

Berichte aus dem
Psychologischen Institut
der Universität Bonn

Band 28 (2002), Heft 1

**Bailing and jailing the unknown way:
A critical examination
of a study reported by Dhami and Ayton (2001)**

Arndt Bröder

Please cite this work as:

Bröder, A. (2002). Bailing and jailing the unknown way: A critical examination of a study reported by Dhami and Ayton (2001). *Berichte aus dem Psychologischen Institut der Universität Bonn*, 28 (1).

CONTENTS

1. Abstract / Zusammenfassung	4
2. Introduction	5
3. Potential pitfalls of comparative model fitting: A case study	6
4. Task and models	7
5. Parameter estimation and model fit	8
6. A simulation study	9
7. Conclusions	12
8. References	13

Diese Arbeit wurde durch die *Deutsche Forschungsgemeinschaft* (DFG, Az. Br 2130 / 1-1) unterstützt.

This research was supported by the *Deutsche Forschungsgemeinschaft* (DFG, grant Br 2130 / 1-1) to the author.

ABSTRACT

Behavioral decision research on multi-attribute decision making is plagued with the problem of drawing inferences about cognitive strategies based on behavioral data. "Comparative model-fitting" may be a viable alternative to traditional approaches like "Process Tracing" or "Structural Modeling". However, this methodology is not without its pitfalls when formal criteria for parameter estimation are missing. To illustrate the problem, the model evaluation procedure described by Dhami and Ayton (2001) is applied to simulated data sets. It is shown that the conclusions drawn from this procedure are not valid.

ZUSAMMENFASSUNG

Die deskriptive Entscheidungsforschung zu Multi-Attribut-Entscheidungen hat das Problem, von Verhaltensdaten auf kognitive Strategien zu schließen. "Vergleichende Modellanpassung" könnte eine brauchbare Alternative zu traditionelleren Ansätzen wie "Process Tracing" und "Structural Modeling" sein. Diese Methode ist jedoch nicht ohne Gefahren, wenn sie auf formale Rechtfertigung der Parameterschätzung verzichtet. Um das Problem zu illustrieren wird eine von Dhami und Ayton (2001) beschriebene Prozedur auf simulierte Daten angewendet. Es wird gezeigt, dass die Schlussfolgerungen aus dieser Prozedur nicht valide sind.

**Bailing and Jailing the unknown way: A critical examination of a study reported
by Dhami and Ayton (2001)**

Since psychologists have acknowledged people's behavioral deviations from „normative“ decision models like SEU theory, a multitude of decision rules and heuristics have been proposed in the literature that are cognitive models of information integration in multi-attribute decision problems. For example, this plethora of models contain noncompensatory ones like „Elimination by Aspects“ (Tversky, 1972), the „Lexicographic Rule“ (Fishburn, 1974), a variant of which is „Take The Best“, introduced by Gigerenzer and colleagues (Gigerenzer, Hoffrage, & Kleinbölting, 1991), the conjunctive and disjunctive decision rules (Einhorn, 1970), the „Satisficing Heuristic“ (Simon, 1957), or the "Matching Heuristic" (Dhami & Ayton, 2001). Compensatory models, on the other hand, are the „Additive Differences Rule“ (e.g. Tversky, 1969), the „Weighted Additive Rule“ or "Franklin's Rule" (Gigerenzer et al., 1999), „Dawes' Rule“ (equal weighting of attributes, Dawes & Corrigan, 1974) among others. This enumeration is far from being complete. A large list of different models is compiled in Svenson (1979, 1983). Gigerenzer et al. (1999) with their „adaptive toolbox“ approach have added even more heuristics to this collection. Since all these models formulate assumptions about cognitive information integration processes during decision making, their value as descriptive models of human thought has to be evaluated empirically which certainly involves the task of bridging the gap between these assumptions and potentially observable behavioral data. This bridging problem is of fundamental importance for psychology and other disciplines because „a theory that is not stated in terms of the data analysis to be used cannot be tested“ (Guttman, 1981, p.63). From a slightly different but similar perspective, Herbert Simon stated that „our methods for gathering data to test our theories must fit the formal shapes of the theories“ (Simon, 1992, p. 159). In a sense, Guttman's opinion implies Simon's and it constitutes a somewhat stronger criterion because he demands that theories themselves are incomplete if they do not refer to the data relevant for their own evaluation, whereas Simon just demands data collections that are formally related to the theories. However, I tend to prefer Simon's more pragmatic criterion because the precise „terms of data analysis“ receive their significance only in relation to specific data collection methods (e.g. experiments), and these are often not defined at the time of theory formulation.

Hence, the invention of clever testing methods often comes *after* theory formulation, but then these methods should ideally conform to Simon's criterion and avoid ad hoc methods whenever possible.

In research on multi-attribute decision making, the connection between the theories and the data for their evaluation often is rather loose. Many criticisms have been raised against the two dominating methodologies in the field that are often referred to as „Structural Modeling (SM)“ and „Process Tracing (PT)“ (see Harte & Koele, 2001, for a recent introduction). The manifold criticisms of both approaches are described in detail in Bröder (2000) and will not be reiterated here.

A third approach of increasing popularity might be called „comparative model fitting“ which aims at theory-based comparisons of the explanatory power of competing models for a given data set (e.g. Dhimi & Ayton, 2001; Dhimi & Harries, 2001; Hoffrage, Hertwig, & Gigerenzer, 2000; Rieskamp & Hoffrage, 1999). The advantages of this approach are that (1.) precise competing models have to be specified explicitly before data analysis and (2.) no ad hoc measures are needed to evaluate the models. However, often other ad hoc assumptions about model parameters have to be made which may lead to seriously distorted inferences about the underlying models. This will be particularly serious if no error theory accompanying the cognitive model is explicitly specified.

Potential pitfalls of comparative model fitting: A case study

Recently, Roberts and Pashler (2000) have criticized the widespread use of „model-fitting“ in psychological science and argued that a good fit of a model *per se*, defined as a high degree of successful data reconstruction, does not in itself guarantee a sound inference about the validity of the model. Such an inference is only valid if it can be shown that the model fit is clearly superior to the fit of other plausible competing models of comparable complexity (e.g. same number of parameters). Thus, all inferences about the validity of theories are made with reference to other theories. Unless we test our favourite models against „straw men“, the superior model fit certainly is a valuable argument in favour of the models.

In this sense, the work reported by Dhimi and Ayton (2001) is a positive example because they compared the fit of several plausible models to behavioral data.

In addition, their work is of great public interest because they tried to model bailing decisions of British court magistrates and drew the conclusion that many of these decisions were compatible with a simple, noncompensatory „Matching Heuristic“ (MH) which stands in opposition to the normative ideal of thoroughly weighting all evidence prescribed by the law.

Task and models

It is not easy to describe Dhami and Aytons (2001) complex investigation in a nutshell, but I will try to do it. The authors presented 27 hypothetical cases for bailing decisions to 81 British magistrates (some cases were presented repeatedly, but this will be ignored here for simplicity). The 27 cases consisted of 9 pieces of cue information each that were orthogonally varied in the sample. The information about each case was presented in a short text passage written in normal English. Information about gender, race, age, the bail decision of the police etc. were given. In the later analysis, the cues were dichotomized. For each hypothetical case, the magistrate's decision was registered („bail vs. jail“). In order to assess the individual decision strategies, three different models were formulated and fitted to actual decisions: Franklin's Rule (FR), Dawes' Rule (DR), and the Matching Heuristic (MH). Franklin's Rule is a compensatory strategy in which all cues are weighted according to their subjective importance and summed up. If this sum exceeds a certain critical value, a „jailing“ decision is made. Dawes' Rule is an equal weight linear model in which critical cue values are summed up regardless of their importance. If this sum exceeds a critical value, a punitive decision is made. Dawes' Rule is compensatory, too, although the relative importance of the cues is ignored. The Matching Heuristic, on the other hand, belongs to the family of fast and frugal decision heuristics in the tradition of Gigerenzer et al.'s (1999) simple heuristics approach. MH is actually a sophisticated disjunctive rule which takes into account the importance hierarchy of the cues: If there is any critical cue value in the k most important cues, a jailing decision is made. Hence, there are two simple stopping criteria in MH: A predefined criterion defines the maximum number of cue values looked up, and the other criterion stops search dependent on the cue values. If there is a critical one within the search set, information search is terminated. Here, k is a free parameter denoting the number of cues looked up (1 to 9 in the example), and therefore,

MH actually is a *family* of heuristics which differ in its actual value. FR, DR, and MH are considered as plausible models of information integration in this situation, and their ability to fit the bailing decisions was compared.¹

Parameter estimation and model fit

How was the model fit assessed? In this situation with a dichotomous criterion the fit of the models can easily be defined as the percentage of correct matches between model and actual decisions. This defines a simple distance criterion which must be minimized by choosing optimal parameter values. In other areas of model assessment, for example multinomial modeling (Hu & Batchelder, 1994; Riefer & Batchelder, 1988), the model parameters and the distance criterion are linked by a formal *distance function* which is minimized by appropriate parameter estimation methods. In order to create „predictions“ of the three models, several parameters had to be determined: Critical values, cue weights, and a criterion value must be estimated for Franklin's Rule; Critical values and a criterion are necessary for Dawes' Rule, and a cue importance hierarchy as well as critical values are necessary for MH. Since there were no prespecified values, these parameters had to be determined from the data somehow.

For each cue in Franklin's Rule, the critical cue value was defined as the one accompanied by a higher proportion of punitive decisions which in turn was used as the weight of the cue. For example, if 65% of male defendants were punished, whereas only 50% of female defendants were punished, „male“ was the critical value of gender, and the cue weight was 65%. The threshold of the weighted sum was determined by summing up the weighted sums of all 27 patterns and dividing by 27. Dhami and Ayton (2001, p. 154) write that „this is a reasonable method for calculating the threshold value because each magistrate made roughly an equal number of punitive and nonpunitive decisions“. However, if the authors intended to model this fact accurately, the *median* might have been a better choice because it is defined as the value splitting a sample in two halves. The best choice, in my opinion, would have been the percentile value of the *actual* proportion of punitive decisions of each judge. In this case, the fairest evaluation

¹ It has become popular to speak of successful model *predictions* instead of model *fits*. In my view, this term is completely misleading when models with free parameters which are estimated are fitted to empirical data after the fact. "Prediction" can only take place when the parameters are known or if hypotheses about their values are tested.

of Franklin's Rule would probably have resulted. But even this suggestion is an arbitrary one and not based on any formal relation between the model and the data (distance function), so I will not elaborate on it further. What I want to say is that this method of determining critical value and threshold is not justified formally and therefore, we have no guarantee that this procedure actually minimizes the (unknown) distance function which relates parameters to the simple criterion of an overlap between empirical decisions and those reproduced by a model with optimally chosen parameter values.

The critical cue values for Dawes' Rule were determined like those for Franklin's Rule, and the threshold was determined accordingly (mean number of critical cue values across the 27 cases). Hence, the same problem of potentially non-optimally chosen parameters arises here. The critical values and the cue validity hierarchy for MH were determined in a similar fashion.

After that, decisions across the 27 cases were „predicted“ by these models and the fit to the actual decisions was determined. Note, that nine different MH versions were tested ($k=1$ to 9), and the best fitting MH with the lowest k was chosen as the representative of MH. The result was that a high percentage of judges were classified as MH-users (32.1%) and that MH was the best fitting model altogether (73.98% decisions correctly fitted versus 73.57% and 69.36% for Franklin's and Dawes' Rules, respectively). The authors concluded „that the Matching Heuristic model best captured magistrates' bail decision making policies“ (Dhimi and Ayton, 2001, 158) and therefore „the present study has shown that the current practice of magistrates' bail decision making is far from ideal practice.“ (p. 163).² Dhimi and Ayton add a lengthy discussion about potential implications for decision research and the justice system.

A simulation study

Given these serious implications, one has to ask: How valid are Dhimi and Ayton's (2001) conclusions about the decision behavior of the judges? As has been suggested above, their method contained some arbitrary decisions in determining the

² It is peculiar that no significance test is reported to corroborate this statement, and the error bars in Exhibit 4 of Dhimi and Ayton (2001) suggest that there is apparently no reason to doubt that the null hypothesis of no differences between the fit of strategies can be maintained.

model parameters. The way of ascertaining cue weights, critical values, and decision criteria were not justified by any formal connection between the models and the measures actually used for these decisions. The lack of formal justification leaves the desideratum of at least an *empirical* validation of the fitting procedure, for example via simulation studies.

In the following, I will report such a simulation study which unfortunately suggests that the method employed by Dhimi and Ayton (2001) was not valid. But before, two formal arguments against the study must be mentioned: First, the competition between the models was not fair because the Matching Heuristic has one more free parameter, namely k . If any of the matching models ($k=1,2,\dots,9$) fitted the data better than Franklin's Rule or Dawes' Rule, the response pattern was classified as generated by the Matching Heuristic. The cue validity hierarchy was determined in the same fashion as the weights for FR. With these weights, FR can only reproduce one fixed proportion of bailing decisions, whereas MH can reproduce at least 9 proportions depending on the choice of the parameter k . The Matching Heuristic is simply a disjunctive decision rule including the k most important attributes, so by adding attributes, we only increase the proportion of bailing decisions in the stimulus sample, thus generating up to nine different proportions that can be „predicted“ by the heuristic, whereas Franklin's Rule and Dawes' Rule are restricted to one proportion! Hence, this will lead to a bias in favour of the Matching Heuristic because it is presumably much more flexible (Cutting, 2000). But besides from that, we simply do not know which consequences are implied by the ad hoc determination of the cue hierarchy which was based on percentages of punitive decisions. If we assume that even highly trained judges have occasional inconsistencies in their decisions, any random fluctuation may have an unknown effect on the values that are used to estimate the relative importance of the cues. This may decrease the chances of Franklin's Rule or Dawes' Rule to fit the data, even if they were generated by one of these two strategies.

In order to control for this possibility, we conducted a simulation study. One hundred data sets with the 27 hypothetical cases were generated by Franklin's Rule, Dawes Rule, a pure Random strategy, and the Matching Heuristic (with $k= 1$ to 6), respectively. That is, we created bailing decisions of hypothetical judges who followed a known strategy. If Dhimi and Ayton's (2001) model fitting method is valid, the

majority of response patterns should be classified according to the strategy which generated it.

The arbitrary cue weights for the nine cues were 1.0 to 9.0 for Franklin's Rule and 5.0 for each cue in Dawes' Rule. In addition, a random error component of on average 10% strategy-inconsistent choices was added to model occasional inconsistencies of the simulated judges. The thresholds varied randomly between 12.5 and 22.5 in a uniform fashion for Franklin's Rule and Dawes' Rule, respectively. Hence, we created data of which we knew *in advance* which strategy generated them. This made it possible to investigate the validity of the fitting procedure used by Dhami and Ayton. After data generation, the model fitting procedure was administered exactly as described in Dhami and Ayton (2001) and outlined above. The striking results can be found in Figure 1.

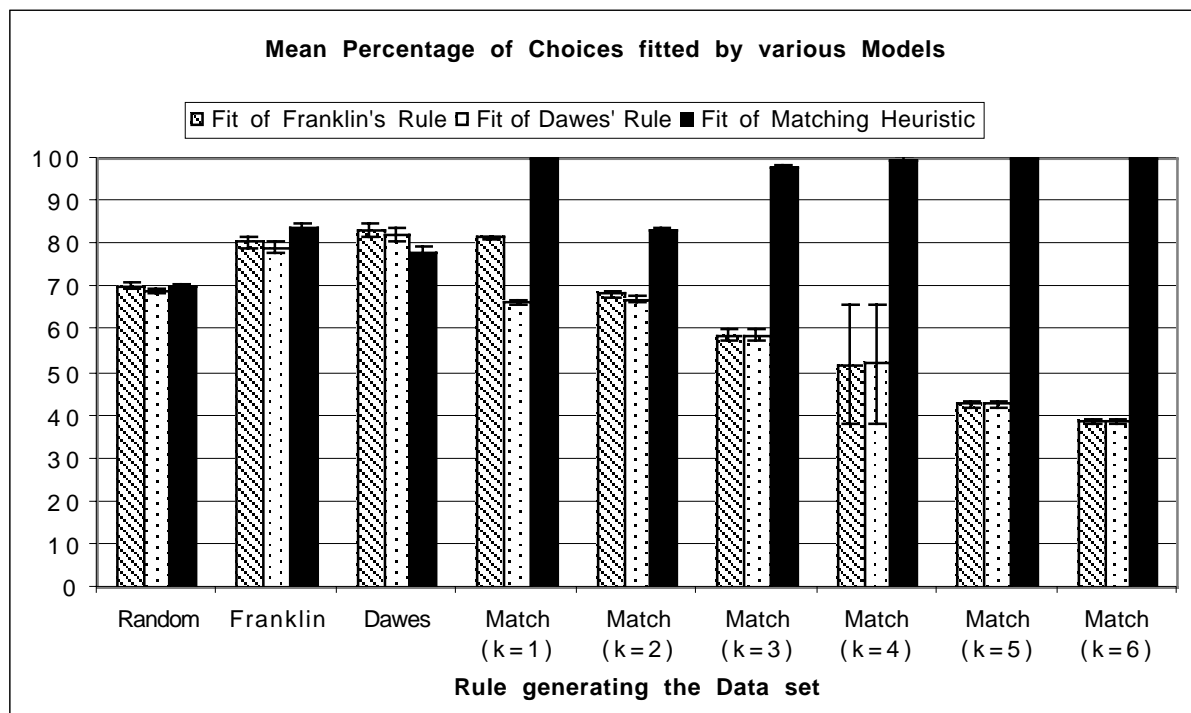


Figure 1: Mean percentage of choices fitted by various models when generated from known strategies

Figure 1 shows the mean percentage of choices fitted by Franklin's Rule, Dawes' Rule, and the Matching Heuristic ($k=1$ to 6), respectively. As can be seen, the Matching Rule nearly always conforms to the highest percentage of empirical choices with the

exception of data generated by Dawes' Rule. The Matching Rule always shows the best fit when data were actually created by this heuristic, but it also shows the best fit for data generated by Franklin's Rule as well as by a random guessing strategy. Strikingly, the Random Choice results are those most closely resembling the data presented by Dhami and Ayton (2001, Exhibit 4, p.158)!

Of course, I do not want to suggest that the judges in their study decided in a random fashion, but I want to emphasize the point that apparently the fitting method is not valid. In line with Cutting (2000) I think "that a model that performed better than another one in fitting random data, however badly, was doing something beyond what a psychological model should do." (p. 13). Hence, Dhami and Ayton's (2001) conclusions about „fast and frugal“ bailing decisions are at best premature.

Conclusions

We may speculate that this problem may result from the missing specification of an error model and unknown consequences of decision errors on the determination of the cue hierarchy in the fitting method. Whatever the reason, the result clearly points to the necessity of clarifying the formal link between theory/model and observable data. Or, at least, there has to be an *empirical* validation of the assessment method. We demonstrated that the former is missing and that the latter was not convincing in Dhami and Ayton's work. Even plausible assumptions can lead us astray if we do not validate our classification method with either formal or empirical arguments.

I did not choose this study because its method is particularly flawed as compared to others but because of the serious impact it might have on the British justice system. While psychological researchers should in principle use validated methods whenever they do research, this is especially important if the conclusions drawn from their work possibly influence public policy!

REFERENCES

- Bröder, A. (2000). A methodological comment on behavioral decision research. *Psychologische Beiträge*, 42, 645-662.
- Cutting, J. E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology*, 44, 3-19.
- Dawes, R.M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81, 95-106.
- Dhami, M. & Ayton, P. (2001). Bailing and jailing the fast and frugal way. *Journal of Behavioral Decision Making*, 14, 141-168.
- Dhami, M. K. & Harries, C. (2001). Fast and frugal versus regression models of human judgement. *Thinking and Reasoning*, 7, 5-27.
- Einhorn, H.J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, 73, 221-230.
- Fishburn, P. 1974. Lexicographic order, utilities and decision rules: a survey. *Management Science*, 20, 1442-1471.
- Gigerenzer, G., Hoffrage, U. & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506-528.
- Gigerenzer, G., Todd, P. & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Guttman L. 1981. What is not what in theory construction. In: *Multidimensional data representations: When and why* (p.47-68), I. Borg (ed.). Ann Arbor, MI: Mathesis Press.
- Harte, J.M. & Koele, P. (2001). Modelling and describing human judgement processes: The multiattribute evaluation case. *Thinking and Reasoning*, 7, 29-49.
- Hoffrage, U., Hertwig, R. & Gigerenzer, G. (2000). Hindsight Bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26, 566-581.
- Hu, X. & Batchelder, W.H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, 59.
- Riefer, D.M. & Batchelder, W.H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, 95, 318-339.
- Rieskamp, J. & Hoffrage, U. (1999). When do people use simple heuristics and how can we tell? In *Simple heuristics that make us smart* (p. 141-167), G. Gigerenzer, P. Todd & the ABC Research Group. New York: Oxford University Press.
- Roberts, S. & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, 107, 358-367.
- Simon, H.A. (1957). *Models of man: Social and rational*. New York: Wiley.
- Simon, H.A. (1992). What is an "explanation" of behavior? *Psychological Science*, 3, 150-161.
- Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance*, 23, 86-112.
- Svenson, O. (1983). Decision rules and information processing in decision making. In: *Human decision making* (p.131-162), L. Sjöberg, T. Tyszka & J. Wise (eds). Bodafors, S: Doxa.
- Tversky A. 1969. Intransitivity of preferences. *Psychological Review*, 76, 31-48.
- Tversky A. 1972. Elimination by aspects: A theory of choice. *Psychological Review*, 79, 281-299.