two rules may overlap but only for very few participants.

It should be noted that the inclusion of additional or alternative strategies may change the estimated probabilities. For example, Jansen and van der Maas (2002) tested two additional strategies. Also, Siegler's Rule 3 predicts random guessing for 15 of the 25 items at hand, which makes it difficult to distinguish it from erroneous implementations of the other rules or toolboxes. The main objective of the present analysis was to provide a proof of concept that outlines how the cognitive toolboxes and the adjunctive Bayesian model comparison techniques can be fruitfully applied to test theories in the developmental literature.

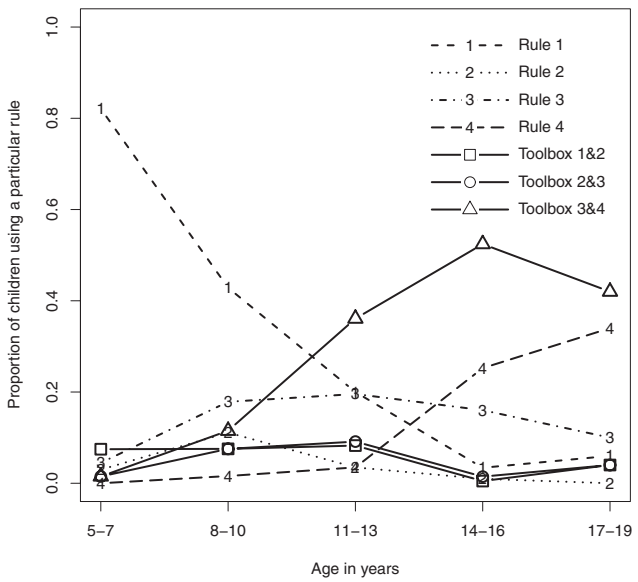

Figure 5. Proportion of children within each age group who use a particular strategy in the balance-scale task by Jansen and van der Maas (2002). Solid lines indicate the use of a toolbox (i.e., the combination of two consecutive rules). Proportions were calculated by assigning the rule with the highest posterior probability to each individual child. Proportions within each age group add up to one.

Function Learning

Function-learning research examines how people learn the functional relationships between two or more variables that vary on a continuous scale (e.g., DeLosh, Busemeyer, & McDaniel, 1997). Recent theories on function learning resemble the idea of a toolbox by suggesting that people's knowledge of functional relationships can be context specific and that several tools, referred to as rules, strategies, or experts, often coexist in parallel and are selected depending on the context. One theory that promotes such a function-learning toolbox is the population of linear experts (POLE) theory (Kalish, Lewandowsky, & Kruschke, 2004). According to POLE, individuals possess a repertoire of (simple) candidate response functions, and one response function is applied on any given trial. The framework of knowledge partitioning (Lewandowsky & Kirsner, 2000) predicts that people rely on context cues to selectively learn and apply functional relationships. Likewise, von Helversen and Rieskamp (2008) argued that depending on the functional relationship between the cues and a criterion, different judgment processes take place. In contrast to these toolbox models, other theories assume that knowledge is not domain specific; the same process occurs independent of the context or the structure of the task (e.g., Busemeyer, Byun, Delosh, & McDaniel, 1997; Kelley & Busemeyer, 2008; Speekenbrink & Shanks, 2010). Because theories on function learning apply to continuous judgments, this example requires that we extend the proposed Bayesian toolbox framework beyond the discrete choice situations analyzed so far.

[4] For these illustrative purposes, here, the rule with the highest posterior probability was assigned to each individual child. Averaging across the relative probabilities of each rule yields similar results.

Data. To test for knowledge partitioning in function learning, Lewandowsky et al. (2002) conducted a series of experiments in which participants learned to predict the value of one continuous variable (i.e., the speed of wildfire spread) from another (i.e., wind speed) in 180 trials with feedback. The underlying link function was U-shaped such that fire speed was high for very low and very high wind speed and was low for medium wind speed.

In the systematic context condition of the experiment, 90 training stimuli from the descending (i.e., left) part of the function were primarily presented in one specific color context, and another 90 training stimuli from the ascending part of the function were primarily presented in another color context (for methodological details, see Experiment 2 of Lewandowsky et al., 2002). Thus, the participants in this condition could partition their knowledge and use a toolbox consisting of two rules to predict fire spread contingent on the color context: a descending function for items in one context and an ascending function for items in the other context. Alternatively, participants could also apply a single yet complicated (i.e., integrated) rule by learning the underlying U-shaped function independent of the color context.

In a second, random-context condition, stimuli on both sides of the U shape occurred in both colors. Thus, participants in this condition could not partition their knowledge contingent on the color context but rather had to revert to the full U-shaped function to solve the task. As a measure of knowledge partitioning, all participants eventually responded to the same transfer set of 74 new item pairs that were presented twice, once in each context. We used this transfer set to test if participants applied the single rule or if they partitioned their knowledge and applied a toolbox.

Method. To test if people applied a context-dependent toolbox or a single all-purpose tool, we implemented a Bayesian model comparison. Despite the many theoretical differences between function learning and the previous examples from the developmental and decision-making literature, similar Bayesian model comparison techniques apply because the theories share a common toolbox assumption.

There are, however, some differences with respect to the details of the respective tools and how they are selected. In the previous examples, the probability of selecting one tool over the other was a free parameter that was estimated from the data, and the magnitude of this parameter was of potential interest in itself. In Lewandowsky et al.'s (2002) function-learning experiment, this parameter was replaced by an explicit theory of how the tools are selected. Accordingly, it was assumed that the color context determined which rule was selected. Therefore, the toolbox consisting of the two context-specific tools was implemented as two separate linear functions that linked wind speed (w) to fire spread (y) for all stimuli i. The two functions were indexed by an indicator variable c that represented the context (i.e., the color) of the stimuli:

$$\gamma_i = \gamma_{co} + \gamma_{c1} \times \mathrm{w_i}, \tag{5}$$

so that a total of four γ parameters (two for each single tool) were estimated from the data.

As the alternative to this toolbox model, we defined the single function-learning tool as a quadratic function, irrespective of the color context:

$$\gamma_i = \gamma_0 + \gamma_1 \times w_i + \gamma_2 \times w_i^2. \tag{6}$$

Here, three γ coefficients had to be estimated from the data.

The models were implemented in BUGS using the so-called Zellner's g prior (Ntzoufras, 2009; Zellner, 1986). The g prior assigns the γ parameters a multivariate normal distribution with a prior variance that is $1/n$ times as important as the variance of the maximum likelihood estimate for γ. This prior is said to contain as much information as a single observation, and it is a popular default choice in Bayesian model selection for linear regression (see also Wetzels, Grasman, & Wagenmakers, in press). Wind speed and fire spread were normalized before entering the model. Just as in the previous case, a binary model indicator variable was established in BUGS to estimate the probability of one model over the other on the individual level.

Results. Figure 6 shows that in the systematic context condition that allows toolbox use, the data from eight of 24 participants clearly provides evidence for a knowledge-partitioning toolbox over a single all-purpose quadratic model. In contrast, the quadratic model provided a better description for all but one participant in the random context condition.[5] Because the experiment was set up such that the test items clearly differentiated between the two strategies, for most participants, the evidence was clear and decisive, as indicated by the large BFs.

Discussion. The results show how the cognitive toolbox concept readily applies to the function-learning research. Some participants in the systematic context condition used a toolbox (i.e., they partitioned their knowledge), whereas participants in the random control condition were best described by a single judgment strategy. These findings are in line with the conclusions of Lewandowsky et al. (2002). Our Bayesian approach was able to uncover and quantify important interindividual differences: Not all participants in the systematic context condition partitioned their knowledge—a substantial number of participants also applied an all-purpose quadratic rule and apparently ignored the context. This heterogeneity seems plausible in light of the original hypothesis that both strategies work well in the systematic context condition. Note that in the present study, strategy selection was determined by the context. Despite this difference from the previous examples where selection was a free parameter, the Bayesian approach could nevertheless be applied in the same vein.

Categorization

The idea of a cognitive toolbox is also prevalent in the categorization literature. Little and Lewandowsky (2009), for instance, showed empirically that people use different categorization rules depending on the context. Ashby and Maddox (2005) concluded in their review that people commonly use multiple and qualitatively different rules or strategies to categorize objects. The idea that categorization responses may rest upon qualitatively different tools or mixtures of experts can also be found in recent categorization models such as RULEX (Nosofsky, Palmeri, & McKinely, 1994), PRAS (Vandierendonck, 1995), and ATRIUM (Erickson &

[5] The evidence in favor of knowledge partitioning in the systematic context condition becomes even stronger if the toolbox model is implemented as a mix of two quadratic rather than two linear functions. See online supplemental material for the results of this implementation.

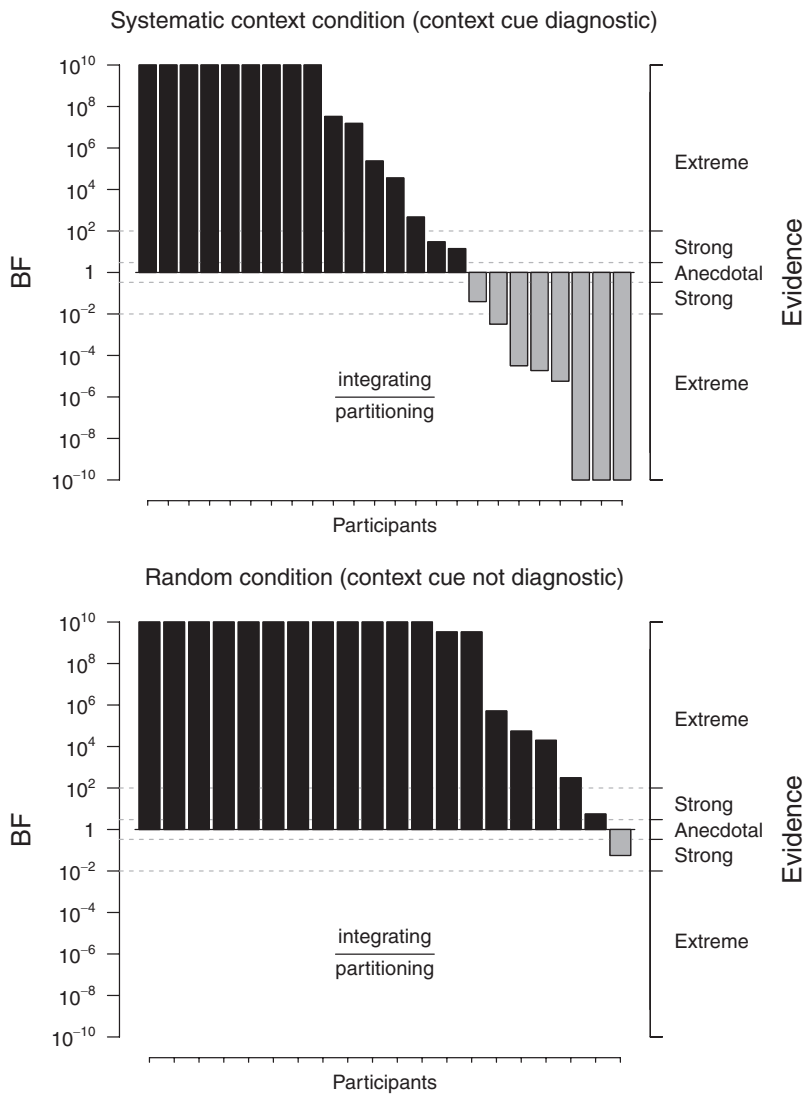

Figure 6. Reanalysis of Lewandowsky, Kalish, and Ngang's (2002) Experiment 2 data. Bars represent Bayes factors (BFs, logarithmic scale) of the integrating all-purpose quadratic rule over knowledge partitioning for each individual participant in the systematic context condition (upper panel) and the random context condition (lower panel). Positive values (black bars) indicate stronger evidence for the all-purpose quadratic rule. Negative values (gray bars) indicate stronger evidence for knowledge partitioning. BFs are truncated at 10^{10} and 10^{-10}.

Kruschke, 1998).[6] All these models assume different cognitive processes underlying categorization. These recent models differ, however, from many early theories on categorization that assume a single, all-purpose categorization process independent of context (Ashby & Maddox, 2005). In sum, the categorization literature is another domain in which a rigorous test of the cognitive toolbox idea is crucial for theoretical advancement.

Data. To assess the extent to which categorization can be described by a single all-purpose tool versus a mixture of different context-specific tools, Yang and Lewandowsky (2004, Experiment 1) had participants learn to classify geometric objects into one of two categories. The objects were described along two continuous dimensions x and y that represented the position of a vertical segment and the height of a rectangle, respectively. Category membership depended on a specific parallel boundary rule that described a complex interaction between the x and y dimensions. Each object was further presented in one of two colors, the so-called context cue that is similar to the experimental manipulation of Lewandowsky et al. (2002), discussed in the previous example.

The categorization study featured two experimental conditions that varied between participants: In the systematic context condition, color in itself did not predict category membership, but

[6] In contrast to a pure toolbox model, ATRIUM assumes that the categorization responses produced by the different modules are subsequently weighted and integrated into a final response.

same-color objects could be classified by a somewhat simple rule that consisted of a linear combination of the *x* and *y* dimensions. As this rule was different for each color, participants in this condition could either learn the complex all-purpose parallel boundary rule or they could learn to use a toolbox consisting of two rules that each worked well in the respective color contexts. In the random context condition, color was randomly assigned to the object. As color carried no systematic information in this condition, here participants could only apply the all-purpose complex parallel boundary rule. To test which classification rule was applied, a similar procedure was used as in the function-learning experiment: All participants saw the same set of 40 novel test objects once within each color context, so that the parallel boundary rule and the toolbox made different predictions.

Method. To test if participants' classification strategies could be described based on a cognitive toolbox consisting of two different, context-dependent rules, we compared it to the alternative parallel boundary model.[7] As this comparison closely resembles the previous toolbox examples, similar Bayesian techniques as before were applicable after some initial data recoding. We estimated the posterior probability of the single parallel boundary model over the toolbox individually for each of the 48 participants (24 in each condition).

Results. As shown in Figure 7, in the systematic context condition, about half the participants (10 of 24) were better described by a toolbox, indicating knowledge partitioning. In contrast, in the random context condition, all but two participants were best described by the parallel boundary rule. As in the previous case of function learning, for most individuals in the experiment, the obtained evidence for either model was decisive.

Discussion. The example above demonstrates how the theoretical concept of cognitive toolboxes can be fruitfully applied to models of category learning and that the Bayesian techniques for testing toolboxes also inform theories in this research area. Our analysis confirms the results reported by Yang and Lewandowsky (2004) that participants used qualitatively different strategies depending on the experimental condition and that most participants in the systematic context condition nevertheless used an all-purpose classification strategy instead of a toolbox. Furthermore, the results provide BF estimates of the parallel boundary model over the toolbox model, that have an intuitive and transparent interpretation and that go beyond the explorative *k*-means cluster analysis in the original study that left some participants unclassified.

Adding Tools to the Toolbox—The Case of Strategy Sprawl

So far, the toolboxes we tested consisted of only two strategies. Could a toolbox with more tools describe the data in the examples above even better, or should we avoid adding tools and stick to simple toolboxes? Perhaps some children in the experiment by Jansen and van der Maas (2002) actually used a mix of three or even more rules to solve the balance-scale task. Maybe it is advisable to add a third tool to the decision-making toolbox in the experiment by Rieskamp and Otto (2006). Certainly, adding yet another tool would increase the flexibility of any toolbox. It is less clear, however, if this increase in complexity actually yields additional insights into the underlying cognitive processes or if it would foster strategy sprawl.

Data. To test if adding another tool to the toolbox is justified, we returned to the data of Rieskamp and Otto (2006) and extended the toolbox with a tallying strategy (TALLY_ε; Dawes, 1979). TALLY predicts that pairs of options are compared along their cues and that the option that is superior on the majority of cues is chosen. In essence, TALLY is a rather simple compensatory strategy that has been successfully applied to predict people's inferences (Bröder, 2000; Mata et al., 2011). Adding TALLY_ε provides an interesting case as the strategy was not reinforced by Rieskamp and Otto. Thus, even though the extended toolbox $\text{TB}_{\text{TTB,WADD,TALLY}}$ is essentially more flexible in fitting the observed choice data, this increase might not translate into stronger evidence in this case.

Method and results. To test the toolbox extended with TALLY_ε, we applied the same Bayesian method as before to compare the toolbox consisting of three strategies to a toolbox with only two strategies. As the original experiment was not designed to test for TALLY_ε, the strategy sometimes does not discriminate between the option pairs, and its predictions often overlap with those of one of the other two strategies. As this effectively decreases the number of discriminating data points, smaller (i.e., less decisive) BFs are expected compared to some of the previous examples. For example, always using TALLY_ε with no error yields a BF of 15 (compensatory condition) and 8 (noncompensatory condition) in favor of the larger toolbox that includes TALLY_ε. In comparison, always applying WADD_ε with no error yields a BF of 12 and 13, respectively, for the smaller toolbox.

Figure 8 shows that within these limits, most participants in Rieskamp and Otto's (2006) experiment were better described by the smaller toolbox, in particular so in the compensatory experimental condition. Adding TALLY_ε as an additional tool apparently was mostly not justified in the light of the observed data.

Discussion. The results showed how a toolbox with only two strategies outperformed a toolbox with more strategies, thus providing an example of how the Bayesian method tackles strategy sprawl and restricts the number of strategies within a toolbox. Although a larger toolbox can fit more diverse data, in the present case, the increased model complexity was not offset sufficiently by the increased goodness of fit. Rather, a simpler, more parsimonious model was best in predicting behavior.

Testing a Toolbox Against Alternative Models

The previous section showed how toolboxes of different sizes can be compared against single cognitive strategies. The same Bayesian approach allows us to test a toolbox against alternative cognitive models that are conceptually different and that are not nested within the toolbox.

As an example for such a comparison, we return to the literature on decision making. Here, exemplar models have been suggested as an alternative to rule-based models. Past research found that exemplar models are able to predict people's cognitive processes in various domains, including memory (Hintzman, 1988), automatization (Logan, 1988), likelihood judgments (Dougherty, Gettys,

[7] See Appendix A for the implementation of the categorization model at hand.

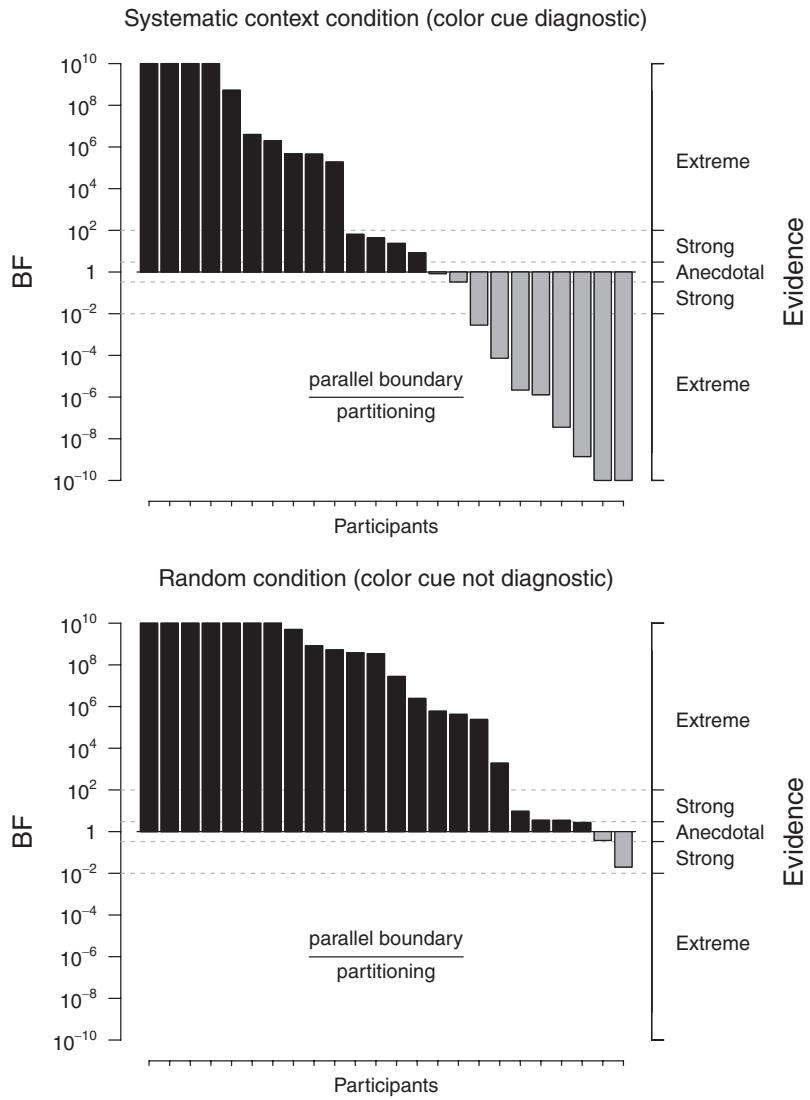

Figure 7. Reanalysis of the data from Yang and Lewandowsky's (2004) Experiment 1. Bars represent Bayes factors (BFs, logarithmic scale) of the parallel boundary rule over knowledge partitioning (i.e., toolbox use) for each individual participant in the systematic context condition (upper panel) and in the random condition (bottom panel). Positive values (black bars) indicate stronger evidence for the parallel boundary rule; negative values (gray bars) indicate stronger evidence for knowledge partitioning. BFs are truncated at 10^{10} and 10^{-10}.

& Ogden, 1999), and attention (Logan, 2002). Juslin and Persson (2002) also proposed exemplar-based models for describing judgment processes (see also Juslin, Jones, Olsson, & Winman, 2003; Juslin, Olsson, & Olsson, 2003). Unlike rule-based strategies such as TTB or WADD, exemplar models rely on a similarity-based judgment process. Exemplar models assume that judgments are made by comparing the present situation (probe) with similar situations (exemplars) stored in memory. Exemplar models require memory and retrieval processes to perform the task. Thus, they are usually considered conceptually different from rule-based models that rely on abstract knowledge (Juslin, Jones, et al., 2003; Nosofsky & Johansen, 2000).

Data and method. We compared an exemplar model of choice against a toolbox based on the empirical choice data of Rieskamp and Otto (2006) already analyzed above. The exemplar model was adapted from Juslin, Jones, et al. (2003) with one free parameter to model the attention weight s that represents the subjective importance of each cue when determining the similarity to previously encountered options.[8] The attention weight varied on a scale from 1 (*not important*) to 0 (*very important*). For simplicity and consistency with past research, we assumed identical attention weights for all cues (e.g., Persson & Rieskamp, 2009; von Helversen & Rieskamp, 2008). We set the prior on s to be uniformly

[8] See Appendix B for the conceptualization of the exemplar model on hand.

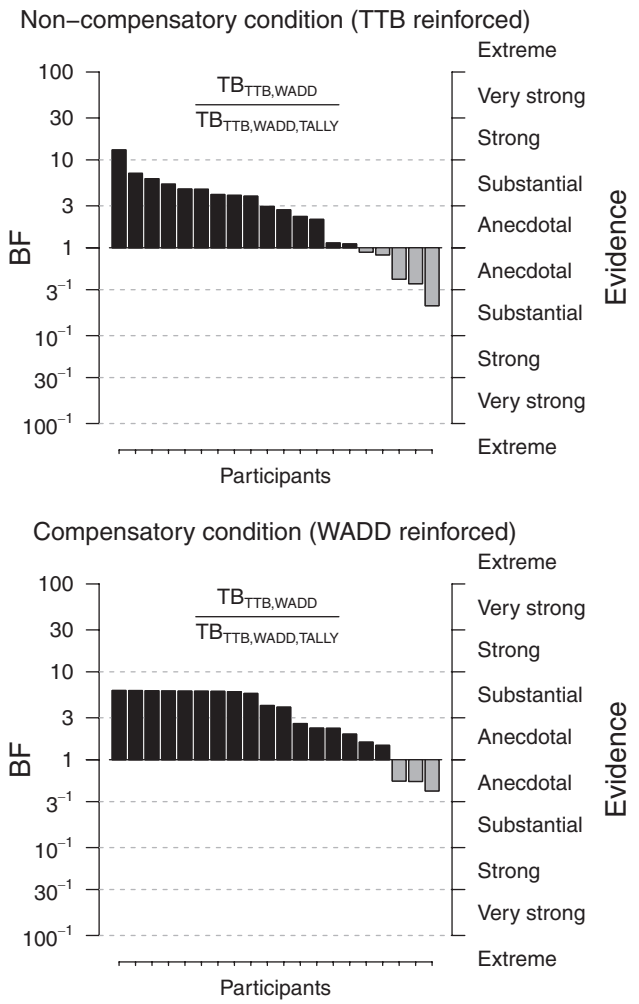

Figure 8. Bayes factor (BF, logarithmic scale) of the two-strategy toolbox $TB_{TTB,WADD}$ over the three-strategy toolbox $TB_{TTB,WADD,TALLY}$. Bars represent estimates for each participant in the experiment by Rieskamp and Otto (2006), ordered by BF. In the noncompensatory condition (upper panel), take the best (TTB) was reinforced; in the compensatory condition (lower panel), the weighted-additive (WADD) strategy was reinforced. Positive values (black bars) indicate stronger evidence in favor of the small toolbox, whereas negative values (gray bars) indicate stronger evidence for the larger toolbox.

distributed in the range of 0 to 1; the $TB_{TTB,WADD}$ model consisting of $WADD_\varepsilon$ and TTB_ε was implemented as before.

Results and discussion. Only four out of the 20 participants who were reinforced to use TTB in the noncompensatory condition were better described by an exemplar model. As shown in Figure 9 (upper panel), the evidence was substantial for only two participants. In contrast to this, the evidence in favor of $TB_{TTB,WADD}$ was much stronger for the 16 remaining participants.

In the compensatory condition, where participants were reinforced to use WADD, the evidence against the exemplar model was even stronger. As shown in Figure 9 (lower panel), for 19 out of 20 participants, the BF strongly favored the toolbox model.

The toolbox's superiority over the exemplar model is also evident when comparing the choices predicted by both models to the choices made by the participants. Figure 10 shows that when the model parameters were sampled from the estimated joint posterior distribution, the predictions of $TB_{TTB,WADD}$ provide a better match to the observed choices than the exemplar model. This posterior predictive evaluation further reveals that the exemplar model is still a reasonable model, as, for most participants, its predictive accuracy is clearly above chance level. In sum, for the experimental data of Rieskamp and Otto (2006), the toolbox model described the choice data better than the exemplar model. This result seems plausible as the participants were encouraged to use either TTB_ε or $WADD_\varepsilon$ and the toolbox model contains both strategies.

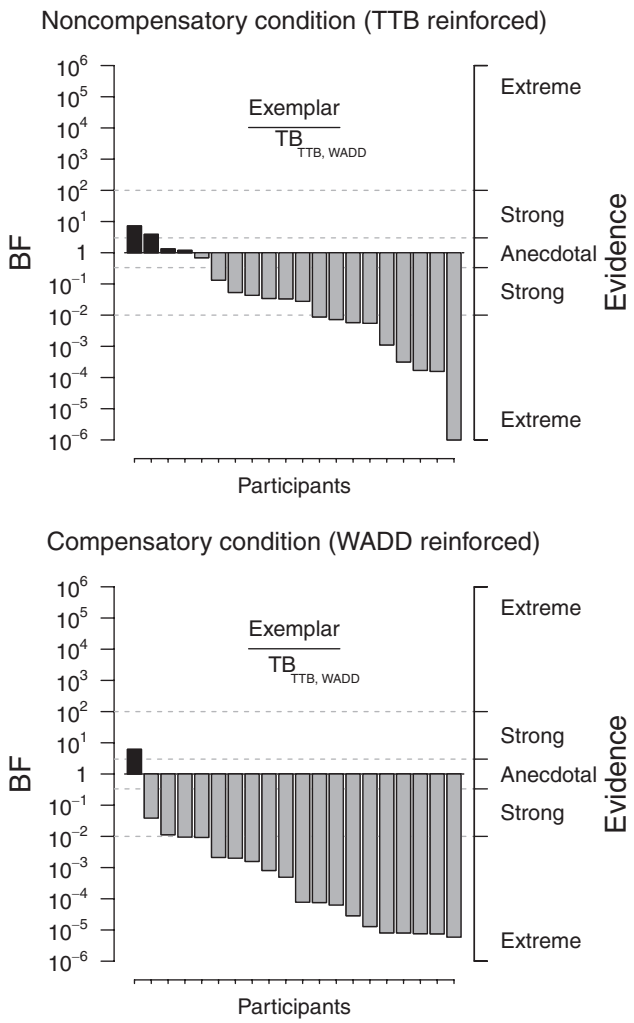

Figure 9. Bayes factor (BF, logarithmic scale) of the exemplar model over the toolbox $TB_{TTB,WADD}$. Columns represent estimates for each participant in the experiment by Rieskamp and Otto (2006), ordered by BF. In the noncompensatory condition (upper panel), take the best (TTB) was reinforced; in the compensatory condition (lower panel), the weighted-additive (WADD) strategy was reinforced. Positive values (black bars) indicate stronger evidence in favor of the exemplar model, whereas negative values (gray bars) indicate stronger evidence for $TB_{TTB,WADD}$. BFs are truncated at 10^{-6}.

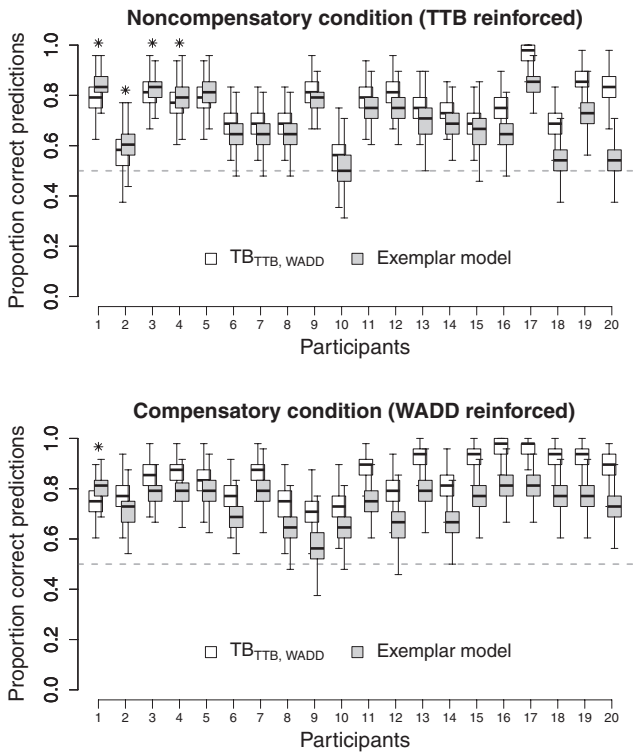

Figure 10. Posterior predictive evaluation of the toolbox $TB_{TTB,WADD}$ and the exemplar model for each participant in the experiment by Rieskamp and Otto (2006), ordered by Bayes factor (asterisks indicate Bayes factor > 1). In the noncompensatory condition (upper panel), take the best (TTB) was reinforced; in the compensatory condition (lower panel), the weighted-additive (WADD) strategy was reinforced. A value of .5 indicates chance level, whereas a value of 1 would show that all choices were correctly predicted. The figure shows that for most participants, $TB_{TTB,WADD}$ achieved a higher predictive accuracy than the exemplar model. Error bars extend to 1.5 times the interquartile range.

Part III: Testing Toolbox Models on the Group Level

When testing a cognitive theory, the ultimate goal is often to draw conclusions about a specific group of people or even the general population. One way to achieve this would be to pool the data across all individuals and estimate the models based on the aggregated data set as if it originated from a single person. Although feasible, this complete pooling approach ignores possible variations and differences among individual decision makers (Gelman & Hill, 2007). Furthermore, the averaged data might not be representative of any of the single individuals who produced the data (Estes, 1956; Heathcote, Brown, & Mewhort, 2000).

One way to address this problem would be to test theories only on the individual level, similar to the approach we used before. However, assuming that the choices of one person are completely independent from those of another person makes it difficult to generalize the results and to test and compare models on the group level. This complete independence assumption also neglects commonalities between individuals; it is plausible that parameter estimates of one individual may also inform the parameter estimates of another, presumably similar person. If people are analyzed independently, this information is lost. Neglecting possible similarities between individual estimates also entails the risk of extreme estimates for individual people that may be unlikely given the distribution of the model estimates on the group level (Gelman & Hill, 2007). These have led to increased interest in hierarchical or partial-pooling approaches, as they are able to describe both the commonalities and the differences between individuals (Cohen, Sanborn, & Shiffrin, 2008; Gelman & Hill, 2007; Nilsson, Rieskamp, & Wagenmakers, 2011; Rouder & Lu, 2005). In the hierarchical approach, a balance between complete pooling and complete independence is achieved by assuming that the individual parameter estimates for each individual stem from higher level group distributions. As these group-level distributions are estimated themselves, it is not necessary to determine the actual degree of pooling beforehand. Rather, the similarity between group members and the degree to which the individual estimates are mutually informative follow from the observed data and the structure of the hierarchical model.

As we outline in several concrete steps below, this principle is very useful as a more elaborate way of testing and comparing cognitive toolboxes on the group level. We first conducted a model recovery study on to the group level and then applied the method to empirical data.

Simulated Choice Data on the Group Level

For the hierarchical group-level approach, we extended our initial simulation by repeatedly simulating 20 synthetic participants who each made 80 pairwise choices. Within each group of 20 individuals, we set the proportion of participants who applied Strategy A_ε (denoted φ_A) to either 1 (all group members applied A_ε) or 0.9. Here, $\varphi_A = 0.9$ indicated that 18 of 20 participants chose according to A_ε (i.e., $\beta = 1$) and the remaining two chose according to Strategy B_ε (i.e., $\beta = 0$). As a second factor, we varied the mean application error E between groups from 0.025 to 0.975 in steps of 0.05. Within each group, the individual ε values assigned to each individual slightly varied around E (0.025 was added to E for half of the individuals and subtracted for the other half) to allow for some variance between individuals. The combination of different parameter values resulted in a total of 40 independent groups. For each of these synthetic groups, we estimated the posterior probability of a simple A_ε model over a more complex toolbox $TB_{A,B}$ on the group level.

Hierarchical extension of the Bayesian method. To develop the hierarchical extension of the Bayesian method previously applied on the individual level, we defined a normally distributed group-level distribution for each parameter of the choice strategies. Prior distributions were assigned to the respective means μ and variances σ of these group-level distributions. As the possible parameter values for the application error ε and the mixture proportion parameter β on the individual level only ranged from 0 to 1, they had to be rescaled into normally distributed values through a probit transformation to allow for proper aggregation on the group level (Rouder & Lu, 2005). For easier interpretation, the obtained group-level parameters were later retransformed to the original rate scale of 0 to 1.

The prior distributions for μ and σ were set such that the resulting values would be uninformative on the original rate scale. Specifically, for the prior on μ, we used a normal distribution with

mean 0 and variance 1. The prior on σ was uniform with a minimum of 0.001 and a maximum of 4 on the probit scale to prohibit extreme distributions on the original rate scale. In this way, group-level distributions were assigned to all model parameters, including the individual application error of the single strategy, ε_A; the application error for the toolbox strategies, ε_{TB}; and the mixture parameter β in the toolbox. In a final step, we implemented a transdimensional model indicator on the group level in BUGS, indicating the overall probability of the models across all participants.

Results of the group-level model comparison. Figure 11 shows that when all 20 decision makers used A_ε ($\varphi_A = 1$) with only a small mean application error ($E = 0.025$), the estimated BF of A_ε over $TB_{A,B}$ on the group level indicated extreme evidence ($BF > 1{,}000$) in favor of A_ε. Figure 11 also shows that the evidence for the data-generating Strategy A_ε remains very strong even for relatively high application errors in the simulation. However, if only two individuals within a group of 20 ($\varphi_A = 0.9$) used B_ε and not A_ε, the BF clearly favored the toolbox model. For small application errors, the BF in favor of the toolbox was extreme (i.e., $BF > 10^6$), and even for relatively high application errors up to about 60%, the evidence in favor of the toolbox was still decisive.

Figure 11 further shows that if the application errors become very high (i.e., between about 0.7 and 0.9), the toolbox becomes slightly less probable as compared to A_ε, even though 2 of 20 participants in the simulation used B_ε (i.e., $\varphi_A = 0.9$.) This is probably because the number of genuine B_ε choices in these simulated data became very small relative to A_ε, so that eventually the more parsimonious single A_ε model was preferred. With an application error near 1, indicating random choice, the models can no longer be differentiated.

Discussion. The hierarchical Bayesian approach showed strong evidence in favor of a simple model if all individuals within a group applied the simple model. Once even a minority used a different strategy, the Bayesian analysis indicated strong evidence in favor of a more complex model incorporating the heterogeneity in strategy use within the group. These results are plausible as the single A_ε strategy could not account for the choices of those group members who used Strategy B_ε consistently. As a consequence, the Bayesian analysis correctly inferred that it was very unlikely that the single model generated all of the observed data.

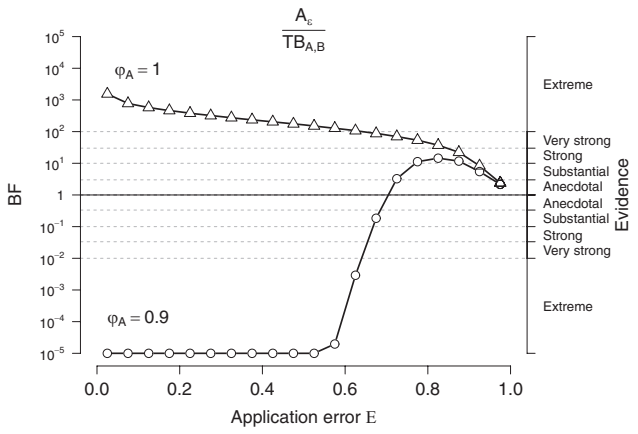

Figure 11. Evidence of the single strategy A_ε over the toolbox $TB_{A,B}$ for a group of 20 simulated decision makers making 80 choices each depending on the mean application error E and the proportion φ_A of A_ε users over B_ε users set in the simulation. The left *y*-axis indicates the advantage of A_ε over $TB_{A,B}$, expressed as the Bayes factor (BF, logarithmic scale). The right *y*-axis provides an approximate interpretation of evidence strength. Here, higher values indicate evidence in favor of A_ε, whereas negative values indicate evidence for $TB_{A,B}$. Values are truncated at 10^{-5}.

Comparing Toolbox Models on the Group Level Based on Empirical Data

The hierarchical Bayesian analysis readily applies to empirical data across a wide range of research areas that rely on the notion of cognitive toolboxes. We illustrate how the Bayesian approach can enhance data analysis and theory building on the group level with the decision-making experiment of Rieskamp and Otto (2006). The goal was to decide if the group of all 20 decision makers within each experimental condition was better described by a toolbox $TB_{TTB,WADD}$ or just a single-decision strategy ($WADD_\varepsilon$ or TTB_ε). It is unclear how the parameter estimates obtained from a no-pooling approach on the individual level should be aggregated on the group level. In contrast, the proposed hierarchical, partial-pooling approach provides a principled way of drawing conclusions on the group level by taking all individual data into account simultaneously.

The above group-level simulation already indicated that a single strategy is only preferred on the group level if individual decision makers were very homogeneous, that is, if almost all used a single strategy. Furthermore, the above individual analysis of the Rieskamp and Otto (2006) data also showed that not all participants used the reinforced single-decision strategy. Rather, a few individuals were clearly identified as toolbox users on the individual level. Thus, we can predict that the more flexible $TB_{TTB,WADD}$ model should be more probable for the Rieskamp and Otto data on the group level.

Method and results. The model comparison was implemented based on routines similar to those used for estimating the simulated choice data. In line with the hypothesized results, the group-level BF for $TB_{TTB,WADD}$ over the reinforced single-strategy model was in excess of 10^6 (i.e., the numerical limit of our estimation routine) for both experimental conditions. This clearly indicates that on the group level, the toolbox model described the data better than any of the single strategies.

Discussion. Although at first glance the estimated BFs in favor of the toolbox on the group level may appear rather high, the results are in line with the findings of the previous group-simulation study: Unlike the flexible toolbox model, a single strategy cannot explain the heterogeneity between group members. Thus, although most individual decision makers in the compensatory condition (14 of 20) were best described by the single $WADD_\varepsilon$ model, the same cannot be said of the remaining minority of participants. Because the $TB_{TTB,WADD}$ model can account for both cases, the Bayesian approach correctly indicates that it is the more probable model on the group level. The same reasoning also holds for the noncompensatory experimental condition, where even more participants were classified as toolbox users on the individual level. The hierarchical group-level extension of the proposed Bayesian method illustrates the strength of the toolbox approach in general: A toolbox model is able to describe the

substantial heterogeneity across and within individuals of how they solve the problems they face.

Posterior group-level distributions. In addition to the model comparison, the hierarchical Bayes model also allowed us to draw detailed conclusions about all model parameters on the group level. Here, the group-level distribution of the mixture proportion of $WADD_\varepsilon$ over TTB_ε was of particular interest as it allowed us to quantify the prevalence of using one strategy over the other across all individuals (see Dennis, Lee, & Kinnell, 2008, for a similar procedure). As shown in Figure 12, in the noncompensatory condition, a mean mixture proportion of .71 in favor of TTB_ε was found on the group level. The highest density interval of this posterior distribution (HPD_{95}) ranged from .60 to .81. This clearly indicates that decisions in this group were based more on TTB_ε than on $WADD_\varepsilon$. In the compensatory condition, this group-level mixture proportion was estimated at .10 (HPD_{95} of .04–.16), indicating a clear preference for using $WADD_\varepsilon$ over TTB_ε in this condition. As the highest posterior density intervals of the two estimates were far apart, the probability of using TTB_ε over $WADD_\varepsilon$ clearly differed between the two experimental conditions, illustrating the effect of the experimental manipulation.

The obtained hierarchical group-level estimates are different from estimates obtained by merely averaging across all β estimates in the individual analysis. For example, in the compensatory condition, such a simple average yields a mean β of 0.15 with a 95% confidence interval from 0.098 to 0.197 across the 20 participants. This underestimates the prevalence of using $WADD_\varepsilon$ on the group level; the Bayesian group-level estimate does not even fall within the confidence interval of the simple mean. For data that contain outliers or if the number of observations differs between individuals, the bias induced by simply averaging across individual data may become even larger, further underscoring the advantages of the hierarchical Bayesian approach (see also Gelman & Hill, 2007).

Group-Level Comparison Against Alternative Models

Finally, we extended the group-level approach to examine if the decisions of a group of people as a whole were better described by a toolbox model or by an alternative, nonnested model. Comparing qualitatively different models on the group level is an important challenge that is significant for theory building in many areas in psychology.

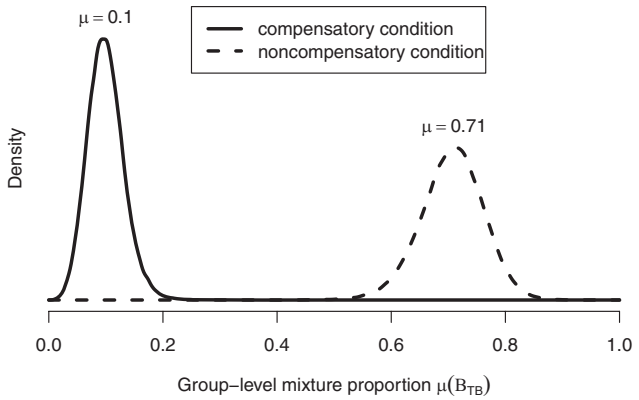

Figure 12. Posterior distribution of the mean group-level mixture proportion $\mu(B_{TB})$ of take the best (TTB_ε) over the weighted-additive ($WADD_\varepsilon$) strategy for both experimental conditions in the study by Rieskamp and Otto (2006). Higher mixture proportions indicate a higher prevalence of TTB_ε over $WADD_\varepsilon$ on the group level.

Data. Similar to the nonnested model comparison on the individual level, we again used the empirical choice data of Rieskamp and Otto (2006). The results on the individual level suggest that the toolbox $TB_{TTB,WADD}$ will be more probable on the group level than the exemplar model. However, putting this prediction to the test requires a hierarchical Bayesian approach that allows us to quantify the evidence in favor of one model over the other in a principled way.

Method and results. The implementation of the hierarchical $TB_{TTB,WADD}$ model was similar to the one outlined in the group analysis above. For the exemplar model, we extended the model implementation previously used to estimate individual participants with a group-level normal distribution to partially pool the individual attention parameters s_k. Priors were assigned to the mean and variance of this group-level distribution. The basis of the model comparison was again a transdimensional model indicator implemented in BUGS.

As expected, the group-level BF for $TB_{TTB,WADD}$ over the exemplar model was in excess of 10^6 for both the compensatory and the noncompensatory conditions. This indicates decisive evidence in favor of $TB_{TTB,WADD}$ on the group level.

Discussion. The results show how toolbox models can be rigorously compared to alternative models of cognition. The same principles can also be applied to toolboxes containing other tools or to alternative cognitive models across various fields of research in psychology as long as these models can be thoroughly specified. The hierarchical approach provides a principled method for partially pooling the data from each single group member such that meaningful conclusions can be drawn on the aggregate level. At the same time, the individual estimates are mutually informed by the estimates of other group members.

This group-level extension readily applies to other research areas in psychology. For example, in Jansen and van der Maas's (2002) balance-scale experiment, most children ages 5–7 years were best described by Siegler's (1976) Rule 1, and hardly any child used Rule 4. Thus, it would be interesting to test if Rule 4 should even be included within the set of possible models for children in this age group. Likewise, even though many children above the age of 10 years seem to use a toolbox consisting of Rules 3 and 4, it would be interesting to test if the probability of applying Rule 4 over Rule 3 within the toolbox increases across the different adolescents' age groups, an analysis that would provide important insights on how children's and adolescents' cognitive strategy use develops over time. Although both questions can be readily answered by means of the outlined hierarchical analysis, for brevity, we do not report the results of such an analysis here. Likewise for brevity, we do not present a hierarchical analysis on the group level for the function-learning and categorization research described above, as our main goal was to outline and illustrate the theoretical and methodological advantages of using a (Bayesian) toolbox approach.

General Discussion

Researchers in psychology have often built their theories on the idea of a strategy repertoire or toolbox that contains a set of cognitive strategies. This approach provides a fruitful framework for modeling cognitive processes, as it accounts for the variability between individuals and it can explain why the same person often behaves differently in nearly identical situations. Although the framework has been applied across several areas in psychology with considerable success, it comes with a number of methodological and theoretical challenges that make it difficult to rigorously test and compare toolbox theories as a whole. To overcome these challenges, we outlined a unifying Bayesian approach that we tested based on model recovery studies and empirical data from various research fields including judgment and decision making, cognitive development, function learning, and perceptual categorization. The flexibility of the approach was further illustrated through comparisons of toolboxes of different size and through comparisons with qualitatively different models. In all these cases, the approach provided novel insights and yielded precise, consistent, and readily interpretable results that rely on a common comparison metric that incorporates both model complexity and goodness of fit.

Trade-Off Between Goodness of Fit and Parsimony

A larger cognitive toolbox may provide a better fit to the observed data because of its greater flexibility, but it may not necessarily provide the best explanation of the underlying cognitive processes (Myung, 2000). Our results indicate that the Bayesian method is well suited for trading off complexity against goodness of fit and hence preventing strategy sprawl: A person who consistently uses one strategy is best described by a simple model that resembles this particular strategy as compared to a more complex toolbox that includes the single model. However, once a person starts using a mix of strategies, the Bayesian approach can account for this situation by endorsing a toolbox.

The same principle holds for analyses on the group level: If most individuals within a given group consistently use the same strategy, a simple model that matches this strategy is preferred over a toolbox that allows for individual differences in strategy use between group members. However, once some group members also use other strategies, a group-level toolbox that allows for such heterogeneity is preferred. In the latter case, Bayesian techniques further provide the posterior probability of selecting each of the tools within the toolbox on the group level.

Bayesian Model Selection

In the Bayesian framework, model complexity can be thought of as the number of different predictions the model can make. As mentioned earlier, a complex model can generate many different predictions, so that the prior probability of each of these predictions is relatively small (Jefferys & Berger, 1992). In contrast, the predictions of a simple model are relatively restricted, so that the prior probabilities are relatively high. This effect carries over to the posterior probability of the model because the likelihood of the data given a specific combination of parameters is weighted by their prior probability. Thus, even though greater flexibility may help to increase the likelihood of the observed data for a specific subset of the parameter space, this increase is counteracted by a lower prior probability of the observed data in the remainder of the parameter space.

This mechanism can be illustrated for the decision-making area with a simple review of the comparison between TTB_{ε} and $TB_{TTB,WADD}$. TTB_{ε} has one free parameter (the application error ε_{TTB}), whereas $TB_{TTB,WADD}$ has two parameters (the application error ε_{TB} and β, the probability of using TTB_{ε} over $WADD_{\varepsilon}$). Because of its higher flexibility, $TB_{TTB,WADD}$ can account for different kinds of observed choice data, but consequently, the model does not make firm predictions before the data come in (cf. Vanpaemel, 2010). In the Bayesian framework, the higher flexibility of $TB_{TTB,WADD}$ is discounted because the prior probabilities must be defined for each possible combination of parameter values. As the joint prior distribution across the parameter space must integrate to one, the prior probability of each prediction decreases with the total number of possible predictions. It is due to this principle that more flexible models are implicitly penalized. Nevertheless, the data may sometimes warrant additional flexibility: If the likelihood for the more flexible toolbox is much higher than the likelihood for the simple model, the penalty for additional flexibility will be offset.

Numerical Estimation Techniques

The suggested Bayesian approach to testing toolbox models requires one to estimate the BF and the posterior model probabilities. For elaborate cognitive models such as toolboxes, these quantities usually cannot be obtained by means of closed-form mathematical solutions. Instead, they must be approximated through numerical techniques such as MCMC methods. These methods can be implemented relatively easily using one of several software packages; here, we used the BUGS language implemented in the software package WinBUGS that runs on any conventional desktop computer and provides a long series of samples from the desired posterior distributions within a reasonable amount of time, even for relatively complex hierarchical models.

Alternative Methods for Comparing Toolbox Models

Apart from the Bayesian techniques outlined here, many alternative methods exist to compare and select between competing cognitive models. All these methods have to negotiate the compromise between goodness of fit and parsimony, but they achieve this compromise in different ways. Here, we discuss the most prominent methods, with particular focus on how they are similar to and different from the Bayesian technique discussed above.

Bayesian information criterion. One popular method is to quantify and correct for model complexity solely through the number of free parameters. This procedure provides the basis for statistical indices such as the Akaike information criterion (AIC; Akaike, 1973; Burnham & Anderson, 2002) and the Bayesian information criterion (BIC; Masson, 2011; Myung, 2000; Raftery, 1995; Schwarz, 1978; Wagenmakers, 2007; for a discussion on the differences between the AIC and BIC, see, e.g., Karabatsos, 2006; Vrieze, 2012).

As the name suggests, the BIC was derived as an approximation of the BF. This approximation is particularly good if the models

under comparison are *nested,* such that one is a simplified version of the other (Kass & Wasserman, 1995). Another advantage of the BIC is that it is relatively straightforward to compute; the BIC is given by BIC $= -2 \times \ln L + k \times \ln n$, where $\ln L$ is the log maximum likelihood, k is the number of free parameters, and n is the sample size. This simplicity, however, comes at a cost: The BIC ignores interactions between parameters and is blind to differences in the parameters' functional form (Karabatsos, 2006; Myung & Pitt, 1997, 2009; Pitt, Myung, & Zhang, 2002).

To illustrate that the BIC may sometimes provide a poor approximation of the BF, consider again the previous simulation study that compared a toolbox $TB_{A,B}$ consisting of two strategies A_ε and B_ε against a single strategy A_ε on the individual level. Here, using the same priors as before, synthetic data consisting of 60 pairwise choices—80% in line with Strategy A_ε and the remaining 20% in line with Strategy B_ε with no application error (i.e., $\varepsilon_A = \varepsilon_B = 0$)—yield a BF of 5.7 in favor of the toolbox $TB_{A,B}$. In contrast, the BIC, estimated by means of maximum-likelihood techniques, approximates a BF of 11.5, thus overestimating the evidence in favor of the toolbox by a factor of two.[9]

The limitations of the BIC in this context become even more apparent when altering the model's functional form. Consider, for instance, the study of Rieskamp and Otto (2006). Here, people using a $WADD_\varepsilon$ strategy will most likely have a higher application error than people using the simple TTB_ε strategy because the former requires, on average, a larger number of processing steps. One way to model this is by setting $\varepsilon_{WADD} > \varepsilon_{TTB}$ within the toolbox—a constraint that limits the possible parameter range without affecting the *number of parameters*. Estimating this modified toolbox based on the same simulation data as before ($N = 60$ choices, 80% in line with TTB, and no implementation error) yields a BF of 5 in favor of the constrained toolbox. In comparison, a toolbox with two independent (i.e., unconstrained) error terms yields a BF of only 2.2. This decrease seems justified because the simulated choice data fell within the range of the more constrained toolbox. In contrast, the BIC approximation yields a BF of 1.5 in both cases, irrespective of the change in functional form.

In other contexts, different changes in functional form may be justified for theoretical reasons or based on prior knowledge about the task. For example, if participants in a given experiment were encouraged to use a strategy B instead of A, the prior probability of B_ε will be higher than that of A_ε. To reflect this, the previous uniform prior distribution on the β parameter that describes this probability within the toolbox may be replaced, for instance, with a beta(1,4) prior distribution that increases the prior probability to Strategy B_ε. With respect to the same simulation data of 60 choices described above, this change should decrease the posterior probability of the toolbox because most observed choices were actually in line with A. In line with this intuition, the BF now indicates the opposite from before, namely, a preference against the toolbox, indicating that now the data are 3 times more probable under A_ε than under the toolbox. The BIC estimate, on the other hand, remains unchanged and still suggests a BF of 1.5 in favor of the toolbox. More extreme changes in functional form (e.g., stronger priors or range restrictions for some of the parameters) may lead to even bigger deviations between the Bayesian approach and the BIC approximation.

Nevertheless, in many situations, the BIC may provide a good approximation of the BF. For example, in the previous analyses of the data on function learning (Lewandowsky et al., 2002) and categorization (Yang & Lewandowsky, 2004), the BIC yields conclusions that were very close to those reached by the exact Bayesian methods that required much more effort to implement. In many model comparison situations, it might be difficult, though, to decide in advance whether the BIC approximation will lead to accurate conclusions.

In sum, parameter-counting model selection methods such as the BIC have the advantage of simplicity and may work well in many situations. Unfortunately, though, these methods are insensitive to functional form complexity, and—as illustrated above—this means that in certain situations, they may fail completely (e.g., as for order-restricted inference; Hoijtink, 2001).

Cross-validation. One model comparison technique that implicitly takes complexity into account beyond the mere number of parameters is cross-validation (Stone, 1974). In cross-validation, the observed data are split into two (or more) subsamples. One calibration sample is used to estimate the models' parameters, and in a second step, the predictive accuracy of the fitted model is tested using the second crucial validation sample (Browne, 2000). Although cross-validation provides a rough method for trading off model complexity against fit, the best way to split data into different samples is unclear, and crucially, different splitting methods can lead to different results and conclusions (see Shao, 1993, 1997). Alternative cross-validation approaches, such as leave-one-out, have been shown to overfit the data (i.e., to select models that are overly complex, just as the AIC does; see Stone, 1977). Finally, in contrast to Bayesian model comparison, cross-validation does not quantify model preference in terms of probability, making it difficult to calibrate and interpret the results.

Accumulative one-step-ahead prediction error. As with cross-validation, the idea behind one-step-ahead accumulative prediction error (APE; Dawid, 1991; Luan, Schooler, & Gigerenzer, 2011; Rissanen, 1986b; Wagenmakers, Grünwald, & Steyvers, 2006) is to assess a model's worth by the adequacy of its predictions. Assume that data points come in one by one and that every time, the model's goal is to make the best prediction for the very next data point. In contrast to leave-one-out cross-validation, the calibration set grows as the data accumulate. Thus, one-step-ahead model predictions are relatively poor at first, but they improve as the calibration set grows in size. We then sum all the prediction errors and prefer the model that has the smallest summed one-step-ahead prediction error.

When prediction error is measured by logarithmic loss, APE is identical to BF model selection. When prediction error is measured with quadratic loss, APE only approximates BF model selection though. In addition, APE implements the principle of *predictive minimum description length* (Rissanen, 1986a). Hence, APE is an attractive method for model selection, with firm theoretical underpinnings. One drawback of APE is that with quadratic loss, the end results are not available as probabilities or odds, and this hinders the interpretation of the results. In addition, with quadratic loss, the results depend on the order in which the data arrive—a generally undesirable property (Wagenmakers et al., 2006). Although there is often a natural ordering to experimental data (e.g., Participant 1

[9] The online supplemental material provides a more systematic outline of this comparison.

is tested prior to Participant 2, and Trial 1 is completed before Trial 2), this ordering is often considered inconsequential, as the data for different participants or trials are assumed to be interchangeable. To obtain an order-independent answer, one can calculate the final APE as an average of APEs for many random orderings of the same data set (Kontkanen, Myllymaki, & Tirri, 2001; Rissanen, 1986a). The drawback of this procedure is that it greatly increases the computational burden. Finally, a thorough comparison of APE and the BF has yet to be performed.

Minimum description length. Another approach is to penalize model complexity using the principle of minimum description length (MDL; e.g., Grünwald, 2000, 2007; Grünwald, Myung, & Pitt, 2005; Rissanen, 1987, 2007). The principle can be instantiated in different ways, but in general, MDL methods seek to quantify the extent to which a model can be used to extract regularities from the data in order to minimize the amount of information required for the data's description. The model that achieves the greatest compression of the data is preferred. BFs and MDLs are derived from very different theoretical frameworks, but they nevertheless tend to yield the same conclusion. Consider, for instance, the differential geometry version of the MDL (e.g., Pitt et al., 2002), which is given by

$$\text{MDL} = -\ln L + \frac{k}{2} \times \ln\left(\frac{n}{2\pi}\right) + V.$$

This equation is effectively the same as the BIC, except for the additional term V. This term quantifies model complexity in terms of the number of different predictions that the model can make. The term V is defined as $= \ln \int d\theta \sqrt{\det[I(\theta)]}$, where θ indicates the parameter space and $I(\theta)$ is the Fisher information matrix (Rissanen, 1996). Inclusion of V allows the MDL to account for functional form complexity, something that the BIC cannot. The BF approach used in this article can account for functional form complexity, just as the MDL does. Furthermore, Balasubramanian (1997) showed that the differential geometry version of the MDL can be recast as a finite-series approximation of the BF model selection using Jeffreys's (1961) prior.

In sum, MDL and BF model selections are closely related in terms of their inferences. We prefer BF model selection over MDL mainly for two reasons. First, in practice, it is much more difficult to apply the MDL than it is to apply BF model selection, a difference that is partly due to the availability of MCMC and partly due to the BF model selection having always been the more popular method of the two. Second, the Bayesian framework deals with uncertainty in parameters (for fixed data), whereas some versions of the MDL deal primarily with uncertainty across the sample space, considering data that may have been obtained but were not (Wallace & Dowe, 1999). More generally, the Bayesian inference machine provides a coherent account of model selection.

Rigorously Testing the Toolbox Approach

Although the Bayesian approach yields coherent and intuitive results, its successful implementation comes with a series of requirements. One is that each toolbox must be precisely defined in mathematical terms. This includes a specification of how strategies within the toolbox are selected. When the selection process is not defined in detail, a pragmatic solution is to estimate the selection probability from the observed data. We applied this generic way of specifying the selection process in our simulation study and outlined how it can be fruitfully used to analyze various empirical data. However, researchers sometimes have specific theories on how strategies are selected (Gigerenzer & Brighton, 2009; Lovett, 1988; Marewski & Schooler, 2011; Siegler & Shipley, 1995). For example, in the above cases of function learning and categorization, a clear-cut theory existed on which specific tool was selected depending on the context. Toward the goal of advancing toolbox theories in psychology, further theory building on how the strategy selection problem can be solved is an important area for future research (Marewski & Schooler, 2011; Rieskamp & Otto, 2006). A more detailed understanding of the conditions that influence which strategy will be selected could provide the basis for more rigorous theoretical tests.

In other cases, models may have additional free parameters that must be estimated from data, such as the weights or the rank order of cues in judgments and decision making (Bergert & Nosofsky, 2007; Lee & Newell, 2011). However, regardless of the detailed model implementation and selection processes, the same Bayesian model comparison techniques suggested here could be applied without loss of generality.

Once the available tools and the selection processes are properly specified, the outlined Bayesian approach can also be used to test toolbox models against qualitatively different psychological theories. In areas where toolbox models are proposed, there are often alternative theories with qualitatively different theoretical assumptions. In these cases, an important empirical question is which model can best explain the underlying cognitive processes. The Bayesian approach outlined here allows researchers to rigorously address this question.

Additional Model Selection Criteria

Testing different theories against each other by means of quantitative (Bayesian) techniques does not replace essential qualitative tests of theories, though. This includes an evaluation of the plausibility of the theories' underlying theoretical assumptions, the overall exploratory power and descriptive adequacy, the theoretical justification of and consistency with previous knowledge of the underlying mental processes, and also the interpretability and usefulness of the model predictions (Myung & Pitt, 1997; Pitt et al., 2002). These qualitative model comparison criteria are important for avoiding the selection of theoretically implausible models just because they predict data better than even more miserable models (Kass & Raftery, 1995). Here, posterior predictive checks such as those outlined above become important as they allow one to estimate the absolute goodness of fit. Likewise, different models often lead to qualitatively different predictions that should be tested.

Conclusion

In summary, the above examples show how the theoretical framework of cognitive toolboxes can be traced across various, seemingly unrelated lines of research in psychology. Although the framework has many virtues, it requires persuasive model comparison techniques that allow rigorous testing of the toolbox approach. The outlined Bayesian techniques provide a coherent and

widely applicable method for analyzing this class of models. Once applied, the obtained results can be easily communicated and interpreted, and they provide multivariate posterior distributions of the estimated model parameters (and not just point estimates). The approach further provides the opportunity to make prior assumptions explicit and to integrate prior knowledge. Thus, we can draw fine-grained conclusions that can enhance the development of toolbox models across many fields in psychology. As there are many ways to Rome, it is important to understand which roads are traveled more and which ones less.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. doi:10.1146/annurev.psych.56.091103.070217

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation, 9,* 349–368. doi:10.1162/neco.1997.9.2.349

Ball, C. T., Langholtz, H. J., Auble, J., & Sopchak, B. (1998). Resource-allocation strategies: A verbal protocol analysis. *Organizational Behavior and Human Decision Processes, 76,* 70–88. doi:10.1006/obhd.1998.2798

Bergert, F. B., & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 107–129. doi:10.1037/0278-7393.33.1.107

Bröder, A. (2000). Assessing the empirical validity of the "take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1332–1346. doi:10.1037/0278-7393.26.5.1332

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1539–1553. doi:10.1037/0278-7393.21.6.1539

Brown, N. R., Cui, X. J., & Gordon, R. D. (2002). Estimating national populations: Cross-cultural differences and availability effects. *Applied Cognitive Psychology, 16,* 811–827. doi:10.1002/acp.830

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44,* 108–132. doi:10.1006/jmps.1999.1279

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: University of Chicago Press.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Busemeyer, J. R., Byun, E., Delosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. R. Shanks (Eds.), *Knowledge, concepts and categories: Studies in cognition* (pp. 408–437). Cambridge, MA: MIT Press.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121,* 177–194. doi:10.1037/0096-3445.121.2.177

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459. doi:10.1037/0033-295X.100.3.432

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review, 100,* 204–232. doi:10.1037/0033-295X.100.2.204

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 57,* 473–484. doi:10.2307/2346151

Cohen, A. L., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review, 15,* 692–712. doi:10.3758/PBR.15.4.692

Costa-Gomes, M. A., & Crawford, V. P. (2006). Cognition and behavior in two-person guessing games: An experimental study. *American Economic Review, 96,* 1737–1768. doi:10.1257/aer.96.5.1737

Coyle, T. R., Read, L. E., Gaultney, J. F., & Bjorklund, D. F. (1998). Giftedness and variability in strategic processing on a multitrial memory task: Evidence for stability in gifted cognition. *Learning and Individual Differences, 10,* 273–290. doi:10.1016/S1041-6080(99)80123-X

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

Dawes, R. M. (1979). Robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582. doi:10.1037/0003-066X.34.7.571

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 53,* 79–109.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 968–986. doi:10.1037/0278-7393.23.4.968

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59,* 361–376. doi:10.1016/j.jml.2008.06.007

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery, 3,* 409–425. doi:10.1023/A:1009868929893

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review, 115,* 199–211. doi:10.1037/0033-295X.115.1.199

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106,* 180–209. doi:10.1037/0033-295X.106.1.180

Einhorn, H. J. (1970). Use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin, 73,* 221–230. doi:10.1037/h0028695

Eisenberg, P., & Becker, C. A. (1982). Semantic context effects in visual word recognition, sentence processing, and reading: Evidence for semantic strategies. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 739–756. doi:10.1037/0096-1523.8.5.739

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112,* 912–931. doi:10.1037/0033-295X.112.4.912

Erev, I., & Roth, A. E. (2001). On simple reinforcement learning models and reciprocation in the prisoner dilemma game. In G. Gigerenzer & R. Selten (Eds.), *The adaptive toolbox* (pp. 215–232). Cambridge, MA: MIT Press.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127,* 107–140. doi:10.1037/0096-3445.127.2.107

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140. doi:10.1037/h0045156

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review, 99,* 689–723. doi:10.1037/0033-295X.99.4.689

Flavell, J. H. (1982). On cognitive development. *Child Development, 53,* 1–10. doi:10.2307/1129634

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733–760.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62,* 451–482. doi:10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103,* 650–669. doi:10.1037/0033-295X.103.4.650

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall.

Glöckner, A., Betsch, T., & Schindler, N. (2010). Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making, 23,* 439–462. doi:10.1002/bdm.668

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247. doi:10.1037/0096-3445.117.3.227

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44,* 133–152. doi:10.1006/jmps.1999.1280

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

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Han, C., & Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association, 96,* 1122–1132. doi:10.1198/016214501753208780

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review, 7,* 185–207. doi:10.3758/BF03212979

Hilbig, B. E. (2010). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review, 17,* 923–930. doi:10.3758/PBR.17.6.923

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95,* 528–551. doi:10.1037/0033-295X.95.4.528

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research, 36,* 563–588. doi:10.1207/S15327906MBR3604_04

Jansen, B. R. J., & van der Maas, H. L. J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review, 17,* 321–357. doi:10.1006/drev.1997.0437

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81,* 383–416. doi:10.1006/jecp.2002.2664

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80,* 64–72.

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

Juslin, P., Jones, S., Olsson, H., & Winman, A. (2003). Cue abstraction and exemplar memory in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 924–941. doi:10.1037/0278-7393.29.5.924

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General, 132,* 133–156. doi:10.1037/0096-3445.132.1.133

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26,* 563–607. doi:10.1207/s15516709cog2605_2

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review, 111,* 1072–1099. doi:10.1037/0033-295X.111.4.1072

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology, 50,* 123–148. doi:10.1016/j.jmp.2005.07.003

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. doi:10.1080/01621459.1995.10476572

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90,* 928–934. doi:10.1080/01621459.1995.10476592

Kelley, H., & Busemeyer, J. (2008). A comparison of models for learning how to dynamically integrate multiple cues in order to forecast continuous criteria. *Journal of Mathematical Psychology, 52,* 218–240. doi:10.1016/j.jmp.2008.01.009

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning* (pp. 13–55). Cambridge, MA: MIT Press.

Kontkanen, P., Myllymaki, P., & Tirri, H. (2001). Comparing prequential model selection criteria in supervised learning of mixture models. In T. Jaakkola & T. Richardson (Eds.), *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics* (pp. 233–238). Los Altos, CA: Morgan Kaufmann.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118,* 97–109. doi:10.1037/a0020762

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Oxford, England: Elsevier Academic Press.

Kruschke, J. K. (2012). Posterior predictive checks can and should be Bayesian: Comment on Gelman and Shalizi, "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology*. Advance online publication. doi:10.1111/j.2044-8317.2012.02063.x

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55,* 1–7. doi:10.1016/j.jmp.2010.08.013

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review, 11,* 343–352. doi:10.3758/BF03196581

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. (2010). *A course in Bayesian graphical modeling for cognitive science*. Unpublished manuscript. Retrieved from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Exper-*

imental Psychology: General, 124, 83–97. doi:10.1037/0096-3445.124.1.83

Lewandowsky, S., Kalish, M., & Ngang, S. K. (2002). Simplified learning in complex situations: Knowledge partitioning in function learning. *Journal of Experimental Psychology: General, 131,* 163–193. doi:10.1037/0096-3445.131.2.163

Lewandowsky, S., & Kirsner, K. (2000). Knowledge partitioning: Context-dependent use of expertise. *Memory & Cognition, 28,* 295–305. doi:10.3758/BF03213807

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association, 92,* 648–655.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association, 103,* 410–423. doi:10.1198/016214507000001337

Little, D. R., & Lewandowsky, S. (2009). Beyond nonutilization: Irrelevant cues can gate learning in probabilistic categorization. *Journal of Experimental Psychology: Human Perception and Performance, 35,* 530–550. doi:10.1037/0096-1523.35.2.530

Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review, 95,* 492–527. doi:10.1037/0033-295X.95.4.492

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review, 109,* 376–400. doi:10.1037/0033-295X.109.2.376

Loomes, G., Moffat, P. G., & Sugden, R. (2002). A microeconomic test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty, 24,* 103–130. doi:10.1023/A:1014094209265

Lovett, A. (1988). Choice. In J. R. Anderson & C. Lebiere (Ed.), *The atomic components of thought* (pp. 255–296). Mahwah, NJ: Erlbaum.

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review, 118,* 316–338. doi:10.1037/a0022684

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28,* 3049–3067. doi:10.1002/sim.3680

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. doi:10.1023/A:1008929526011

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118,* 393–437. doi:10.1037/a0024143

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43,* 679–690. doi:10.3758/s13428-010-0049-5

Mata, R., von Helversen, B., & Rieskamp, J. (2011). When easy comes hard: The development of adaptive strategy selection. *Child Development, 82,* 687–700. doi:10.1111/j.1467-8624.2010.01535.x

Meng, X. L. (1994). Posterior predictive *p*-values. *Annals of Statistics, 22,* 1142–1160. doi:10.1214/aos/1176325622

Milinski, M., & Wedekind, C. (1998). Working memory constrains human cooperation in the prisoner's dilemma. *PNAS: Proceedings of the National Academy of Sciences, USA, 95,* 13755–13758. doi:10.1073/pnas.95.23.13755

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. doi:10.1006/jmps.1999.1283

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4,* 79–95. doi:10.3758/BF03210778

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116,* 499–518. doi:10.1037/a0016104

Newell, B. R. (2005). *Re*-visions of rationality? *Trends in Cognitive Sciences, 9,* 11–15. doi:10. 1016/j. tics.2004.11.005

Newell, B. R., & Lee, M. D. (2011). The right tool for the job? Comparing an evidence accumulation and a naïve strategy selection model of decision making. *Journal of Behavioral Decision Making, 24,* 456–481. doi:10.1002/bdm.703

Nilsson, H., Rieskamp, J., & Wagenmakers, E.-J. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology, 55,* 84–93. doi:10.1016/j.jmp.2010.08.006

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 3–27. doi:10.1037/0096-1523.17.1.3

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375–402.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review, 101,* 53–79. doi:10.1037/0033-295X.101.1.53

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. A. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective, & Behavioral Neuroscience, 1,* 360–370. doi:10.3758/CABN.1.4.360

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552. doi:10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The use of multiple strategies in judgment and choice. In N. J. Castellan (Ed.), *Individual and group decision making* (pp. 19–39). Hillsdale, NJ: Erlbaum.

Persson, M., & Rieskamp, J. (2009). Inferences from memory: Strategy- and exemplar-based judgment models compared. *Acta Psychologica, 130,* 25–37. doi:10.1016/j.actpsy.2008.09.010

Piaget, J. (1952). *The origins of intelligence in children*. New York, NY: International University Press. doi:10. 1037/11494-000

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421–425. doi:10.1016/S1364-6613(02)01964-2

Pitt, M. A., Myung, I. J., & Zhang, S. B. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. doi:10.1037/0033-295X.109.3.472

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 11–196). Cambridge, England: Blackwell.

Rieskamp, J. (2006). Perspectives of probabilistic inferences: Reinforcement learning and an adaptive network compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 1355–1370. doi:10.1037/0278-7393.32.6.1355

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica, 127,* 258–276. doi:10.1016/j.actpsy.2007.05.004

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135,* 207–236. doi:10.1037/0096-3445.135.2.207

Rissanen, J. (1986a). A predictive least-squares principle. *IMA Journal of Mathematical Control and Information, 3,* 211–222. doi:10.1093/imamci/3.2-3.211

Rissanen, J. (1986b). Stochastic complexity and modeling. *Annals of Statistics, 14,* 1080–1100. doi:10.1214/aos/1176350051

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 49,* 223–239.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42,* 40–47. doi:10.1109/18.481776

Rissanen, J. (2007). *Information and complexity in statistical modeling*. New York, NY: Springer.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108,* 370–392. doi:10.1037/0033-295X.108.2.370

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12,* 573–604. doi:10.3758/BF03196750

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Sewell, D. K., & Lewandowsky, S. (2011). Restructuring partitioned knowledge: The role of recoordination in category learning. *Cognitive Psychology, 62,* 81–122. doi:10.1016/j.cogpsych.2010.09.003

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association, 88,* 486–494. doi:10.1080/01621459.1993.10476299

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica, 7,* 221–242.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284. doi:10.1080/03640210802414826

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8,* 481–520. doi:10.1016/0010-0285(76)90016-5

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current Directions in Psychological Science, 3,* 1–5. doi:10.1111/1467-8721.ep10769817

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In T. Simon & G. Halford (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31–76). Hillsdale, NJ: Erlbaum.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41,* 1–20. doi:10.1146/annurev.ps.41.020190.000245

Speekenbrink, M., & Shanks, D. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General, 139,* 266–298. doi:10.1037/a0018620

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 36,* 111–147. doi:10.2307/2984809

Stone, M. (1977). Asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 39,* 44–47.

Todd, P. M., & Gigerenzer, G. (2001). Shepard's mirrors or Simon's scissors? *Behavioral and Brain Sciences, 24,* 704–705. doi:10.1017/S0140525X01650088

Todd, P. M., Gigerenzer, G., & the ABC Research Group. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

Townsend, J. T. (1975). The mind–body equation revisited. In C. Cheng (Ed.), *Philosophical aspects of the mind–body problem* (pp. 200–218). Honolulu, HI: Honolulu University Press.

Tversky, A., & Kahneman, D. (1981, January 30). The framing of decisions and the psychology of choice. *Science, 211,* 453–458. doi:10.1126/science.7455683

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550–592. doi:10.1037/0033-295X.108.3.550

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review, 111,* 757–769. doi:10.1037/0033-295X.111.3.757

van der Maas, H. L. J., Quinlan, P. T., & Jansen, B. R. J. (2007). Towards better computational models of the balance scale task: A reply to Shultz and Takane. *Cognition, 103,* 473–479. doi:10.1016/j.cognition.2007.01.009

Vandierendonck, A. (1995). A parallel rule activation and rule synthesis model for generalization in category learning. *Psychonomic Bulletin & Review, 2,* 442–459. doi:10.3758/BF03210982

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498. doi:10.1016/j.jmp.2010.07.003

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General, 137,* 73–96. doi:10.1037/0096-3445.137.1.73

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17,* 228–243. doi:10.1037/a0027127

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50,* 149–166. doi:10.1016/j.jmp.2006.01.004

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology, 60,* 158–189. doi:10.1016/j.cogpsych.2009.12.001

Wallace, C. S., & Dowe, D. L. (1999). Refinements of MDL and MML coding. *Computer Journal, 42,* 330–337. doi:10.1093/comjnl/42.4.330

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (in press). A default Bayesian hypothesis test for ANOVA designs. *American Statistician*.

Yang, L. X., & Lewandowsky, S. (2004). Knowledge partitioning in categorization: Constraints on exemplar models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 1045–1064. doi:10.1037/0278-7393.30.5.1045

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233–243). Amsterdam, the Netherlands: North-Holland.

(*Appendices follow*)

Appendix A

Model Implementation for the Categorization Data by Yang and Lewandowsky (2004)

The two continuous, arbitrarily scaled dimensions x and *y* of the 40 novel transfer items in the experiment by Yang and Lewandowsky (2004) were normalized. Next, all items below the diagonal (i.e., $\gamma < x - 0.5$) were mirrored along the identity vector y = x. This way, items above the hypothetical boundary vector f(x) = x + 0.93 belonged to Category *A,* and items below belonged to Category *B*. In this transformed stimulus space, the true parallel boundaries rule (i.e., the all-purpose model) was formalized as a single boundary vector with unity slope $\hat{\gamma}_i = \alpha + x_i$, where $\hat{\gamma}_i$ is the prediction for each item *i* and the intercept α is a free parameter of the model that determines where individuals set the boundary.

The probability of a correct classification was modeled based on an exponential choice function (i.e., Luce's choice rule),

$$p(A)_i = \frac{1}{1 + e^{\theta \cdot d_i}},$$

where $p(A)_i$ is the probability of predicting Category *A,* θ is a free parameter indicating the degree of error when applying the boundary rule, and d_i is the difference between the true and the predicted γ value, calculated as $d_i = \gamma_i - \hat{\gamma}_i$.The model is estimated by comparing $p(A)$ against participants' actual answers. The alternative, context-dependent categorization model was similar except that *all* items within one of the two categories were mirrored along the diagonal.

Appendix B

Conceptualization of the Exemplar Model

The exemplar model proposed by Juslin, Jones, Olsson, and Winman (2003) predicts people's choices based on the assumption that a person choosing between two options retrieves similar choice situations from memory. In particular, a pair of options (exemplars) is retrieved where each option is described by a vector of cue values that can be positive (i.e., +), negative (i.e., −), or unknown (i.e., ?). The retrieved exemplars can be described by a so-called cue configuration. For each cue in a configuration, nine combinations of cue values are possible (i.e., +/+, +/−, +/?, −/+, etc.). When making inferences, the cue configuration of the current pair (probe) is compared to the configuration of all previous pairs (*exemplars*) by determining the similarity between the configurations, defined as $s(x,y) = \Pi_{m=1}^{M} d_{xym}$, where d_{xym} takes a value of 1 if the combination of cue values of the probe *x* corresponds with the combination of cue values of the exemplar *y* for cue *m*; otherwise, it takes the value s_m, which is an attention weight parameter varying between 0 and 1 (cf. Juslin, Jones, et al., 2003). The attention weights represent the cues' subjective importance; the smaller the value, the greater the importance. For simplicity, we assumed that the attention weights are identical for all cues (see Persson & Rieskamp, 2009; von Helversen & Rieskamp, 2008). Finally, the probability that the first option A from the option pair A and B will be chosen is determined by

$$p(A) = \frac{\sum_{j \in A} s(x, y_j)}{\sum_{j \in A} s(x, y_y) + \sum_{j \in B} s(x, y_j)},$$

where the index $j \in$ A denotes that the sum is reached over all exemplars y_j where Option A was the correct choice, whereas the index $j \in$ B denotes that the sum is reached over all exemplars y_j where Option B was the correct choice.

Received October 25, 2011
Revision received August 30, 2012
Accepted August 30, 2012 ■