Judgment and Decision Making, Vol. 6, No. 8, December 2011, pp. 782–799

Diagnostic Task Selection for Strategy Classification in Judgment and Decision Making: Theory, Validation, and Implementation in R

Marc Jekel[*] Susann Fiedler[†] Andreas Glöckner[†]

Abstract

One major statistical and methodological challenge in Judgment and Decision Making research is the reliable identification of individual decision strategies by selection of diagnostic tasks, that is, tasks for which predictions of the strategies differ sufficiently. The more strategies are considered, and the larger the number of dependent measures simultaneously taken into account in strategy classification (e.g., choices, decision time, confidence ratings; Glöckner, 2009), the more complex the selection of the most diagnostic tasks becomes. We suggest the Euclidian Diagnostic Task Selection (EDTS) method as a standardized solution for the problem. According to EDTS, experimental tasks are selected that maximize the average difference between strategy predictions for any multidimensional prediction space. In a comprehensive model recovery simulation, we evaluate and quantify the influence of diagnostic task selection on identification rates in strategy classification. Strategy classification with EDTS shows superior performance in comparison to less diagnostic task selection algorithms such as representative sampling. The advantage of EDTS is particularly large if only few dependent measures are considered. We also provide an easy-to-use function in the free software package R that allows generating predictions for the most commonly considered strategies for a specified set of tasks and evaluating the diagnosticity of those tasks via EDTS; thus, to apply EDTS, no prior programming knowledge is necessary.

Keywords: Comparative model fitting; strategy classification; diagnostic task selection.

1 Introduction

The identification of individuals' decision strategies has always challenged behavioral decision research. There are at least three traditional approaches. Structural modeling applies a regression based approach to identify the relation between the distal criterion variable, proximal cues, and peoples' judgments (e.g., Brehmer, 1994; Brunswik, 1955; Doherty & Kurz, 1996; see Karelaia & Hogarth, 2008, for a meta-analysis); process tracing methods, for example, record information search (e.g., Payne, Bettman, & Johnson, 1988) or use think aloud protocols (e.g., Montgomery & Svenson, 1989; Russo, Johnson, & Stephens, 1989) to trace the decision process (see Schulte-Mecklenbeck, Kuehberger, & Ranyard, 2011, for a review); whereas comparative model fitting approaches investigate the fit of data and predictions of different models to determine the model or decision strategy employed (e.g., Bröder, 2010; Bröder & Schiffer, 2003; see also Pitt & Myung, 2002).

Comparative model fitting in particular has gained popularity in recent Judgment and Decision Making (JDM) research. In this paper, we discuss the problem of diagnostic task selection when using this strategy classification method. We suggest the Euclidian Diagnostic Task Selection (EDTS) method as a standardized solution. We report results from a comprehensive model recovery simulation that investigates the effects of different task selection procedures, number of dependent measures and their interaction on the reliability of strategy classification in multiple-cue probabilistic inference tasks.

2 Task selection in strategy classification based on comparative model fitting

The principle of strategy classification based on comparative model fitting (referred to in the following as *strategy classification*) is comparing a vector of choice data D_a consisting of n choices for person a to a set of predictions P_a of a set of strategies S. The strategy that "explains" the data vector best is selected. Strategies in set S have to be sufficiently specified to allow the definition of a complete vector of predictions P_a. Vector P_a can consist of sub-vectors for predictions on different dependent measures. Some strategies have free parameters to capture

We thank Jonathan Baron and two anonymous reviewers for their suggestions and comments. This research was partially funded by the priority program of the German Science Foundation (DFG) SPP 1516 "New frameworks of rationality" project numbers GL 632/3-1 and BR 2130/8-1.

[*] Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113, Bonn, Germany. E-mail: jekel@coll.mpg.de.

[†] Max Planck Institute for Research on Collective Goods.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

individual differences. Aspects that have to be considered to achieve a reliable strategy classification are: a) that all relevant strategies are included in the strategy set (e.g., Bröder & Schiffer, 2003), b) that overfitting due to model flexibility is avoided (e.g., Bröder & Schiffer, 2003), c) that appropriate model selection criteria are used (e.g., Hilbig, 2010; Hilbig, Erdfelder, & Pohl, 2010; Pitt & Myung, 2002; Pitt, Myung, & Zhang, 2002), and d) that diagnostic tasks are selected that allow differentiating between strategies (e.g., Glöckner & Betsch, 2008a). In the current paper, we investigate the influence of a more or less diagnostic task selection in more detail.

We are particularly interested in the consequences of representative sampling as opposed to diagnostic task selection. Tasks are to a varying degree representative of the environment and/or they are more or less diagnostic with respect to strategy identification (Gigerenzer, 2006). Representative sampling means that experimental tasks are sampled based on the probability of them occurring in the environment to which results should be generalized to (Brunswik, 1955).[1] Representative sampling is important with respect to external validity for at least two reasons. First, if one wants to generalize findings on rationality or accuracy of people's predictive decisions from an experiment to the real world, it is essential to draw a representative and hence generalizable sample.[2] One could, for instance, not claim that the calibration of a person‘s confidence judgments is bad if this conclusion is based on a set of "trick questions" that in fact are more difficult than they seem and that rarely appear in the real world (Gigerenzer, Hoffrage, & Kleinbölting, 1991).[3] A second aspect concerns interactions between task selection and strategy use. If the selection of tasks disadvantages the use of certain strategies (i.e., in contrast to its application in the real world), people are less likely to employ it, which leads to a general underestimation of its frequency of application.[4]

On the contrary, in diagnostic sampling, tasks are selected that differentiate best between strategies, that is, for which the considered strategies make sufficiently different predictions. Diagnostic task selection has not been given sufficient attention in some previous work. For example, the priority heuristic as a non-compensatory model for risky choices (Brandstätter, Gigerenzer, & Hertwig, 2006) was introduced based on a comparative model test. In 89 percent of the choice tasks used in the study, the priority heuristic made the same prediction as one of the established models (i.e., cumulative prospect theory with parameters estimated by Erev, Roth, Slonim, & Barron, 2002). Subsequent analyses showed that the performance of the heuristic dramatically drops when more tasks are implemented, for which the heuristic and prospect theory make different predictions (Glöckner & Betsch, 2008a). More research showed that conclusions about the heuristic being a reasonable process model for the majority of people were premature (Ayal & Hochman, 2009; Fiedler, 2010; Glöckner & Herbold, 2011; Hilbig, 2008; Johnson, Schulte-Mecklenbeck, & Willemsen, 2008). To circumvent such problems in future, diagnostic task selection should be given more attention. However, diagnostic task selection becomes a complex problem if multiple strategies and multiple dependent measures are considered simultaneously as described in the next section. Afterwards we suggest and evaluate a standardized method that allows selecting a set of very diagnostic tasks from all possible tasks based on a simple Euclidian distance calculation in a multi-dimensional prediction space.

3 Strategy classification based on multiple measures

Strategy classification methods were commonly based on choices only. However, strategies are often capable of perfectly mimicking each others' choices. Noncompensatory heuristics, for example, are submodels of the weighted additive strategy with specific restrictions of cue weights. This problem is even more apparent when, in addition, strategies are considered that do not assume deliberate stepwise calculations (Payne, et al., 1988). Recent findings on automatic processes in decision making (Glöckner & Betsch, 2008c; Glöckner & Herbold, 2011) suggest also taking into account cognitive models assuming partially automatic-intuitive processes (Glöckner & Witteman, 2010). Important classes of models are evidence accumulation models (Busemeyer & Johnson, 2004; Busemeyer & Townsend, 1993; Roe,

[1] Dhami, Hertwig, and Hoffrage (2004) equate representative sampling with "*probability sampling*, in which each stimulus has an equal probability of being selected" (p. 962, emphasis original; see also Hoffrage & Hertwig, 2006). Although it can be questioned whether equal probability sampling is a sound implementation of Brunswik's representative sampling at all (i.e., the probability of a task to appear in a study should match the probability of the task to appear in the real world), we use equal probability sampling for matters of convenience and lack of knowledge of the "true" sampling probabilities for the tasks used in the simulation reported below.

[2] See the correspondence criterion of rationality (Todd & Gigerenzer, 2000).

[3] Note that the frequency of an event in an environment is not per se an index of its significance. That is, rare events that lead to irrational behavior can be highly significant due to their consequences (e.g., severe punishment for not solving "trick questions") or due to selective oversampling of these—then no longer—"rare" events (e.g., oversampling of "trick questions" to take advantage of irrational behavior).

[4] The same is of course true for methodological approaches that hinder the application of certain strategies. It has, for instance, been shown that in some situations the classic mouselab paradigm hinders the application of weighted compensatory strategies (Glöckner & Betsch, 2008c).

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

Busemeyer, & Townsend, 2001), multi-trace memory models (Dougherty, Gettys, & Ogden, 1999; Thomas, Dougherty, Sprenger, & Harbison, 2008), and parallel constraint satisfaction (PCS) models (Betsch & Glöckner, 2010; Glöckner & Betsch, 2008b; Holyoak & Simon, 1999; Simon, Krawczyk, Bleicher, & Holyoak, 2008; Thagard & Millgram, 1995). As an example, we include a PCS strategy in our simulation.

Based on the idea that multiple measures can improve differentiation, the multiple-measure maximum-likelihood (MM-ML) strategy classification method (Glöckner, 2009, 2010; Jekel, Nicklisch, & Glöckner, 2010) was developed. MM-ML simultaneously takes into account predictions concerning choices, decision time, and confidence. MM-ML defines probability distributions for the data-generating process of multiple dependent measures (e.g., choices, decision times and confidence) and determines the (maximum) likelihood for the data vector D_a given the application of each strategy in the set S and multiple further assumptions (for details, see Appendix A).

It was shown that the MM-ML method leads to more reliable strategy classification than the choice based method (Glöckner 2009).[5] It has, for instance, been successfully applied to detect strategies in probabilistic inference tasks (Glöckner, 2010) and tasks involving recognition information (Glöckner & Bröder, 2011).

4 Simulation

We used a model recovery simulation approach to investigate the effects of task diagnosticity, numbers of dependent measures, and the interaction of the two on the reliability of strategy classification. We thereby simulated data vectors for hypothetical strategy users with varying noise rates and tried to recover their strategies employing the MM-ML method. In accordance with Glöckner (2009), we simulated probabilistic inferences for six different cue patterns (i.e., a specific constellation of cue predictions in the comparison of two options; see Figure 1, right), which are repeated ten times each resulting in a total of 60 tasks per simulated person.[6] The choice of the cue patterns was manipulated to test our predictions with respect to representative sampling and diagnostic task selection based on a standardized method. In practice, the selection of the most diagnostic cue patterns for a set of strategies is not trivial and to the best of our knowledge no standard procedures are available. We suggest a method to determine the cue patterns that differentiate best between any given set of strategies and test whether the method increases reliability in strategy classification.

4.1 Design

We generated data based on five strategies in probabilistic inference tasks with two options and four binary cues. We varied the validity of the cues in the environment, the degree of noise in the data generating process, the number of dependent measures included in the model classification, and the diagnosticity of cue patterns that were used. As dependent variables, we calculated the proportion of correct classifications—the identification rate—and the posterior probability of the data-generating strategy.[7] Ties and misclassifications were counted as failed identification. This results in a 5 (data generating strategy) × 3 (environment) × 4 (error rates for choices) × 3 (noise level for decision times and confidence judgments) × 3 (number of dependent measures) × 4 (diagnosticity of tasks) design. For each condition, we simulated 252 participants, resulting in 544,320 data points in total.

4.1.1 Data-generating strategies

For simplicity, we rely on the same data-generating strategies used in previous simulations (Glöckner, 2009) namely: parallel constraint satisfaction (PCS), take-the-best (TTB), equal weight (EQW), weighted additive ($\text{WADD}_{\text{corr}}$), and random (RAND) strategy, which are described in Table 1.

4.1.2 Environments

We used three environments: a typical non-compensatory environment with one cue clearly dominating the others (cue validities = [.90 .63 .60 .57]),[8] a compensatory environment with high cue dispersion (cue validities = [.80 .70 .60 .55]), and a compensatory environment with low cue dispersion (cue validities = [.80 .77 .74 .71]).

4.1.3 Error rates for choices and noise level for confidence and time

For each simulated participant, a data vector D_a was generated, based on the prediction of the respective data-

[5] MM-ML has been implemented as an easy-to-use function in the open-source statistical package R (Jekel et al., 2010) and in a function for STATA (Glöckner, 2009). The most recent implementations are provided on request by the authors of this paper.

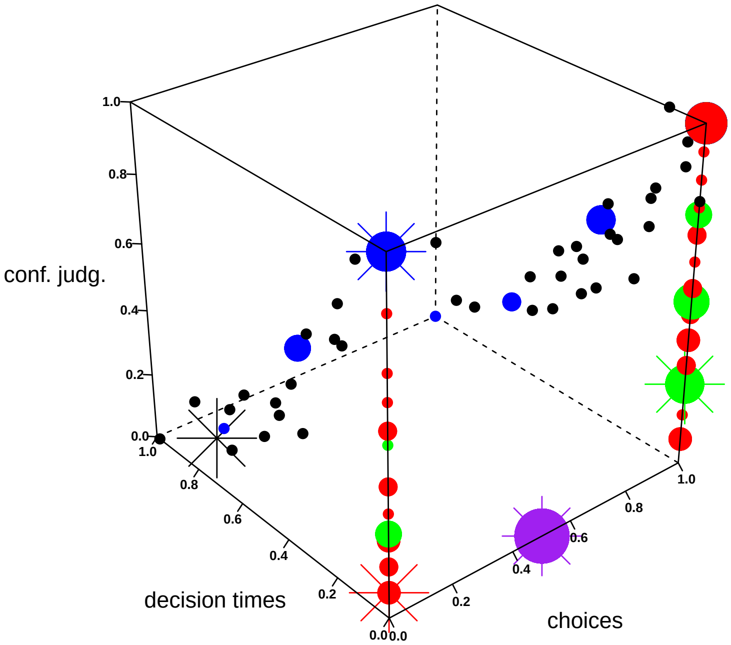
[6] Completing 60 tasks takes about 5–15 minutes, which allows mixing them with sufficient distractors to avoid interactions of task selection and strategy use mentioned in section 2.

[7] The posterior probability of the data-generating strategy can be calculated from the BIC values as described in Equation 3 in Appendix A (Wagenmakers, 2007).

[8] The environment is non-compensatory because the most valid cue can never be overruled by less valid cues if compensatory strategies such as $\text{WADD}_{\text{corr}}$ that takes (chance corrected) validities into account or PCS are applied.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

Figure 1: Prediction for 40 qualified cue patterns generated from five strategies (black = PCS, blue = TTB, red = EQW, green = WADD$_{\text{corr}}$, purple = RAND) in the rescaled prediction space with the three dependent measures (i.e., choices, decision times, confidence judgments) as coordinate axes. The size of the dots is (logarithmically) related to the number of predictions (i.e., density) at the respective coordinates. The five stars represent the predictions of the strategies for the (exemplary) cue pattern shown in the right side of Figure 1.



	Option	
	A	B
Cue1 (val = .80)	–	+
Cue2 (val = .70)	+	–
Cue3 (val = .60)	+	+
Cue4 (val = .55)	+	–

Note. Positive and negative cue predictions are represented by $+$ and $-$, respectively.

Table 1: Description of the strategies used in the simulation.

Strategies	Description
PCS	People construct a mental representation of the task (modeled as connectionist network; see Glöckner & Betsch, 2008b) based on the constellation of available and salient cues and their subjective estimation of validities. Spreading activation mechanisms accentuate initial advantages of one option over the other and make this option appear more attractive in that cues favoring it are highlighted and cues speaking against it are devalued. The most attractive option (i.e., option with the highest activation) is chosen.
TTB	Cues are searched sequentially by means of validity. The search is stopped when a cue discriminates between options. The option that is indicated by the discriminating cue is chosen.
EQW	The cue values for both options are added up. The option with the higher sum is chosen.
WADD$_{\text{corr}}$	The weighted sum of cue values is computed with cue validities for both options being corrected for chance level (i.e., validities – .5). The option with the highest weighted sum is chosen.
RAND	One option is chosen at random.

Note. We used PCS with fixed parameters and a quadratic cue transformation function: decay = .10; $w_{o1-o2} = -.20$; $w_{c-o} = .01/-.01$ [positive vs. negative prediction]; $w_v = ((v - .50) \times 2)^2$, stability criterion = 10^{-6}; floor = -1; ceiling = 1 (see Glöckner, 2010, for details).

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

generating strategy plus noise. The vector consisted of a sub-vector for choices, decision times, and confidence. For the choice vector, (exact) error rates were manipulated from 10% to 25% at 5%-intervals. For example, an error rate of 10% leads to 6 out of 60 choices that are inconsistent with the predictions of the strategy. It was randomly determined which six choices were flipped to the alternative choice for each simulated participant.[9]

Normally distributed noise was added to the predictions of the strategies for the decision time and confidence vectors (normalized to a mean of 0 and a range of 1). The three levels of noise on both vectors differed with respect to the standard deviation of the noise distribution σ_{error} = [1.33 1 0.75], which is equivalent to a manipulation of the effect size of d = [0.75 1 1.33]. Note that adding normally distributed noise $N(\mu = 0, \sigma_{error})$ to a normalized prediction vector leads to a maximum (population) effect size of $d = \frac{\mu_{max} - \mu_{min}}{\sigma_{pooled}} = \frac{1}{\sigma_{pooled}}$. Note also that the term $\mu_{max} - \mu_{min}$ is the difference between the means of the most distant populations from which realizations of the dependent measures are sampled and which reduces to 1 due to normalizing prediction vectors. The pooled standard deviation of those populations is equal to the standard deviation of the noise distribution (i.e., $\sigma_{pooled} = \sigma_{error}$) because random noise is the only source of variance within each population. Thus, a standard deviation of (e.g.) $\sigma_{error} = \sigma_{pooled} = 1.33$ leads to a maximum effect size of $d = \frac{1}{1.33} \approx 0.75$ between the most distant populations of the dependent measures.

4.1.4 Number of dependent measures

The strategy classification using MM-ML was based on varying numbers of dependent measures including (a) choices only, (b) choices and decision times, or (c) choices, decision times and confidence judgments.

4.1.5 Diagnosticity in Cue Patterns

We manipulated the diagnosticity of cue patterns used in strategy classification by using a) the Euclidian Diagnostic Task Selection (EDTS) method that determines the most diagnostic tasks given a set of strategies and the number of dependent measures considered, b) two variants of this method that generate medium and low diagnostic tasks, and c) representative (equal probability) sampling of tasks.

Probabilistic inference tasks with two options and four binary cues (i.e., [+ –]) allow for 240 distinct cue patterns. To prepare task selection, the set was reduced to a qualified set of 40 cue patterns by excluding all option-reversed versions ($n = 120$) and versions that were equivalent except for the sign of non-discriminating cues (i.e., [– –] vs. [+ +]). Then, strategy predictions for each of the three dependent measures were generated and rescaled to the range of 0 to 1 (for details, see Appendix B). The rescaled prediction weights for each strategy and each qualified task are plotted in the three-dimensional space that is spanned by the three dependent measures (Figure 1).

EDTS (Table 2) is based on the idea of cue patterns being diagnostic if predictions for strategies differ as much as possible. The pairwise diagnosticity is thereby measured as Euclidian distances between the predictions of two strategies for each cue pattern in the three-dimensional prediction space (Figure 1). The main criterion for cue pattern selection is the average diagnosticity of a cue pattern which is the mean of its Euclidian distances across all possible pairwise strategy comparisons in the space (i.e., PCS vs. TTB, PCS vs. EQW, ...). For statistical details, see Appendix C, and for a discussion of EDTS-related questions, see Appendix E.

For the high diagnosticity condition, we selected six cue patterns according to the EDTS procedure. For the medium and low diagnosticity condition, we selected cue patterns from the middle and lower part of the by diagnosticity sorted list of cue patterns generated in step 4 of EDTS. Cue patterns were sampled uniformly at random for the representative sampling condition.[10]

4.1.6 EDTS function in R

We have implemented EDTS as an easy-to-use function in the free software package R (2011). You can specify your own environment (i.e., number of cues and validities of cues), generate the set of unique pairwise comparisons between cue patterns for your environment (as described in 4.1.5), derive predictions for all strategies on choices, decision times, and confidence judgments for those tasks (as described in 4.1.1), and apply EDTS to calculate the diagnosticity of each task (as described in 4.1.5); see Appendix D and F for a detailed description of the EDTS function.

By applying the EDTS function, you can find the most diagnostic tasks from a specified environment, set of

[9] Note that some strategies predict guessing for some (or all) types of tasks (e.g., RAND). The choices of a simulated participant applying (e.g.) RAND were determined probabilistically in a first step—with choice A vs. B being equally likely. A choice was flipped to the alternative option if it was randomly selected for an error application in a second step.

[10] The ordering of cue patterns in EDTS dependends on the number of dependent measures taken into account. For the conditions with different numbers of dependent measures the Euclidian distances were calculated in the respective P-dimensional space (e.g., two-dimensional space if choices and decision times were included).

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

Table 2: Euclidian Diagnostic Task Selection (EDTS).

Steps	Description
1	Determine all predictions of all strategies and all distinct cue patterns and rescale them—except for choices—to a range from 0 to 1 (per strategy).
2	Determine pairwise Euclidian distances in the P-dimensional prediction space (P = number of dependent measures) between all strategy predictions for each cue pattern.[11] Rescale distances for each pairwise comparison to a range from 0 to 1, resulting in *diagnosticity* scores for each cue pattern.
3	Calculate the *average diagnosticity* of each cue pattern as its mean Euclidian distance across all pairwise strategy comparisons.
4	Sort cue patterns by their average diagnosticity and select n cue patterns from the top of this list (in our example n = 6).
5	Double-Check: Identify the maximum of diagnosticity scores for each pairwise comparison of strategies across n cue patterns. Compare the maximum to a threshold. If one maximum is below the aspired threshold, replace the last cue pattern(s) with one of the following cue patterns until the threshold is reached for all strategy comparisons. If no such cue pattern is found, repeat the procedure with a lower threshold.

strategies, and set of measures for *future* studies. You can also (systematically) alternate the number and validities of cues to find the environment that produces tasks that optimally distinguish between a set of strategies. Finally, you can also use the EDTS function to evaluate the diagnosticity of tasks, thus the reliability of strategy comparisons, and thus the reliability of conclusions from *past* studies.

4.2 Hypotheses

Based on previous simulations (Glöckner, 2009), we predict that additional dependent measures for MM-ML lead to higher identification rates and posterior probabilities for the data-generating strategy. We further expect that less diagnostic cue patterns lead to lower identification rates and posterior probabilities. We also hypothesize an interaction effect between diagnosticity and the number of dependent measures, that is, less diagnostic cue patterns benefit more from adding further dependent measures. For practical purposes, we are particulary interested in the size of the effect of each manipulation to assess the extent to which common practices influence results.

5 Results

5.1 Identification Rate

The overall identification rates for each type of task selection averaged across all environments and all strategies based on choices only are displayed in Figure 2 (left). As expected, cue patterns with high diagnosticity selected according to EDTS lead to the highest identification followed by representative sampling; cue patterns with medium and low diagnosticity were consistently even worse in identification. (see Figure 2, middle) or close (left and right) in identification. All types of task selection benefit from adding a second (see Figure 2, middle) and a third (see Figure 2, right) dependent measure. Representative sampling and the conditions with low and medium diagnosticity benefit most from adding a third dependent measure.[12]

Hence, results are descriptively in line with our hypotheses. For a statistical test of the hypotheses, we conducted a logistic regression predicting identification (1 = identified, 0 = not identified) by number of dependent measures, diagnosticity of tasks, environment, generating strategy, epsilon rate for choices, effect size for decision times, and confidence judgments (Table 3, first model).[13]

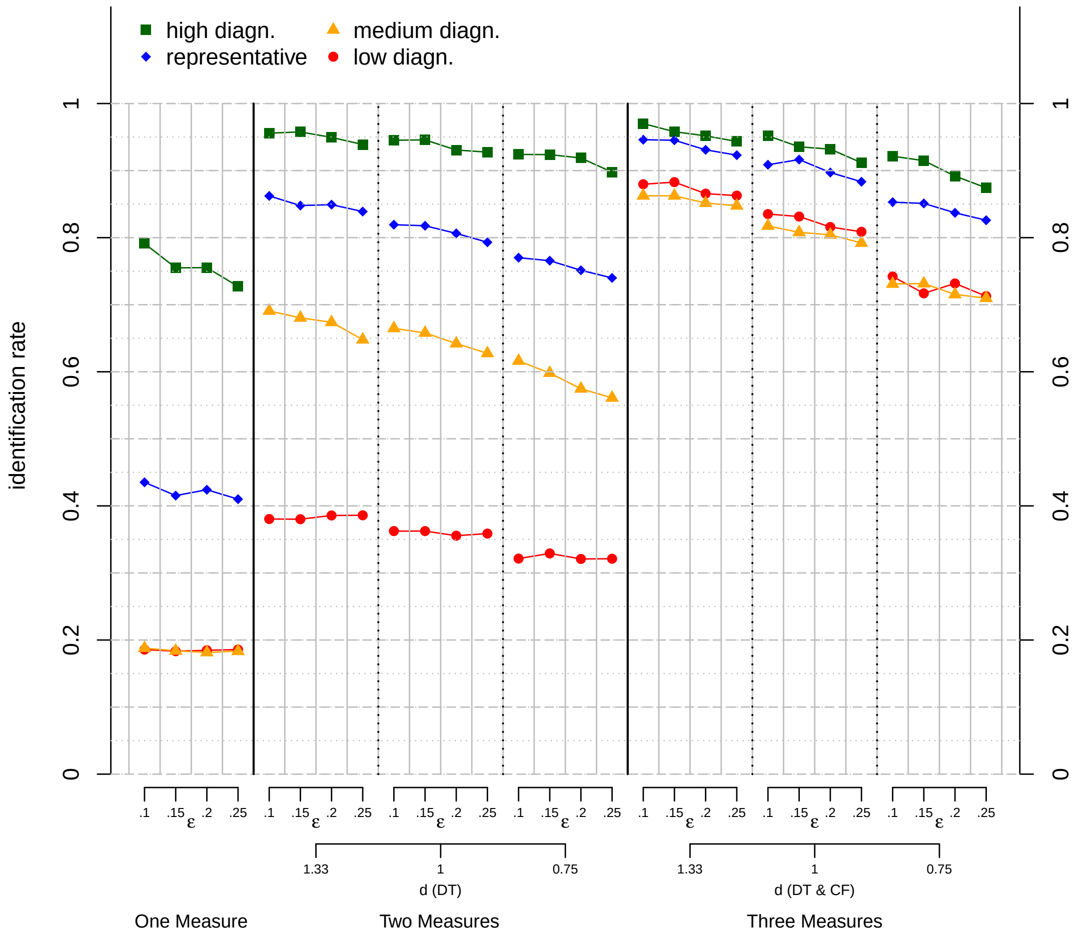
[11] Note that the selection mechanism is not limited to three dependent measures; the dimensionality of the space can be expanded (reduced) by adding (subtracting) dependent measures. Each pair of strategies must make different predictions on at least one dependent measure to disentangle strategies. Note also that it is possible to weight each dimension differently in order to scale the impact of each measure on the diagnosticity score (see Appendix E and w_{d_p} in formula 6, Step 2, Appendix C for details).

[12] We checked that the pattern of average identification rates displayed in Figure 2 is *not* driven by a single strategy (e.g., RAND) or only some of the strategies.

[13] The interpretation of p-values in the model is not warranted: the number of participants and therefore the test power can be arbitrarily varied. On the contrary, effect sizes can be interpreted in order to identify the relative importance (i.e., incremental explained variance) of each variable when all other variables are controlled for. Addition-

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

Figure 2: Identification rates for each type of task selection averaged across strategies and environments based on a) choices (left), b) choices and decision time (middle), and c) choices, decision time, and confidence (right).



The term ϵ refers to the error rate for choices. The middle and right graphs are separated by the effect size for decision time (DT) resp. decision time and confidence (DT & CF), d indicates the (maximum) effect size.

Results of the logistic regression indicate changes in the ratio of the odds for a successful strategy identification. For example, the odds ratio for the first dummy variable indicating that two dependent measures were used (i.e., choices and decision times), as compared to choices only (i.e., control group), is 7.39. This implies that the odds for identification increase by the factor 7.39 from using choices alone to using choices and decision times.[14] Adding decision time and confidence increases the odds ratio for identification by a factor of 20.91 (compared to choices only).

The odds for identification decrease by the factor of 0.29 (i.e., reduction to less than one third; see Footnote 14) when using representative sampling instead of high diagnostic sampling according to EDTS. The reduction from high to medium and low diagnostic sampling is even more pronounced.

Finally, less diagnostic pattern selection mechanisms benefit more from adding further dependent variables, as indicated by the odds ratios for the interaction terms between number of dependent measures and task diagnosticity. In particular, when all three dependent measures are considered, identification dramatically increases for representative sampling as well as medium and low diagnostic tasks, so that the disadvantage of representative sampling decreases to 3% (Table 4).

Hence, in line with our hypothesis, we replicate the finding that identification increases with number of de-

ally, the sign of the predictors can be interpreted in order to identify the relation between the proposed factors and the dependent variables (i.e., identification and posterior probability).

[14] If the odds for correct identification with choices only was .761/.239 = 3.18 (see Table 4), the odds for correct identification with choices and decision time would be $3.18 \times 7.39 = 23.50$, which translates into an odds ratio of .959/.041. Odds ratios below 1 indicate a reduction of the odds. The magnitude of the effects can be compared by calculating the inverse value for odds ratios below 1 (i.e., 1/odds ratio).

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

Table 3: Logistic regression predicting successful identification in strategy classification (Model 1) and linear regression predicting posterior probability of the data generating strategy (Model 2).

	Model 1	Model 2	
	Identification	Posterior probability	
Independent variables	Odds ratios	B	Incr. R^2
Intercept		.718	
# of Dependent Measures (control = 1 [choices only])			.193
2 [Choices and decision times] (1 = yes)	7.39	.199	
3 [Choices, decision times, and confidence judgments] (1 = yes)	20.91	.323	
Task Diagnosticity (control = high diagnosticity)			.169
Representative sampling (1=yes)	0.29	–.123	
Medium diagnosticity (1=yes)	0.08	–.260	
Low diagnosticity (1=yes)	0.05	–.320	
Environment (control = comp. & high dispersion)			.004
Noncompensatory (1=yes)	0.53	–.051	
Compensatory & low dispers. (1=yes)	0.69	–.031	
Generating Strategy (control = PCS)			.075
TTB (1= yes)	1.12	.026	
EQW (1= yes)	0.95	–.019	
$\text{WADD}_{\text{corr}}$ (1= yes)	0.30	–.139	
RAND (1= yes)	18.85	–.185	
Error in Choices (ϵ) (control = low [.10])			.003
Medium (.15)	0.94	–.010	
High (.20)	0.89	–.025	
Highest (.25)	0.82	–.043	
Maximum Effect Size for Decision Times and Confidence (control = highest [1.33])			.007
Medium (1)	0.83	–.028	
Lower (.75)	0.62	–.063	
Interaction: # of Dependent Measures × **Task Diagnosticity**			.020
Two dependent measures × Representative Sampling	1.68	.034	
Two dependent measures × Medium Diagn.	3.43	.151	
Two dependent measures × Low Diagn.	0.74	-.027	
Three dependent measures × Representative Sampling	3.95	.119	
Three dependent measures × Medium Diagn.	9.42	.190	
Three dependent measures × Low Diagn.	10.25	.188	

Note. Variables are dummy-coded and compared against the control condition. Variables for which interactions are calculated are centered. Nagelkerke's R^2 = .547 for identification rates; Adj. R^2 = .474 for posterior probabilities (N = 544,320, $p < .001$). $p < .001$ for all predictors and model comparisons (full vs. reduced models).

Table 4: Identification rates for the number of dependent measures and task selection vs. representative sampling.

Number of dependent measures	High diagnosticity	Representative sampling	Medium diagnosticity	Low diagnosticity
One	76.1%	41.6%	18.4%	18.4%
Two	93.4%	80.5%	63.6%	35.5%
Three	92.9%	89.3%	79.4%	80.7%

Note. Averaged over strategies, ϵ rates for choices, effect sizes for decision times and confidence judgments, and environments.

pendent measures. High-diagnosticity task-sampling according to EDTS leads to superior identification rates. The disadvantage of representative sampling decreases when more dependent measures are included.

5.2 Posterior probabilities for the data generating strategy

To analyze the effects of our manipulations further, we regressed posterior probabilities on the same factors described above (Table 3, Model 2). As expected, given that identification and posterior probabilities are both calculated from Bayesian Information Criterion (see Appendix A, Equation 2) values, the hypothesized effects of the manipulations are replicated. The independent variables of the linear model explain 47.4% of the variance in posterior probabilities. The number of dependent measures and task diagnosticity explain most of the unique variance[15] in posterior probabilities (19.3% and 16.9%). In comparison to classification based on choices only, two and three dependent measures lead to an increase of .199 and .323 in posterior probabilities. In comparison to high diagnostic cue patterns selected according to EDTS, posterior probabilities are reduced by –.260 and –.320 for cue patterns with medium and low diagnosticity, and by –.123 for representative sampling. Thus, cue pattern selection according to EDTS leads to considerably higher posterior probabilities of the data generating strategies than representative sampling.

6 Discussion and conclusion

Individual level strategy classification in judgment and decision-making is a statistical and a methodological challenge. There was a lack of standard solutions to the complex problem of diagnostic task selection in multi-dimensional prediction spaces. In the current paper, we suggest Euclidian diagnostic task selection (EDTS) as a simple method to select highly diagnostic tasks and show that EDTS increases identification dramatically. Furthermore, we replicate the increase in identification rates by employing multiple dependent measures in multiple-measure maximum likelihood (MM-ML) strategy classification method (Glöckner, 2009, 2010). We find that, under the conditions considered in our simulation, representative task-sampling reduces the odds for successful strategy classification by more than factor 1/3 compared to EDTS. This disadvantage, however, reduces if multiple dependent measures are used. Hence, if representative sampling is advisable for other methodological reasons (see section 2), multiple measures should be used. Unfortunately, this is not possible for all models because many models predict choices only (i.e., paramorphic models of decision making).

Our findings highlight that the issue of diagnosticity in task selection in comparative model fitting should be taken very seriously. To avoid ad-hoc criteria, we suggest using the EDTS method introduced in this article. Furthermore it would be advisable to report average diagnosticity scores for each selected cue pattern to be able to evaluate results better.

Robin Horton (1967a, 1967b, 1993)[16], who investigated the differences between religious and scientific thinking within the framework of Popper's critical rationalism, stated (1967b, p. 172) that "[f]or the essence of experiment is that the holder of a pet theory does not just wait for events to come along and show whether or not it has a good predictive performance."—an approach that might be equated with representative sampling—"He bombards it with artificially produced events in such a way that its merits or defects will show up as immediately and as clearly as possible." We hope that EDTS may help to find those events in a more systematic fashion in future research.

[15] Unique variance is determined by the reduction in variance from the full linear model to the model reduced by the respective factor(s).

[16] We thank Jonathan Baron for making us aware of this work.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

References

Ayal, S., & Hochman, G. (2009). Ignorance or integration: The cognitive processes underlying choice behavior. *Journal of Behavioral Decision Making, 22*, 455–474.

Betsch, T., & Glöckner, A. (2010). Intuition in judgment and decision making: Extensive thinking without effort. *Psychological Inquiry, 21*, 279–294.

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

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137–154.

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

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

Brunswik, E. (1955). Representative design and the probability theory in a functional psychology. *Psychological Review, 62*, 193–217.

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 133–154). Malden, MA: Blackwell Publishing.

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.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin, 130*, 959–988.

Doherty, M. E., & Kurz, E. M. (1996). Social judgment theory. *Thinking & Reasoning, 2*, 109–140.

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

Erev, I., Roth, A. E., Slonim, R. L., & Barron, G. (2002). Combining a theoretical prediction with experimental evidence to yield a new prediction: An experimental design with a random sample of tasks. Unpublished manuscript. Columbia University and Faculty of Industrial Engineering and Management, Techion, Haifa, Israel.

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

Gigerenzer, G. (2006). What's in a sample? A manual for building cognitive theories. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 239–260). Cambridge: Cambridge University Press.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506–528.

Gigerenzer, G., & Todd, P. M. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences, 23*, 727–780.

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. (2010). Multiple measure strategy classification: Outcomes, decision times and confidence ratings. In A. Glöckner & C. L. M. Witteman (Eds.), *Foundations for tracing intuition: Challenges and methods* (pp. 83–105). London: Psychology Press.

Glöckner, A., & Betsch, T. (2008a). Do people make decisions under risk based on ignorance? An empirical test of the Priority Heuristic against Cumulative Prospect Theory. *Organizational Behavior and Human Decision Processes, 107*, 75–95.

Glöckner, A., & Betsch, T. (2008b). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making, 3*, 215–228.

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

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

Glöckner, A., & Bröder, A. (2011). Processing of recognition information and additional cues: A model-based analysis of choice, confidence, and response time. *Judgment and Decision Making, 6*, 23–42.

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.

Glöckner, A., & Hodges, S. D. (2011). Parallel constraint satisfaction in memory-based decisions. *Experimental Psychology, 58*, 180–195.

Glöckner, A., & Witteman, C. L. M. (2010). Beyond dual-process models: A categorization of processes underlying intuitive judgment and decision making. *Thinking & Reasoning, 16*, 1–25.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

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

Hilbig, B. E. (2010). 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, and Cognition, 36*, 123–134.

Hochman, G., Ayal, S., & Glöckner, A. (2010). Physiological arousal in processing recognition information: Ignoring or integrating cognitive cues? *Judgment and Decision Making, 5(4)*, 285–299.

Hoffrage, U., & Hertwig, R. (2006). Which world should be represented in representative design. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 381–408). Cambridge: University Press.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*, 3–31.

Horton, R. (1967a). African traditional thought and Western science: Part 1. From tradition to science. *Africa, 37*, 155–187.

Horton, R. (1967b). African traditional thought and Western science: Part 2. The ‘closed’ and ‘open’ predicaments. *Africa, 37*, 155–187.

Horton, R. (1993). *Patterns of thought in Africa and the West: Essays on magic, religion and science.* Cambridge: Cambridge University Press.

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.

Johnson, E. J., Schulte-Mecklenbeck, M., & Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115*, 263–272.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134*, 404–426.

Krause, E. (1987). *Taxicab geometry: An adventure in non-Euclidean geometry.* Mineola, N. Y.: Dover Publications.

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.

Montgomery, H., & Svenson, O. (1989). A think-aloud study of dominance structuring in decision processes. In H. Montgomery & O. Svenson (Eds.), *Process and structure in human decision making* (pp. 135–150). Oxford, England: John Wiley & Sons.

Paradis, E. (2005). R for Beginners. Available at: http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf.

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.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421–425.

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.

R Development Core Team. (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 141–167). New York, NY: Oxford University Press.

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

Roe, R., Busemeyer, J. R., & Townsend, J. (2001). Multiattribute decision field theory: A dynamic, connectionist model of decision making. *Psychological Review, 108*, 370–392.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*, 759–769.

Schulte-Mecklenbeck, M., Kuehberger, A., & Ranyard, R. (2011). *A handbook of process tracing methods for decision research: A critical review and user’s guide.* New York: Psychology Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Simon, D., Krawczyk, D. C., Bleicher, A., & Holyoak, K. J. (2008). The transience of constructed preferences. *Journal of Behavioral Decision Making, 21*, 1–14.

Thagard, P., & Millgram, E. (1995). Inference to the best plan: A coherence theory of decision. In A. Ram & D. B. Leake (Eds.), *Goal-driven learning* (pp. 439–454). Cambridge, MA: MIT Press.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*, 155–185.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences, 23*, 727–780.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

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

Appendices

Appendix A: The Multiple-Measure Maximum Likelihood strategy classification method (MM-ML)

Appendix A describes the basic math of the MM-ML method; see Glöckner (2009, 2010) and Jekel, Nicklisch, and Glöckner (2010) for a more thorough description of the method, tools, and tutorials on how to apply MM-ML.

To apply MM-ML in probabilistic decision making, it is necessary to select a set of strategies, a set of dependent measures, and a set of cue patterns. For each dependent measure, assumptions have to be made concerning the probability function of the data-generating process. In our simulation study, we use choices, decision times, and confidence judgments as dependent measures and assume choices for six cue patterns which are repeated ten times each. The number of choices in line with a strategy prediction is assumed to be binomially distributed with a constant error rate for each cue patter; (log transformed and order corrected) decision times and confidence judgments are assumed to be drawn from normal distributions around rescaled prediction weights with constant standard deviation per measure.

Given a contrast weight t_{T_i} for the decision time and t_{C_i} for the confidence judgment of task i, further observing a data vector D consisting of a subvector for choices with n_{j_k} being the number of choices of type of tasks j congruent to strategy k and consisting of subvectors for decision time x_{T_i} and confidence judgment x_{C_i} for task i, it is possible to calculate the likelihood L_{total} for the observed data vector under the assumption of an application of strategy k (and the supplementary assumptions mentioned above) for a participant according to (Glöckner, 2009, Equation 8, p. 191):

$$L_{total} = p(n_{jk}, \vec{x}_T, \vec{x}_C | k, \epsilon_k, \mu_T, \sigma_T, R_T, \mu_C, \sigma_C, R_C) = \prod_{j=1}^{J} \binom{n_j}{n_{jk}} (1-\epsilon_k)^{n_{jk}} \epsilon_k^{(n_j - n_{jk})} \times \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma_T^2}} e^{-\frac{(x_{T_i} - (\mu_T + t_{T_i} R_T))^2}{2\sigma_T^2}} \times \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\sigma_C^2}} e^{-\frac{(x_{C_i} - (\mu_C + t_{C_i} R_C))^2}{2\sigma_C^2}}. \quad (1)$$

The error rate for choices, ϵ_k, the overall mean and standard deviation for decision times (μ_T, σ_T) and confidence judgments (μ_C, σ_C) as well as the rescaling factor R_T and R_C (R_T, $R_C \geq 0$) for decision times and confidence judgments that minimize the log-likelihood function are estimated.

The Bayesian Information Criterion (BIC, Schwarz, 1978) is calculated to account for different numbers of parameter (numbers vary because some strategies do not predict differences on all dependent measures or assume a fixed error rate of .50) according to:

$$\text{BIC} = -2\ln(L) + \ln(N_{obs})N_p. \quad (2)$$

N_{obs} represents the number of task types (i.e., six in the simulations) and N_p the number of parameters that need to be estimated for the likelihood. Thus, a strategy with more free variables is punished for its flexibility.

Finally, the posterior probability Pr for a specific strategy k, i.e., the probability of the strategy k as the data-generating mechanism under consideration of the observed data D and under the assumption of equal prior probabilities for all (i.e., K) considered strategies, can be calculated based on the BIC values according to (compare with Wagenmakers, 2007, Equation 11, p. 797):

$$\text{Pr}_{\text{BIC}}(s_k|D_a) = \frac{e^{[-\frac{1}{2}\times \text{BIC}_{s_k}]}}{\sum_{o=1}^{K} e^{[-\frac{1}{2}\times \text{BIC}_{s_o}]}}. \quad (3)$$

Appendix B: Strategy predictions

Predictions of strategies are derived by assuming that TTB, EQW, and WADD$_{\text{corr}}$ are applied in a stepwise manner according to the classic elementary information processes approach (e.g., Payne, et al., 1988). For PCS, predictions are derived from a standard network simulation (Glöckner & Betsch, 2008b; Glöckner, Betsch, & Schindler, 2010; Glöckner & Bröder, 2011; Glöckner & Hodges, 2011) using the parameters mentioned in the Note of Table 1. Table 5 shows the predictions for the cue patterns selected for the high diagnosticity condition in the environment with cue validities of .80, .70, .60, and .55 as an example.

Choices. Choice predictions are determined according to the mechanisms described in Table 1.

Decision times. For TTB, EQW, and WADD$_{\text{corr}}$, the number of computational steps necessary to apply the strategy is used as time prediction. For PCS, the number of iterations of the network necessary to find a stable solution is used as an indicator for decision time.[17]

[17] Note in Table 5 that (e.g.) EQW and WADD$_{\text{corr}}$ have the same set of contrast weights for decision time predictions (i.e., zeroes). This does not mean that the application of both strategies takes the same time (WADD$_{\text{corr}}$ should take longer due to the additional weighting of cues with validities). The application of both strategies is independent of the type of tasks (i.e., all cues are always investigated); thus, contrast predictions do not differ between types of tasks and are set to 0 (i.e., decision times are supposed to stem from a single distribution for

Table 5: Highly diagnostic cue patterns for three dependent measures in a compensatory environment (validities = [.80 .70 .60 .55]) and predictions for each strategy and dependent measure.

	Types of decision tasks					
	1	2	3	4	5	6
	A B	A B	A B	A B	A B	A B
Cue 1 (v = .80)	− +	− +	− +	+ +	− +	+ −
Cue 2 (v = .70)	+ −	+ −	+ −	+ +	+ +	+ −
Cue 3 (v = .60)	+ −	+ −	+ +	+ +	+ −	+ −
Cue 4 (v = .55)	+ −	+ +	+ −	+ −	+ −	+ +
	Choice Predictions					
PCS	A	B	B	A	B	A
TTB	B	B	B	A	B	A
EQW	A	A	A	A	A	A
WADD$_{corr}$	A	A:B	B	A	B	A
RAND	A:B	A:B	A:B	A:B	A:B	A:B
	Time Predictions (contrasts)					
PCS	0.340	0.326	0.026	0.240	−0.273	−0.659
TTB	−0.167	−0.167	−0.167	0.833	−0.167	−0.167
EQW	0	0	0	0	0	0
WADD$_{corr}$	0	0	0	0	0	0
RAND	0	0	0	0	0	0
	Confidence Predictions (contrasts)					
PCS	−0.050	−0.322	−0.159	−0.217	0.072	0.677
TTB	0.167	0.167	0.167	−0.833	0.167	0.167
EQW	0.250	−0.250	−0.250	−0.250	−0.250	0.750
WADD$_{corr}$	−0.167	−0.250	−0.167	−0.167	0	0.750
RAND	0	0	0	0	0	0

Note. Positive cue values are indicated by +, negative cue values by −. A:B represents guessing between options.

Confidence judgments. For TTB, the validity of the discriminating cues is used as a predictor to confidence (Gigerenzer, et al., 1991). For EQW and WADD$_{corr}$, the difference in the (un)weighted sum of cue values for each option is used instead. For PCS, the difference in activations of the options is used as a predictor for confidence judgments.

Appendix C: Euclidian Diagnostic Task Selection (EDTS)

Step 1: Generate standardized prediction vectors

Define a set of K strategies s to be tested, a set of P dependent measures d used for MM-ML, and a set of I qualified cue patterns c (i.e., excluding identical patterns). Calculate prediction vectors for each strategy (see Appendix B) and rescale them to a range of 0 to 1 per strategy. Note that dependent measures of probabilities

all tasks); see also Glöckner (2009, 2010) and Jekel, Nicklisch, and Glöckner (2010) for details.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

(e.g., choices) should *not* be rescaled.[18] The goal is to choose n cue patterns highest in diagnosticity from the set of the I cue patterns. Assume the following notation for raw (indicated by superscript R) contrast weights cw:

$$cw^R_{s_k d_p} = \left[cw^R_{s_k d_p c_1} \quad cw^R_{s_k d_p c_2} \quad \ldots \quad cw^R_{s_k d_p c_I} \right]. \quad (4)$$

Each contrast vector is calculated from the raw values to fit the range from 0 to 1. Contrast weights are rescaled by:

$$cw_{s_k d_p c_i} = \frac{cw^R_{s_k d_p c_i} - \min(cw^R_{s_k d_p \cdot})}{\max(cw^R_{s_k d_p \cdot}) - \min(cw^R_{s_k d_p \cdot})}. \quad (5)$$

Step 2: Calculate diagnosticity scores for strategy comparisons

Compute the diagnosticity scores for each task as the Euclidian distances ED for each strategy comparison and each cue pattern within the space spanned by the vectors of the P dependent measures that are weighted by w_{d_p}. Following this, standardize these distances to a range from 0 to 1. ED between strategy k and o ($k \neq o$) for cue pattern i are calculated by:

$$ED^R_{s_k s_o c_i} = \sqrt{\sum_{p=1}^{P} (w_{d_p} \times (cw_{s_k d_p c_i} - cw_{s_o d_p c_i}))^2}. \quad (6)$$

For each comparison of strategy k and o, rescale $ED^R_{s_k s_o}$ across all I cue patterns to fit the range from 0 to 1 by:

$$ED_{s_k s_o c_i} = \frac{ED^R_{s_k s_o c_i} - \min(ED^R_{s_k s_o \cdot})}{\max(ED^R_{s_k s_o \cdot}) - \min(ED^R_{s_k s_o \cdot})}. \quad (7)$$

[Rationale for rescaling: Euclidian distances for each strategy comparison should have the same range to avoid overweighting (resp. underweighting) of strategy comparisons with a high variance (resp. low variance) in Euclidian distances.]

Step 3: Calculate the average diagnosticity scores

Calculate the means for each row of the matrix containing the rescaled Euclidian distances ED_{c_i} to receive the average diagnosticity AD score for each cue pattern by:

$$AD_{c_i} = \frac{2 \times (K-2)!}{K!} \times \sum_{k=1}^{K-1} \sum_{o=k+1}^{K} ED_{s_k s_o c_i}. \quad (8)$$

Step 4: Sort cue patters by average diagnosticity scores and select cue patterns

The set of I cue patterns can be easily sorted by their AD score. The n cue patterns with the highest AD score would be selected.

Step 5: Refine selection

Investigate if the maximum of diagnosticity scores for each strategy comparison is above a threshold t_{min}. To find an appropriate set of cue patterns, the threshold should increase with the number of dependent measures used and decrease with the number of pairwise comparisons. In the simulations, we used a threshold value of $t_{min} = .75$. If a maximum is below the aspired threshold, replace the last cue pattern(s) by one of the following cue patterns until the threshold is reached for all comparisons. If no such cue pattern is found, repeat the procedure with a lower threshold.

[Rationale: A high mean of rescaled Euclidian distances for a cue pattern can be produced by a single high distance for one of the strategy comparisons. Apply step 5 to ensure that there is at least one diagnostic cue pattern for each strategy comparison in the subset (as defined by the threshold).]

Appendix D: Implementation of EDTS as a function in R

EDTS is implemented as an easy-to-use function in R. R (2011) is a software for statistical analysis under the GNU general public license, e.g., it is free of any charge. R is available for Windows, Mac, and UNIX systems. To download R, visit the Comprehensive R Archive Network (http://cran.r-project.org/). To learn more about R, we propose the free introduction to R by Paradis (2005); however, to apply EDTS in R, no sophisticated prior experience with the R syntax is required.

You can download the EDTS.zip folder[19] from http://journal.sjdm.org/vol6.8.html, listed with this article. In the folder *EDTS*, there are two files—*mainFunction.r* and *taskGenerator.r*—and an additional folder *strategies* containing six further R files. In the current version of the EDTS function, it is possible to generate all possible unique pattern comparisons for two-alternative decision tasks with binary cue values (i.e., 1 or –1), to derive predictions for all tasks and a set of default strategies, and to calculate the diagnosticity index for each task as proposed in the article.

To use the EDTS function, you need to copy and paste (or submit) the code provided in the file *mainFunction.r*,

[18] Otherwise, predicted guessing (i.e., $p(A) = .5$) is erroneously recoded as predicted choice for option A (i.e., $p(A) = 1$) if a strategy predicts choices for option B (i.e., $p(A) = 0$) and guessing only.

[19] Software to extract the folder is included in most operating systems or is available as open source software (e.g., 7-zip from http://www.7-zip.org/ for Windows systems).

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

i.e., you can open *mainFunction.r* in a standard text editor, copy the entire code, and paste the code in the open R console. To call the function afterwards, type the command:

> EDTS (*setWorkingDirectory, validities, measures, rescaleMeasures, weightingMeasures, strategies, generateTasks, derivePredictions, reduceSetOfTasks, printStatus, saveFiles, setOfTasks, distanceMetric, PCSdecay, PCSfloor, PCSceiling, PCSstability, PCSsubtrahendResc, PCSfactorResc, PCSexponentResc*)

in the open R console and hit Enter. If an argument of the function is left blank, the default is applied. Arguments, descriptions, valid values, examples and defaults are listed in Appendix F. In the following, we give an example for illustrative purposes.

Example

Assume you want to test which of (e.g.) the four strategies—PCS, TTB, EQW, or RAND—describes human decision making best in a six-cue environment with the cue validities v = [.90 .85 .78 .75 .70 .60] (compare with Rieskamp & Otto, 2006). Your goal is to select the most diagnostic tasks from all possible tasks for an optimal comparison of strategies. Assume further that you will assess choices and decision times as dependent variables in your study; thus, you only need to rescale decision times (see Appendix C). For all the remaining arguments, you want to keep the defaults of the function.

To apply EDTS, you put the unzipped EDTS folder under C:\, open the file *mainFunction.r* with a text editor, copy the entire text and paste it in the open R console. Following, you type in:

> EDTS(validities = c(.90, .85, .78, .75, .70, .60), measures = c(“choice”, “time”) , rescaleMeasures = c(0, 1), strategies = c(“PCS”, “TTB”, “EQW”, “RAND”))

and hit Enter. Three .csv files are created: (1) *tasks.csv* includes all qualified patterns for a pairwise comparisons with six cues (i.e., 364 tasks), (2) *predictions.csv* includes choice and decision time predictions for all strategies (i.e., PCS, TTB, EQW and RAND) and all tasks listed in *tasks.csv*, (3) *outputEDTS.csv* includes the average diagnosticity score (AD), the minimum, maximum, and median diagnosticity of all strategy comparisons. Additionally, “raw” diagnosticity scores for each task and each strategy are provided. Based on the AD scores, you finally select the most diagnostic tasks for the strategy comparisons in a six-cue environment (see Table 2, step 5).

Generalizations

We added two further strategies as default strategies: (1) WADD$_{\text{uncorr}}$ (Rieskamp & Hoffrage, 1999) has been extensively used in past studies and thus can serve as an interesting competitor. WADD$_{\text{uncorr}}$ is identical to WADD$_{\text{corr}}$ but does not correct validities for chance level (e.g., .5 for pairwise comparisons). (2) RAT (Lee & Cummins, 2004) is the rational choice model based on Bayesian inference. It has been included as a further strategy in order to allow comparisons between heuristic models and the rational solution in probabilistic decision making.

Additionally, it is also possible to extend the set of default strategies with your own strategies. To do so, you open the file *predictions.csv* and include a prediction column for each measure and for each task for your own strategies (as defined in *tasks.csv*). The labels of the new columns need to fit the form NameOfYourStrategy.Measure. Additionally, the order and number of columns (i.e., the order of predictions for each measure) need to follow the order of the measures of the other strategies included (i.e., choice, time, and confidence for the default measures).[20] To apply EDTS for your own specified set of strategies, you then include the names of your strategies in the argument *strategies* of the EDTS function and set the argument derivePredictions = 0 (i.e., predictions are not derived and the data matrix defined in *predictions.csv* with your set of strategies is loaded into the program instead).

It is also possible to add further dependent measures. Similar to adding strategies, you insert a further column for each strategy following the form Strategy.Measure for the labeling in the first row of the data matrix. For example, if you want to compare PCS and TTB on choices, decision times, and (e.g.) arousal (Hochman, Ayal & Glöckner, 2010), the file *predictions.csv* consists of 7 columns. In the first column, the number of the task is coded. From the second to the third column, PCS predictions are inserted with the labels PCS.choice, PCS.time, PCS.arousal in the first row of the data matrix. From the fifth to seventh column, TTB predictions are inserted with the labels TTB.choice, TTB.time, TTB.arousal. Thus, predictions for each measure are inserted by strategy and for each strategy the measures are in the same order.

In general, the EDTS function is thus applicable to any strategy for which quantitative predictions on each measure can be derived for a set of tasks. The function can also be applied to tasks differing from the default characteristics (e.g., probabilistic decision-making between three options and/or continuous cues) or from the default type (e.g., preference decisions between gambles)

[20] Note that labels need to be put in double quotation marks (i.e., “”) and values are separated by comma in all .csv files.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

by inserting the predictions for each strategy and measure in the file *predictions.csv* as described above. Thus, the method is not limited to the strategies and tasks used and implemented as defaults in the EDTS function. The experienced R user can thus implement her strategies as R code. To simplify coding, the main EDTS function and strategies are coded in separate files (see folder strategies), and strategies are also coded as functions that are similar in structure (same input variables, etc.).[21]

Appendix E: Open questions and future research

This short Appendix is supposed to make you aware of some open questions. For those researchers who are interested in applying EDTS, this section may sensitize you to critical aspects of EDTS. For those researchers who are interested in optimizing EDTS, the following open questions can be a hint for future studies; the EDTS function provided (see Appendix D and F) may further facilitate this process.[22]

There are alternative selection criteria (e.g., maximum or median) that may be used for task selection instead of the mean proposed and validated in the current study. For example, strategy comparisons may be more effective if the most discriminating task for each comparison (= maximum) is selected. However, there are two opposing forces at work: the number of tasks increases rapidly if the set of strategies increases (i.e., 5 strategies $= 10$ tasks, 6 strategies $= 15$ tasks, 7 strategies $= 21$ tasks, etc.). This can lead to less repetition of the selected tasks if the number of tasks that can be presented in a study is limited. Less repetition can then lead to a less reliable strategy classification dependent on the error rate. It is therefore an open question if the gain of diagnosticity for single comparisons outweighs the loss of reliability due to less repetition of the tasks. To facilitate comparison between several diagnosticity statistics, the output of the EDTS function includes several diagnosticity statistics (mean, median, maximum, and minimum) and the "raw" diagnosticity scores for each strategy comparison and each task.

There is no need to restrict EDTS to *Euclidian* distances as the metric for diagnosticity scores. It is an open question if other metrics lead to reliable strategy classification as well (or even better ones). We have implemented the option to calculate diagnosticity scores based on Taxicab/Cityblock metrics (Krause, 1987) in the EDTS function as well.

Finally, there may be reasons to weight the impact of each dependent measure on the diagnosticity score differently. For example, it may be reasonable to reduce the impact of dependent measures that are less reliable and thus favor more reliable measures in diagnostic task selection. It is an open question if different weighting schemes (e.g., weighting of each measure relative to a reliability index) lead to higher identification rates. We have implemented the option to weight measures differently in the EDTS function.

[21] We are happy to collect further strategies programmed by other users to extend the set of strategies implemented in the EDTS function; please send your files to the first author (jekel@coll.mpg.de). We plan to provide future extensions to the EDTS function as a download from a website that will be announced via the JDM-society mailing list.

[22] We thank our reviewers for making us aware of these issues.

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

Appendix F: EDTS() function in R

Arguments, descriptions, valid values, examples, and defaults.

Argument[23]	Description	Valid Values	Example	Default
setWorking-Directory	indicate the directory of the EDTS folder	platform specific requirements	"C:/Files/EDTS/" (be aware of "")	"C:/EDTS/"
validities	set the validities of the cues (note: the software extracts the number of the cues automatically from the number of validities specified)	any numeric value between .5 and 1	c(.9, .85, .80) (be aware of c())	c(.80, .70, .60, .55)
measures	set the measures used for EDTS	any character(s)	c("choice", "time") (be aware of "")	c("choice", "time", "confidence")
rescale-Measures	indicate which measure to be rescaled to a range from 0 to 1 for EDTS (see Appendix C, formula 5)	rescale measure? 1 = yes, 0 = no	c(0, 1) (be aware of c())	c(0, 1, 1)
weighting-Measures	weighting of measures before calculating Euclidian distances (see w_{d_p} in formula 6)	any numeric value	c(1, 1, 1) (be aware of c())	c(1, 1, 1)
strategies	set the strategies for EDTS	any character(s)	c("PCS", "TTB", "EQW") (be aware of "")	c("PCS", "TTB", "EQW", "WADDcorr", "RAND", "RAT", "WADDuncorr")
generateTasks	indicate if all possible unique tasks should be generated	generate all unique tasks? 1 = yes, 0 = no	1	1
reduceSet-OfTasks	remove all tasks that only differ in the sign for non-discriminating cues (i.e., – – vs. + +)	reduce the set of tasks? 1 = yes, 0 = no	1	1
derive-Predictions	inductate if predictions are to be derived for strategies and tasks	derive predictions? 1 = yes, 0 = no	1	1
printStatus	indicate if the current status of the EDTS function is to be printed	print the current status of the function? 1 = yes, 0 = no	0	1
saveFiles	indicate if the the data created by the EDTS function (i.e., tasks, predictions, EDTS output) is to be saved as .csv files	create files? 1 = yes, 0 = no	0	1

⋮

(table continued on next page)

https://doi.org/10.1017/S1930297500004216 Published online by Cambridge University Press

(table continued from last page)

⋮

setOfTasks	the set of tasks for which predictions and diagnosticity scores are to be derived can be inserted directly into the function within the R environment	first column (task number) $= 1, 2, 3, \ldots n$, second and third column (cue pattern) $= 1$ or -1	cbind(c$(1, 1, 1, 1, 2, 2, 2, 2)$, c$(1, -1, 1, -1, 1, 1, 1, 1)$, c$(1, 1, -1, 1, 1, -1, 1, 1)$) (be aware of cbind$()$ for binding vectors columnwise and c$()$ for creating vectors)	“none”
distance-Metric	define the metric for EDTS calculations; this is an experimental (!) option to simplify further validation studies on different distance metrics for EDTS	“Euclidian” or “Taxicab”	“Euclidian” (be aware of “”)	“Euclidian”
PCS specific options (see Glöckner & Betsch, 2008b; Glöckner & Bröder, 2011)				
PCSdecay	decay of node activation	any positive numeric value	.05	.1
PCSfloor	minimum node activation	any numeric value	-1	-1
PCSceiling	maximum node activation	any numeric value	1	1
PCSstability	specifiy (decimal point) sensitivity for a stable PCS solution; higher values result in more iterations (i.e., sensitivity is the inverse of the specified value)	any numeric value	10ˆ6	10ˆ6
PCSsub-trahendResc	rescaling of the validities v: w_v = ((v – PCSsubtrahendResc) × PCSfactorResc)$^{\text{PCSexponentResc}}$	any numeric value	0	.5
PCSfactor-Resc	see description PCSsubtrahendResc	any numeric value	1	2
PCSexponent-Resc	see description PCSsubtrahendResc	any numeric value	1.9	2

[23] Note that hyphens within arguments (i.e., -) are included only for reasons of limited space in the table.