Judgment and Decision Making, Vol. 6, No. 8, December 2011, pp. 814–820

Methodological notes on model comparisons and strategy classification: A falsificationist proposition

Morten Moshagen* Benjamin E. Hilbig†

Abstract

Taking a falsificationist perspective, the present paper identifies two major shortcomings of existing approaches to comparative model evaluations in general and strategy classifications in particular. These are (1) failure to consider systematic error and (2) neglect of global model fit. Using adherence measures to evaluate competing models implicitly makes the unrealistic assumption that the error associated with the model predictions is entirely random. By means of simple schematic examples, we show that failure to discriminate between systematic and random error seriously undermines this approach to model evaluation. Second, approaches that treat random versus systematic error appropriately usually rely on relative model fit to infer which model or strategy most likely generated the data. However, the model comparatively yielding the best fit may still be invalid. We demonstrate that taking for granted the vital requirement that a model by itself should adequately describe the data can easily lead to flawed conclusions. Thus, prior to considering the relative discrepancy of competing models, it is necessary to assess their absolute fit and thus, again, attempt falsification. Finally, the scientific value of model fit is discussed from a broader perspective.

Keywords: falsification, error, model testing, model fit.

1 Introduction

The comparative evaluation of theories is an issue of fundamental importance in all sciences. In general, many disciplines proceed by submitting a particular theory or derived hypothesis to empirical tests and evaluating it through the logic of verification and falsification. Although such tests can be constructed to differentiate between models (experimentum crucis) given that opposing predictions can be derived (Platt, 1964), it is more common that their comparison proceeds more indirectly. Specifically, underlying assumptions or predictions derived from each particular model are tested independently. Over time, instances of confirmation and disconfirmation are accumulated for each model. According to the classical falsificationist logic (Popper, 1959), a model that repeatedly fails relevant tests is eventually discarded. Thereby, the question of which is the better theory or model is answered indirectly: In the long run, it is the model which makes testable and falsifiable predictions and endures critical tests of these. There are numerous implementations of this approach in JDM research and well-stated arguments have been formulated in favor of testing critical properties or central assumptions of single models (for recent examples see Birnbaum, 2008; Fiedler, 2010). Indeed, a typical variant is to conduct series of investigations which successively shed light on the determinants and/or bounding conditions of certain effects or theories.

However, discontent with testing properties of single models in isolation has been voiced. The line of argument can be summarized as follows (Gigerenzer & Brighton, 2009; Marewski & Olsson, 2009): It is problematic to test a specific hypothesis derived from a single model against the indefinite number of unspecified alternatives. Rather, it is argued that we need to compare alternative models directly. In line with such arguments, a popular approach is to specify several competing models and directly compare these in terms of their ability to account for empirical data (Shiffrin, Lee, Kim, & Wagenmakers, 2008). One particular variant specific to JDM research is the strategy classification approach which attempts to identify the decision strategy an individual most likely used (Bröder, 2000, 2002; Rieskamp & Hoffrage, 2008; Rieskamp & Otto, 2006). Following the idea that people adaptively select from a set of strategies (Gigerenzer & Selten, 2001; Payne, Bettman, & Johnson, 1988, 1993), models are compared on the level of individual subjects[1]

*University of Mannheim, Schloss, EO 254, 68133 Mannheim, Germany. Email: moshagen@uni-mannheim.de.

†University of Mannheim, Germany, and Max-Planck Institute for Research on Collective Goods, Germany.

[1]It has repeatedly been stated that individual-level analyses are preferable over aggregate results (Gigerenzer & Brighton, 2009; Glöckner, 2009; Pachur, Bröder, & Marewski, 2008), given clear individual differences in judgment and decision making (e.g., Bröder, 2003; Hilbig, 2008a). In essence, however, neither an individual-level nor an aggregate view should be considered superior per se (Cohen, Sanborn, & Shiffrin, 2008), and converging evidence from both views certainly is most conclusive (Hilbig, Erdfelder, & Pohl, 2011).

and the superior model is retained as a description of how the decision maker proceeded.

In the current paper, we focus on comparative model testing in general and the more JDM-specific procedure of strategy classification in particular. Following the notion that a good test of a theory is one that implements a sufficiently high hurdle to be overcome by this theory (e.g., Meehl, 1967), we identify two major shortcomings in existing approaches to comparative model evaluation: (1) failure to distinguish between random and systematic error and (2) neglect of global model fit. As we will argue and demonstrate, these seriously question the conclusions that may be drawn.

2 Systematic versus random error

One approach to model evaluation is to assess which of several models makes most correct predictions in terms of observable choices. The adherence rate denotes the proportion of observed choices that are in line with the predictions of a model, given that the latter makes a prediction.[2] For example, the recognition heuristic (Goldstein & Gigerenzer, 2002) predicts that people choose recognized over unrecognized options when judging which scores higher on some criterion (e.g., which of two cities has more inhabitants). The rate of adherence to this heuristic is simply the proportion of cases in which a participant chose the recognized option, while the error rate is defined as the proportion of choices that conflict with the heuristic's predictions (i.e., 100% minus the adherence rate). When we compare competing models, we regard the one yielding the highest adherence rate as the data generating model (e.g., Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, 2010). At the same time, a model need not yield perfect adherence (100%), because choices will be marred by some execution errors resulting from demands of the task, fatigue, slips of the finger etc.

The question of which maximal error rate a model should be allowed to produce is subject to idiosyncrasies of researchers, however. Since an adherence rate of 50% would be observed for purely random patterns in binary choices, this is the lowest useful criterion (Glöckner, 2009; Rieskamp, 2008). However, for choice patterns approaching simple random responding, it would be dubitable to conclude systematic execution of any strategy at all. Some have therefore suggested applying stricter criteria (Bröder & Schiffer, 2003; Glöckner, 2009). Nonetheless, a general reservation against applying a single error-threshold to all models or strategies is that their application may be not equally difficult, and that the amount of execution errors may also depend on the particular task.[3]

Irrespective of the error threshold applied, adherence rates make a strong and very problematic assumption regarding the type of error which occurs. It is implicitly taken for granted that the error is entirely random and that only its average size—across all items or trials—matters. In the above example, it is merely considered *how many* of, say, 100 paired-comparison choices the recognition heuristic predicts correctly. However, an at least equally relevant question is *which* of these 100 choices are explained by the model. The adherence rate ignores the latter aspect. As an upshot, model refutation becomes extremely difficult: Since almost any (completely implausible) model can easily produce above-chance-level adherence rates (Hilbig, 2010b), how should we expect to falsify a model? By contrast, assessment of whether a model or strategy adequately describes observed choices is a question of the degree of *systematic* error. The crucial question is not merely how much overall error a model implies, but whether the error really is random and, consequently, of equal magnitude across all items. The main flaw inherent in adherence rates is the neglect of different item types and their respective error-rates. Only by considering these separately can we identify systematic error.

Returning to the above example, the recognition heuristic often yields adherence rates greater than 80% (Pachur et al., 2008; Pohl, 2006) and thus a relatively small average error. However, as argued above, the crucial question is whether the probability of choosing as predicted by the recognition heuristic is roughly 80% across *all* items (allowing for attributing the remaining 20% to random error). In fact, however, this is not the case (Hilbig & Pohl, 2008; Pohl, 2006). Participants often adhere to the recognition heuristic whenever it implies a factually correct choice (e.g., when the recognized of two cities really has more inhabitants than the unrecognized one), but deviate from the heuristic's prediction whenever the choice it implies is factually false (Hilbig, Pohl, & Bröder, 2009). Thus, their adherence varies systematically as a function of item type, which is why it is entirely inappropriate to attribute non-adherence to random error only (Hilbig, Erdfelder, & Pohl, 2010; Hilbig, 2010b).

[2] As a result, the adherence rate will often lead to comparing apples with oranges (Hilbig, 2010b): One model may produce high adherence but make predictions only in very few cases; another may yield slightly less adherence but actually make a prediction in the majority of cases. Clearly, the former need not be considered superior at all.

[3] A viable approach would be to estimate a to-be-expected error-rate for each strategy from independent data: Specifically, each participant could be taught one particular strategy and it could then be assessed how many trials she needs before consistently solving problems in line with this strategy (or how many errors she makes). This could then serve as an estimate for the difficulty associated with executing a strategy.

For exactly these reasons, researchers have specifically tested whether models yield equal adherence rates across different types of items (e.g., Bröder & Eichler, 2006; Hilbig, 2008b; Newell & Fernandez, 2006; Richter & Späth, 2006). These studies represent clear instances of model refutation by testing the critical property of unsystematic errors across experimentally manipulated types of items. The overall adherence rate, by contrast, is uninformative, unlikely to provide an instance of model refutation, and often biased (Hilbig, 2010a). Therefore, models cannot be evaluated (let alone compared to each other) by merely considering whether choices deviate from model predictions. Although this point has been recognized before (e.g., Bröder & Schiffer, 2003), current JDM articles fail to treat it appropriately (e.g., Brandstätter, Gigerenzer, & Hertwig, 2008; Marewski et al., 2010).

Table 1: Cue patterns for three item types and choice predictions of strategies taken from Bröder and Schiffer (2003).

	Item type 1		Item type 2		Item type 3	
	A	B	C	D	E	F
Cue 1	1	0	1	0	1	0
Cue 2	1	1	0	1	1	1
Cue 3	1	0	0	1	1	1
Cue 4	0	1	0	0	0	1
Predictions:						
WADD	A		D		E	
EQW	A		D		Guess	
TTB	A		C		E	

3 Neglect of global model fit

Addressing this severe shortcoming of adherence rates, several researchers developed methods assessing which model or strategy best accounts for the data while allowing for a certain degree of random errors only (Bröder & Schiffer, 2003; Glöckner, 2009; Jekel, Nicklisch, & Glöckner, 2010; Rieskamp, 2008). For each model considered, the empirical distance of the predicted pattern from the observed pattern is measured by means of a distance function (usually a log-likelihood value or a transformation thereof such as the Bayesian Information Criterion, BIC). Because the error is constrained to be equal across all item types, systematic errors lead to model misfit and are thus penalized. The best-fitting model is then deemed to reflect the actual decision making process or strategy.

Despite the indubitable superiority of such procedures over the mere comparison of adherence rates, this particular procedure also bears a caveat: Reliance on relative model fit as criterion requires that the data generating model is among the competitors. However, the observed data may not have been generated by *any* of the models considered, and the model yielding the smallest discrepancy may be entirely invalid (Gelman & Rubin, 1995; Roberts & Pashler, 2000; Zuccini, 2000). Like the average adherence rate, relative model fit is unlikely to allow for model refutation. One model will always fit the data best. Without the ability to falsify candidate models, researchers may uphold a model that is merely least false but still far from adequate.

Although this issue has been openly acknowledged (Bröder & Schiffer, 2003; Glöckner, 2009), no conclusive efforts have tackled the problem. Fortunately, however, it is easy to assess whether a particular model *might* have generated the data: Prior to preferring a certain model over others by drawing on relative fit, we need to establish that each is able to account for the observed data, through testing global goodness-of-fit. Thus, instead of taking for granted the vital requirement that a model by itself should adequately describe the data, we need to test this assumption. As we demonstrate below, failure to consider absolute model fit can easily lead to flawed conclusions.

To illustrate this point, consider the judgment situation depicted in Table 1 (Bröder & Schiffer, 2003): Decision makers infer which of two options (A or B) is superior in terms of some criterion given four probabilistic binary cues (for applications of such task structures see Bröder & Schiffer, 2006; Glöckner & Betsch, 2008; Rieskamp & Hoffrage, 2008). For example, the task might be to judge which of two cities, A or B (options), has more inhabitants (criterion) based on whether or not a city has an international airport, is state capitol, has a university, and has a major-league football team (probabilistic binary cues with different predictive validity, cf. Gigerenzer & Goldstein, 1996). There are three item types (choices between A/B, C/D and E/F in Table 1) which differ in the cue patterns, constructed so as to differentiate between three candidate decision strategies: A weighted additive strategy (WADD; choose the option with the higher sum of cue values weighted by their validities), an equal weight strategy (EQW; choose the option with the higher sum of positive cue values), and a lexicographic take-the-best strategy (TTB; consider cues in order of their validity; choose according to first discriminating cue). Table 1 shows the choice predictions of each strategy.

Using this set-up, we ran a series of simulations, mostly mirroring the procedures of Bröder and Schiffer (2003). We first let each of the three strategies (WADD, EQW, and TTB) generate 1,000 data sets (simulated decision makers) with 30 choices per item type and a constant random error rate of 10%. Additionally, 1,000 data

Table 2: Simulation results generating data by different strategies and classifying data sets by means of the BIC.

	Classification [%]			
Data generated by	WADD	EQW	TTB	Unclassified
WADD + 10% error	**99.6**	0.4	0.0	0.0
EQW + 10% error	2.1	**97.9**	0.0	0.0
TTB + 10% error	0.0	0.0	**100.0**	0.0
Random	0.0	0.1	0.0	**99.9**
3C + 10% error	43.2	0.0	41.8	**15.0**

Note. The target category is highlighted in bold.

sets were generated by a pure guessing strategy to rule out that some strategies would fit random data. However, our main argument raised above is that conclusions based on the mere assessment of relative fit will be flawed if the data generating strategy is *not* within the set of those considered. Thus, we additionally simulated 1,000 data sets generated by a three-cue strategy (3C; compare choice options on each cue; choose the first option to reach three positive cue values). Across the three item types, it predicts the choice pattern "A, guess, E" which is distinct from those of the other strategies (see Table 1). Note that any other strategy could have been used for this demonstration, as long as it predicts a choice pattern distinct from the strategies under consideration.[4]

Parameter estimation for each strategy and data set proceeded by minimizing the log-likelihood ratio statistic G^2 by means of the EM algorithm (Hu & Batchelder, 1994) as implemented in the multiTree software tool (Moshagen, 2010). Following the recommendation of Glöckner (2009), a strategy was no longer considered if it required an average error of 30% or larger. Then, the strategy yielding the smallest BIC was chosen for classification.

The results displayed in Table 2 mirror those of Bröder and Schiffer (2003): For data generated by either of the strategies *within* the set, classifications were almost perfect. The reliability of such classifications can be assessed through the Bayes Factor, which expresses the posterior odds in favor of one model compared to another, given the data (Wagenmakers, 2007). The Bayes Factors between the best and the second best fitting strategy were > 3 (implying positive evidence, Raftery, 1995) for more than 95% of all classifications, suggesting that these were highly reliable. At the same time, random data generation led to practically all data sets remaining unclassified, as is desirable.

However, once data were generated by a strategy *outside* of the set considered, there were substantial misclassifications. When the 3C strategy was the true underlying model, the optimal outcome would have been that no single data set is classified. However, about 85% of data sets were actually classified—all as WADD and TTB in about equal proportions (final row of Table 2). Once again, Bayes Factors were > 3 comparing the best and second best fitting model in over 99% of data sets. Thus, most data sets were clearly and reliably classified, even though no single one was generated by any of the strategies under consideration. Given that researchers can rarely claim to know whether the data generating strategy is in fact within their set, this finding seriously questions any conclusion drawn from such a model comparison procedure based on relative fit.

As a remedy, we call for initially considering absolute model fit to determine whether a model is consistent with the data. Because the classification method of Bröder and Schiffer (2003) is a member of the family of multinomial processing tree models (Batchelder & Riefer, 1999; Erdfelder et al., 2009), absolute model fit for each strategy can be determined by evaluating the asymptotically chi-square distributed log-likelihood ratio statistic G^2 (Hu & Batchelder, 1994).[5] In the above example with the 3C strategy generating the data, using a conventional type-I error of .05[6] yielded exclusion of 95.5% of data sets as unclassifiable. Thus, rather than wrongly considering about 85% of decision makers as WADD or TTB-users, the vast majority was treated appropriately and left unclassified. At the same time, introducing an absolute-fit-threshold only marginally affected classifications whenever data generation followed one of the strategies within the set: Non-classification rates were 7.6%, 5.7%, and 10.3% for data generated by WADD, EQW, and TTB, respectively. As this exercise demonstrates, using absolute model fit to refute candidate models prevents false classifications if the data generating model is not in the set of those considered. At the same time, if the true model is within the set, classifications are only slightly more conservative.

[4] Once several strategies make exactly *the same* choice predictions, there is no possibility to discriminate between them. The only remedy would then be to look for further item types which can discriminate between strategies or—if strategies inherently make the same choice predictions—consider further dependent measures such as reaction times or confidence judgments (Glöckner, 2009; Jekel et al., 2010).

[5] It might be argued that sparse cell counts resulting from individual level analyses question the assumption of an asymptotic chi-square distribution of the G^2 statistic. However, the exact distribution of G^2 under H0 can be easily estimated by means of the parametric bootstrap (Efron & Tibshirani, 1993).

[6] Since multiple disjoint hypothesis are tested using the same data set, it may be reasonable adjust the type-I error of each single test to arrive at a desired family-wise error. Similarly, to avoid rejection of an otherwise correct model or strategy based on trivial differences when the statistical power is very high, a compromise power analyses (Erdfelder, Faul, & Buchner, 2005; Faul, Erdfelder, Lang, & Buchner, 2007) can be applied, adjusting the type-I error such that it is equal to the type-II error.

4 Model fit and validity

If a model or strategy is found to fit the data in absolute terms and also outperforms other (fitting) models, can we conclude that the model under investigation is correct? Unfortunately, not. The seminal work of Wason (1968) provides an instructive example of this fallacy: When asked to identify a rule underlying a sequence of numbers such as "2, 4, 8", people find it difficult to identify the generality of the underlying rule and tend to test overly specific rules such as "the previous number multiplied by two". Although such specific rules may perfectly describe the given sequence, the actual data generating rule may be much more general (e.g., "a triple of numbers"). Thus, considering a fitting model to be the data-generating one is an instance of the classical logical fallacy of affirming the consequent (Trafimow, 2009): The rule "if the model is correct, then the model will fit the data" cannot imply the reverse "the model fits the data, therefore it is correct". If a model fits the data, it is only a candidate that *may* have generated the data (although the validity of this assertion can be made more or less plausible by drawing on additional tests).

More generally speaking, even a perfectly fitting model need not be valid, because at least one of its core assumptions may be entirely wrong (Roberts & Pashler, 2000), and there may also be "infinitely many theoretically distinct models that fit the data equally well" (Voss, Rothermund, & Voss, 2004, p. 1217). In essence, "the danger is [...] to use a good fit as a surrogate for a theory" (Gigerenzer, 1998, p. 200), as model fit is only ever necessary but never sufficient for model validity. In turn, regardless of whether the underlying assumptions of a model are theoretically and empirically justifiable, misfit provides an instance of falsification with regard to the model in question. However, the diagnosticity of model misfit is also limited, because it may stem from very different sources that do not necessarily invalidate the core assumptions of a certain model. Even repeated occurrences of misfit may still be due to inappropriate auxiliary assumptions that are of little relevance to the core ideas of a model (Lakatos, 1970). Nevertheless, model (mis)fit must be acknowledged and failure to fit the data calls for model refinement in the very least.

Since the conclusions that can be drawn from model comparisons—even if they include testing for systematic error appropriately and are based on assessment of absolute model fit as we have called for herein—are necessarily limited, model evaluations will always need to be complemented by other test of critical model properties or tests of competing hypothesis derived from different models (Glöckner & Herbold, 2011; Hilbig & Pohl, 2009; Roberts & Pashler, 2000).[7]

[7]The diagnostic paucity of such model comparisons also implies that theory development should not merely pursue in a statistical bottom-up fashion.

5 Conclusion

In the present paper, we identified two major shortcomings of existing approaches to comparative model evaluation, namely (1) failure to distinguish between random and systematic error and (2) neglect of global model fit. Both of these lessen the chances to falsify models and therefore increase the dangers of drawing inadequate conclusions. The first point refers to studies comparing models by means of the average adherence across items, that is, the proportion of choices in line with a model's predictions (e.g., Brandstätter, Gigerenzer, & Hertwig, 2008; Marewski et al., 2010) or similar measures such as majority choices (Brandstätter, Gigerenzer, & Hertwig, 2006; Kahneman & Tversky, 1979). The second applies to superior approaches which panelize systematic error and compare models based on their relative ability to account for the data (Bröder & Schiffer, 2003; Glöckner, 2009; Rieskamp, 2008). These can warrant valid conclusions only if the data-generating model is in the set of those compared. However, this is typically unknown and whenever it is not the case, failure to test whether models are able to adequately describe observed data in terms of absolute goodness-of-fit can lead to false conclusions.

In summary, we propose to retain the logic of falsification—as is well-implemented when testing critical properties of single models (Birnbaum, 2008; Fiedler, 2010)—when comparing models or strategies in terms of fit. Misfit of a model represents an instance of falsification and should exclude this model from consideration in a model comparison. This, in turn, will secure conclusions drawn from model comparisons against the daunting possibility that "in the land of the blind, the one-eyed [model] is made king". Moreover, we advocate testing critical properties or central assumptions of models directly, instead of pursuing blind competitions. The higher we set the hurdles for our models, the more confidence we can have in those which stand the test of time.

References

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

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115*, 463–501.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). Making choices without trade-offs: The priority heuristic. *Psychological Review, 113*, 409–432.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2008). Risky choice with heuristics: Reply to Birnbaum

(2008), Johnson, Schulte-Mecklenbeck, and Willemsen (2008), and Rieger and Wang (2008). *Psychological Review, 115*, 281–289.

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.

Bröder, A. (2002). Take the Best, Dawe's Rule, and Compensatory Decision Strategies: A Regression-based Classification Method. *Quality & Quantity, 36*, 219–238.

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 611–625.

Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica, 121*, 275–284.

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

Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 904–918.

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

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/ Journal of Psychology, 217*, 108–124.

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1565–1570). Chichester, UK: Wiley.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Fiedler, K. (2010). How to study cognitive decision algorithms: The case of the priority heuristic. *Judgment and Decision Making, 5*, 21–32.

Gelman, A. & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25*, 165–173.

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology, 8*, 195–204.

Gigerenzer, G. & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107–143.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox*. Cambridge, MA: The MIT Press.

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

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 34*, 1055–1075.

Glöckner, A., & Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making, 24*, 71–98.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75–90.

Hilbig, B. E. (2008a). Individual differences in fast-and-frugal decision making: Neuroticism and the recognition heuristic. *Journal of Research in Personality, 42*, 1641–1645.

Hilbig, B. E. (2008b). One-reason decision making in risky choice? A closer look at the priority heuristic. *Judgment and Decision Making, 3*, 457–462.

Hilbig, B. E. (2010a). Precise models deserve precise measures: a methodological dissection. *Judgment and Decision Making, 5*, 272–284.

Hilbig, B. E. (2010b). Reconsidering "evidence" for fast-and-frugal heuristics. *Psychonomic Bulletin & Review, 17*, 923–930.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2010). One-reason decision-making unveiled: A measurement model of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*, 123–134.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37,* 827–839.

Hilbig, B. E., & Pohl, R. F. (2008). Recognizing users of the recognition heuristic. *Experimental Psychology, 55*, 394–401.

Hilbig, B. E., & Pohl, R. F. (2009). Ignorance- versus evidence-based decision making: A decision time analysis of the recognition heuristic. *Journal of Exper-*

imental Psychology: Learning, Memory, and Cognition, 35, 1296–1305.

Hilbig, B. E., Pohl, R. F., & Bröder, A. (2009). Criterion knowledge: A moderator of using the recognition heuristic? *Journal of Behavioral Decision Making, 22*, 510–522.

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of engineering processing tree models with the EM algorithm. *Psychometrika, 59*, 21–47.

Jekel, M., Nicklisch, A., & Glöckner, A. (2010). Implementation of the Multiple-Measure Maximum Likelihood strategy classification method in R: addendum to Glöckner (2009) and practical guide for application. *Judgment and Decision Making, 5*, 54–63.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–292.

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgave (Eds.), *Criticism and the growth of knowledge* (pp. 91–195). Cambridge: Cambridge University Press.

Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G. (2010). From recognition to decisions: extending and testing recognition-based models for multi-alternative inference. *Psychonomic Bulletin & Review, 17*, 287–309.

Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Zeitschrift fur Psychologie/Journal of Psychology, 217*, 49–60.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science, 34*, 103–115.

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*, 42–54.

Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentially of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19*, 333–346.

Pachur, T., Bröder, A., & Marewski, J. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? *Journal of Behavioral Decision Making, 21*, 183–210.

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.

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

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science, 146*, 347–353.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making, 19*, 251–271.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111–163.

Richter, T., & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 150–162.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1446–1465.

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

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

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358–367.

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.

Trafimow, D. (2009). The theory of reasoned action: A case study of falsification in psychology. *Theory & Psychology, 19*, 501–518.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*, 1206–1220.

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

Wason, P. C. (1968). Reasoning about a rule. *The Quaterly Journal of Experimental Psychology, 20*, 273–281.

Zuccini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44*, 41–61.