Use of the recognition heuristic depends on the domain's recognition validity, not on the recognition validity of selected sets of objects

Rüdiger F. Pohl¹ · Martha Michalkiewicz^{1,3} · Edgar Erdfelder¹ · Benjamin E. Hilbig²

Published online: 10 February 2017
© Psychonomic Society, Inc. 2017

Abstract According to the recognition-heuristic theory, decision makers solve paired comparisons in which one object is recognized and the other not by recognition alone, inferring that recognized objects have higher criterion values than unrecognized ones. However, success—and thus usefulness—of this heuristic depends on the validity of recognition as a cue, and adaptive decision making, in turn, requires that decision makers are sensitive to it. To this end, decision makers could base their evaluation of the recognition validity either on the selected set of objects (the set's recognition validity), or on the underlying domain from which the objects were drawn (the domain's recognition validity). In two experiments, we manipulated the recognition validity both in the selected set of objects and between domains from which the sets were drawn. The results clearly show that use of the recognition heuristic depends on the domain's recognition validity, not on the set's recognition validity. In other words, participants treat all sets as roughly representative of the underlying domain and adjust their decision strategy adaptively (only) with respect to the more general environment rather than the specific items they are faced with.

Keywords Recognition validity · Recognition heuristic · Decision making

Introduction

The adaptive toolbox approach to judgment and decision making (Gigerenzer, Todd, & ABC Research Group, 1999) assumes that people possess a repertoire of decision strategies that are tailored to exploit helpful cue information in specific environments. According to that approach, decision makers have learned through their extended experience which tool fits best to which environment. In this way, decision making can adapt to a wide range of potential environments, so that in a specific decision situation, the decision maker only needs to detect the features of the current environment (i.e., structure of the task and valid cues) and then selects the most appropriate strategy. Behavior that utilizes these helpful environmental features has been termed “ecologically rational” (Todd & Gigerenzer, 2012).

We focus here on one of the simplest strategies from this toolbox, the recognition heuristic (RH) (Goldstein & Gigerenzer, 2002) which presumes that whenever one recognizes only one of two objects in a to-be-compared pair, one may infer that the recognized object has the larger criterion value. This strategy is so simple, because it relies on only one cue, namely recognition, which is very easily accessible (cf. Pachur & Hertwig, 2006). Besides, recognition seems to be a valid cue in many domains and thus rather helpful in otherwise rather difficult tasks.

As an example, imagine you are asked which of the following two Swiss cities is more populous, Geneva or Interlaken. And imagine further that you recognize the city name Geneva, but have never heard of Interlaken. In this case,

---

✉ Rüdiger F. Pohl
pohl@psychologie.uni-mannheim.de

¹ Department of Psychology, School of Social Sciences, University of Mannheim, D-68131 Mannheim, Germany

² Cognitive Psychology Laboratory, University of Koblenz-Landau, Koblenz-Landau, Germany

³ Institute of Experimental Psychology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

most people tend to infer that Geneva is the larger of the two (which happens to be correct). This behavior corresponds to what the RH theory predicts (Goldstein & Gigerenzer, 2002). The heuristic will yield accurate judgments so long as recognition is related to the to-be-inferred criterion, in other words, whenever the validity of recognition as a cue is larger than chance. Previous results revealed that decision makers are able to differentiate domains with different recognition validities and adjust their behavior accordingly (Gigerenzer & Goldstein, 2011). In other words, decision makers are able to capitalize on this specific feature and thus behave in an ecologically rational way (Todd & Gigerenzer, 2012).

These findings imply that decision makers have some grasp of how valid the recognition cue is. This, in turn, begs the question: What exactly determines subjective estimates of the recognition validity? One option is that individuals assess recognition validity from the set of objects they are making inferences on, that is, the set's recognition validity (we call this the “local-optimization” hypothesis). Another option is that they refer to the larger underlying domain of all objects from which the set was drawn, that is, the domain's recognition validity (we call this the “global-optimization” hypothesis). The first alternative appears highly effective in adapting to any specific task environment by considering only the actual items and cue validities within the set. Indeed, only through such a “local” focus can one hope to determine the accuracy offered by a strategy such as the RH in the specific task one is facing. Such an expectation is supported by findings from sampling studies showing that “judgments and decisions are often remarkably sensitive to the data given in a stimulus sample” (Fiedler, 2008, p. 186). A potential procedure of how to estimate the recognition validity in a given set of objects, even without receiving feedback after single choices, could proceed as follows. If people have at least some valid knowledge in the respective domain (as is arguably the case for the world cities used herein), they can rely on the degree to which recognition and knowledge are aligned: For any recognized object, additional knowledge might speak for or against that object being the correct choice. Thus, if recognition and knowledge repeatedly coincide in pairs with only one object recognized, one may eventually discard knowledge and rely on recognition alone (based on the insight that recognition validity is likely to be large and knowledge yields little to no additional information). In contrast, if recognition and knowledge repeatedly conflict with each other, one may hesitate to use recognition alone. That way, people might arrive at an indirect estimate of the recognition validity in the specific set of objects and adapt their behavior accordingly.

However, having to re-assess cue validities in each and every object sample from scratch would clearly be cumbersome and potentially impossible (Dougherty, Franco-Watkins, & Thomas, 2008). The second alternative is much more parsimonious in that it registers only the general domain from

which the objects were drawn, retrieves the domain's “global” cue validities, and then selects the appropriate strategies. In fact, most natural samples may be considered roughly representative or at least not severely biased so that global optimization should generally fare well. The caveat in this case is that one may select a strategy that does not actually yield the desired accuracy for the specific task at hand (or forgo use of a more accurate strategy), namely, if the set is not representative of the domain. Fiedler (2008) referred to this as “meta-cognitive myopia,” that is, the “conspicuous reluctance to consider the history and the constraints imposed on the given stimulus sample” (Fiedler, 2012, p. 42; see also Fiedler & Kutzner, *in press*; Hogarth & Soyer, 2015). Ignorantly presuming unbiased samples, however, could have detrimental consequences for choice performance. If recognition is not valid in the set, then sticking to the RH (because it is valid in the domain) could result in worse performance given that other strategies would fare better. Vice versa, if recognition is not valid in the domain, but highly valid in the selected set, not detecting the usefulness of recognition could again be detrimental to one's performance. Our main question thus is whether decision makers base their strategy selection on the domain's cue validities or are able to detect deviations in selected sets and thus let the set's cue validities govern which strategy is chosen.

In the next section, we discuss in more detail what recognition validity is and how it possibly affects choice behavior. Then, in the main part, we report two experiments in which we manipulated the recognition validity in two different ways, such that domain and set validities diverged: In Experiment 1, we used two selected sets of objects with different recognition validities, but drawn from the same domain. In Experiment 2, we used two selected sets with the same recognition validity, but drawn from domains with different recognition validities. Finally, we discuss the implications of our findings for what has been termed “ecological rationality” in decision making (Todd & Gigerenzer, 2012).

Recognition validity

The recognition validity, termed α (Goldstein & Gigerenzer, 2002), is defined as the proportion of cases in which recognition is aligned with the correct inference, out of all cases in which recognition discriminates between the objects in a pair. For example, if recognition points to the correct choice in 70 out of 100 such pairs (and to the wrong choice in the remaining 30), then $\alpha = .70$. This value is well above chance (.50), and thus recognition would be considered a valid, and thus helpful, cue. In other words, the recognition validity reflects the maximum proportion of correct choices one can achieve by consistently following the RH.

The recognition validity is typically approximated as follows (cf. Goldstein & Gigerenzer, 2002). First, a domain is

defined, for example, all Swiss cities with more than 25,000 inhabitants (yielding 43 cities). Second, a sample of participants indicates which of these cities they recognize. Third, all possible pairs of the 43 cities are constructed and the recognition validity is computed for each individual by counting in what percentage of the city pairs consisting of one recognized and one unrecognized city the recognized city is in fact the correct choice. These values are then averaged across all individuals to yield the overall recognition validity for the domain. Alternatively, the Spearman rank correlation r_s between the cities' true size-ranks and their recognition rates can be used to determine the overall recognition validity, namely $\alpha = (1 + r_s)/2$ (see Martignon & Hoffrage, 2002; Pachur, 2010; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, 2012).

Consequently, estimates of the recognition validity depend on several factors, most notably on the exact definition of the domain and on the sample of participants asked whether they recognize the objects.¹ Indubitably, the recognition validity will most likely change if we restrict the domain to cities with more than 50,000 inhabitants (yielding only ten cities in Switzerland), or extend it to cities with more than 10,000 inhabitants (yielding 143 cities in Switzerland). Many such definitions of domains that have been reported in the literature appear rather arbitrary and do not represent natural categories. For example, Hoffrage (2011) reported that recognition rates (for the same objects) depended on the size of the reference class. Similarly, if we ask different samples, say, German versus Canadian participants, to recognize the set of Swiss cities (see Goldstein & Gigerenzer, 2002) the resulting recognition validity will probably differ, too. Similarly, Pohl (in press) asked younger and older children to recognize cities from a list of large world cities and found in two experiments that recognition rates and thus recognition validity differed depending on the children's age.

The conclusion from these observations is that a domain simply does not possess any one “true” recognition validity, because such a value is only derived from averaging across the recognition knowledge of a sample of participants, and not from external, objective sources.² Thus, recognition validity is a rather abstract entity, and we can approach it only by

pooling across a large sample or averaging across studies (thus minimizing error terms). For example, for the domain of large world cities, we found recognition validities of selected sets to range from .57 to .79, $M = .70$ ($SD = .08$), across 15 published data sets. Or, for the domain of single country's largest cities (like Belgium, Canada, Italy, or Switzerland), we found values from .74 to .88, $M = .82$ ($SD = .05$), across 16 published data sets. These mean values might come close to what we could consider a domain's “true” recognition validity, but it is nonetheless clear that recognition validity will not be conclusively determined as any one-point value.

Nonetheless, decision makers seem to have access to a domain's recognition validity, at least with respect to the question of whether recognition is a valid cue or not (see, e.g., Hilbig, Erdfelder, & Pohl, 2010, Data set 7; Pachur, Mata, & Schooler, 2009, Exp. 1; Pohl, 2006, Exp. 1). The corresponding results are reported in the next section. Pachur, Bröder, and Marewski (2008) went even further and asked participants for numerical estimates of the recognition validity (see also Glöckner & Bröder, 2011, 2014). They found in three experiments that subjective estimates of the validities of recognition and three other cues conformed largely to the computed values.³ Similarly, Hoffrage (2011, Exp. 3) asked his participants for subjective estimates of the validity of recognition for objects that were merely recognized versus objects for which further knowledge was available. Participants' mean estimate was larger for objects with additional knowledge and correctly mirrored the computed recognition validities. However, when Hoffrage compared objects from small versus large reference classes, he found that subjective estimates of recognition validity were contrary to the computed values.

Effects of recognition validity

Given that participants—at least on average—are able to accurately estimate the recognition validity, albeit not in all cases (see above), one would in turn expect that they adjust their decision-making behavior accordingly. This ability, to differentiate between domains with different recognition validities, has been termed “environment adaptivity” (Pachur et al., 2009). The empirical evidence shows that decision makers are indeed able to adapt their behavior to features of the respective environment. When recognition is a highly valid cue, participants use the RH on average much more often than when it is an invalid cue. Table 1 summarizes three such findings with respect to adherence rates and the probability of RH use. Adherence rates give the proportion of choosing the recognized objects in pairs with only one object recognized, but are known to be a confounded measure and thus do not adequately reflect true use of the RH (see Hilbig, 2010;

¹ Other influences stem from (a) the questionable dichotomy and stability of recognition judgments (which could be more fine-grained and also change from one assessment to the next; see Castela & Erdfelder, in press; Castela, Kellen, Erdfelder, & Hilbig, 2014; Dougherty et al., 2008; Erdfelder, Küpper-Tetzel, & Mattem, 2011; and Newell & Fernandez, 2006), (b) the historical time of judgment (rank orders could change over time, especially for objects at the lower end of the criterion dimension where absolute differences tend to be smaller), and (c) the extent to which the criterion dimension is well-defined (e.g., several Internet sites provide different rank orders for the size of world cities, depending on the exact definition of what a city is).

² This is different for other probabilistic cues, for example, whether or not a city has an international airport. Such information can be determined objectively and thus does not depend on any specific sample asked. However, subjective knowledge may also differ from reality and lead to other estimates of a cue's validity.

³ Note, however, that cues were explicitly learned in these experiments, so that participants probably had a rather good access to their validity.

Table 1 Results from three studies comparing materials with different recognition validities

Study Experimental conditions	Recognition validity	Adherence rate	Probability of RH use (<i>r</i>)
Pohl (2006, Exp. 1)			
Population of Swiss cities	.86	.89	.66 ^a
Distance of Swiss cities to Interlaken	.51	.54	.08 ^a
Pachur et al. (2009, Exp. 1)			
Population of US cities	.90	.95	.80 ^b
Frequency of infectious diseases	.62	.64	.30 ^b
Hilbig et al. (2010, Data set 7)			
Population of Italian cities	.87	.88	.65
Height above sea level of Italian cities	.53	.59	.20

^a Not reported in Pohl (2006); data later reanalyzed by the author with the *r*-model (Hilbig et al., 2010)

^b Not reported in Pachur et al. (2009); data later reanalyzed by Horn et al. (2015) with a hierarchical extension of the *r*-model, a Beta-MPT model (see Smith & Batchelder, 2010)

Hilbig et al., 2010; Horn, Pachur, & Mata, 2015; Pachur et al., 2012; Pohl, 2011). We thus also report results from model-based analyses using the *r*-model, a multinomial processing tree model that allows an unbiased estimate of RH use (Hilbig et al., 2010; see also Schwikert & Curran, 2014; Horn et al., 2015). Details of the model are reported below.

For example, Pohl (2006, Exp. 1) compared two between-subjects experimental conditions, using 20 large Swiss cities as materials. In one condition he asked participants which of two Swiss cities is the more populous one (“population” group), in the other he asked which of two Swiss cities is closer to the city of Interlaken, which is close to the geographical center of Switzerland (“distance” group). Recognition validity was large in the population group, but at chance in the distance group. Participants’ behavior in terms of adherence rates as well as model-based estimates of RH use (parameter *r*) perfectly reflected this difference. Pachur et al. (2009, Exp. 1) and Hilbig et al. (2010, Data set 7) found similar results.

In addition, two studies provided summaries of such findings in form of scattergrams pitting adherence rate against recognition validity (see Gigerenzer & Goldstein, 2011, Fig. 1; Pachur et al., 2012, Fig. 5-2). Gigerenzer and Goldstein included a large set of 43 studies and reported an impressive correlation of .57 across studies.

Effects of biased selections

However, in all the studies cited above, the selected materials were more or less representative of the underlying domain (or task), so that it remains unclear whether

participants based their assessment of recognition validity on the domain or the selected set. To our knowledge, only three studies—although two of them unintentionally—provide some insight on that question. Pohl (2006, Exp. 4) used the height of mountains as decision domain. The recognition validity for the selected set of mountains happened to be at chance (.49), but participants nevertheless used the RH with a high probability (.80; estimated with the *r*-model). This behavior corresponds to the recognition validity that Goldstein and Gigerenzer (2002) had reported for that domain, namely .85, which probably is a better estimate of the true recognition validity for the height of mountains. Similarly, Pohl (*in press*, Exp. 2) compared children, adolescents, and adults, and used large world cities as materials (that typically have a recognition validity of around .70). The selection of cities turned out to have a recognition validity close to chance ($\approx .50$) for all groups. Yet, all three groups used the RH rather often as indicated by high estimates of the *r*-model, that is, they behaved as if the set had been representative of the underlying domain.

In a recent study, Basehore and Anderson (2016, Exp. 2) deliberately selected three training sets out of the domain of large world cities that typically has a substantial recognition validity of around .70 (see above). The three sets had widely differing recognition validities of .80, .50, and .20 (based on pilot data) and RH use or non-use was trained by giving participants feedback on their choices in a first experimental phase. In a second phase, without feedback and without biased sets, participants showed adherence rates that differed slightly but significantly between the three training conditions, namely .95, .92, and

.83, respectively, thus showing an effect of the training procedure. But again, adherence rates were rather high, reflecting much more the recognition validity in the domain than in the trained, biased sets.

In sum, it appears that individuals are sensitive to diverging recognition validities in different domains and that they adjust their use of the RH accordingly. Further, it appears that individuals base their selection of inference strategies more on the domain's and not the selected set's features. However, the latter conclusion is based on little and more or less casual evidence so far, whereas a thorough investigation by systematically manipulating the domain's and the set's recognition validities is still missing. In the following, we report two such experiments that critically test whether participants adjust their RH use to the domain's or to the set's recognition validity.

Overview of experiments

Both of our experiments followed the same basic procedure (cf. Goldstein & Gigerenzer, 2002; Pohl, 2006). Participants received a number of pairs of cities and were asked for each pair to infer which of the two cities is more populous.⁴ However, the chosen sets of cities were not representative of the underlying domains. Specifically, in Experiment 1, we constructed two sets (from the same domain) with recognition validities that were either below or above the domain's recognition validity. In Experiment 2, we constructed two sets with the same recognition validity, but drawn from two different domains, one with a larger and the other with a lower recognition validity. In addition to the inference task, participants also received a recognition test, in which they judged for each city whether they recognized it or not. To continue the research discussed above (Hoffrage, 2011; Pachur et al., 2008) and as a manipulation check, we also asked participants to provide subjective estimates of the recognition validity in the respective domain.

From the individual recognition judgments, we classified all pairs of cities into knowledge pairs (i.e., both objects recognized), guessing pairs (i.e., neither object recognized), and recognition pairs (i.e., exactly one object recognized, split according to whether the recognized city was the correct or the false choice). The resulting frequencies of pairs in each of these four categories were further split into whether participants made a correct or false inference, thus leading to eight disjoint categories. These frequencies are the input to the r-model to estimate the frequency of RH use and other

parameters (Hilbig et al., 2010). The r-model is a multinomial processing tree model (Batchelder & Riefer, 1999; Erdfelder et al., 2009) that allows the estimation of the proportion of RH use in a direct and unbiased way (Hilbig, 2010). The r-model consists of three trees representing the three possible pairs in the comparison task. It accounts for the observed frequencies of choices in the eight categories by four latent parameters. The most important of these parameters is the probability r of applying the RH, that is, making inferences in recognition cases based on the recognition cue in isolation. In the following analyses, parameters were estimated with multiTree (Moshagen, 2010).

The model was considered to fit the data if its fit value (G^2) was not significant. Following other authors (e.g., Bayen, Erdfelder, Bearden, & Lozito, 2006; Erdfelder, 1984; Moshagen & Erdfelder, 2016), we deliberately chose a lower Type-1 error probability for the model fit to the aggregate data than usual, namely $\alpha = .01$, because otherwise—given the large data set of 36,000 data points in Experiment 1 and 9,000 in Experiment 2—even minute deviations from the model assumptions would result in model misfit. Even with $\alpha = .01$, small deviations from the model (i.e., $w = .10$ as defined by Cohen, 1988) are detected by a $G^2(df = 2)$ goodness-of-fit test with a power exceeding $(1-\beta) = .99$ (computed with G*Power; see Faul, Erdfelder, Buchner, & Lang, 2009). To compare parameter values of the r-model between experimental conditions, we constrained the respective parameters to be equal and tested the resulting decrement in model fit (ΔG^2). If this decrement was significant, the parameters were considered to be different from each other. Just as in case of the omnibus goodness-of-fit tests, $\alpha = .01$ is a reasonable criterion of significance for these $\Delta G^2(df = 1)$ tests, as $\alpha = .01$ is associated with a power exceeding $(1-\beta) = .99$ even for small effects of size $w = .10$. In sum, both our $G^2(2)$ goodness-of-fit tests and the $\Delta G^2(1)$ parameter comparison tests are already very sensitive to small deviations from the null hypotheses in the experiments reported here.

A disadvantage of the G^2 significance tests is that they focus on goodness of fit only and do not take model flexibility into account. In general, one aims at parsimonious models of low complexity with few free parameters that nevertheless fit the data well. To crosscheck whether the r-model variants favored by our G^2 and ΔG^2 test are also those that provide the most parsimonious description of the data structure taking model complexity into account, we additionally compared all variants of r-models (i.e., those with versus those without some or all parameters constrained to be equal across conditions) using two established model selection criteria, the Akaike Information criterion (AIC) and the Fisher Information Approximation (FIA) to the normalized maximum likelihood (see, e.g., Heck, Moshagen, & Erdfelder, 2014; Wu, Myung, & Batchelder,

⁴ Once more relying on the well-established city-size task, our results are readily comparable to many prior studies. Nonetheless, it might also be fruitful if more diverse domains were tested in future studies, including some that bear more resemblance to everyday tasks. Otherwise, one might legitimately question the ecological relevance of the RH.

2010). The AIC criterion combines a badness-of-fit-measure with a penalty term reflecting the number of free parameters in the respective model. For a set of candidate models, AIC selects the model that provides the best approximation to the underlying data structure, in the sense that it minimizes expected information loss (in Kullback-Leibler terms) when approximating the expected frequencies implied by the true model (Wagenmakers & Farrell, 2004). In contrast to AIC, the FIA criterion not only accounts for the number of free parameters but also for model complexity due to functional form, which makes it a particularly attractive model selection criterion provided that the sample size is not too small (see Heck et al., 2014; Hilbig & Moshagen, 2014). For the data we report here, sample size is not a problem.

Following the scarce evidence discussed above (Basehore & Anderson, 2016, Exp. 2; Pohl, 2006, Exp. 4, and in press, Exp. 2), it is plausible to assume that participants will adjust RH use to the domain's recognition validity, ignoring the selected set's recognition validity. In other words, they will behave according to the global- and not the local-optimization hypothesis. So, in Experiment 1 in which two different sets varying in recognition validity were drawn from the same domain, we expected no difference in RH use, whereas in Experiment 2 in which sets with the same recognition validity were drawn from two domains varying in recognition validity we expected clear differences in RH use.

Experiment 1

Methods

Participants and design A total of 120 participants, consisting of 82 females and 38 males, aged between 17 and 55 years ($M = 23.1$, $SD = 5.4$) were recruited at the University of Mannheim. After providing consent and demographic information, participants were randomly assigned to one of two experimental groups. One group received a material set with high and the other with low recognition validity, but drawn from the same domain (see below). Note that our hypothesis was the null, that is, that both groups would show the same RH use.

Material We used the domain of major world cities asking participants to compare pairs of cities with respect to their population. From that domain and based on previous data (Hilbig, 2010; Hilbig et al., 2010; Hilbig, Erdfelder, & Pohl, 2012; Hilbig, Michalkiewicz, Castela, Pohl, & Erdfelder, 2015), we selected one set of 25 cities featuring a high recognition validity of approximately .80 (*high-alpha set condition*)

and a second set of 25 cities featuring a low recognition validity of approximately .50 (*low-alpha set condition*; see Table 3 in Appendix A). Note that α is at chance level in the latter condition, so that recognition is not a valid cue at all. According to a study by Hilbig (2010), the domain of major world cities has a recognition validity of .64. This value is comparable to what was found in other studies using representative sets from that domain ($\alpha \approx .70$). Hence, both selected sets deviated from the underlying domain in terms of recognition validity—one upwards, the other downwards.

To rule out other possible influences, we kept all remaining aspects of the materials constant. First of all, we ensured that both sets included the same number of recognized objects, resembling the recognition proportion of the underlying decision domain (i.e., .60). At the same time, we controlled for knowledge validity. Knowledge validity, labeled β , is defined as the proportion of correct choices for pairs with both objects recognized (Goldstein & Gigerenzer, 2002), and thus is an index of each participant's quality of knowledge in a given set. We selected items such that knowledge validity would be similar across sets and moreover similar to the typical knowledge validity in the underlying domain ($\beta \approx .65$, cf. Hilbig et al., 2010, 2012, 2015). Finally, to avoid an influence of the specific choice of as few as just two specific sets, we created 15 different sets for each of the two conditions (while ensuring that they fulfilled the specific requirements with respect to recognition validity, recognition proportion, and knowledge validity). This way, any behavioral differences in the two conditions could be attributed to differences in recognition validity only.

Procedure Apart from the specific item sets, the procedure was identical for both the high-alpha and the low-alpha conditions. Participants worked on the city-size task, consisting of a recognition task and a paired-comparison task in counterbalanced order. In the recognition task, participants were asked to indicate for the respective set of 25 cities whether or not they had heard of each city before the experiment. For the paired-comparison task, the 25 cities were exhaustively paired resulting in 300 pairs. For each pair, participants were asked to judge which city is more populous. To prevent participants from just clicking through the 300 trials, the comparison task comprised performance-dependent payment. Specifically, participants were informed that they would gain 3 cents for every correct answer, but also lose 3 cents for every wrong answer. However, we provided no feedback during the experiment.

After completion of the city-size task, we asked participants among other things to estimate the recognition validity (cf. Hoffrage, 2011; Pachur et al., 2008). To this end, the concept of validity was explained thoroughly and participants were asked to give individual ratings of the validity of the

Table 2 Parameter estimates of the r-model and subjective recognition-validity estimates in Experiments 1 and 2 (with standard errors in parentheses)

	Experiment 1		Experiment 2	
	High-alpha set	Low-alpha set	High-alpha domain	Low-alpha domain
Domain's α	$\approx .70$	$\approx .70$	$\approx .80$	$\approx .50$
Parameter				
<i>a</i>	.78 (.005)	.52 (.005)	.67 (.01)	.63 (.01)
<i>b</i>	.66 (.005)	.60 (.005)	.70 (.01)	.60 (.01)
<i>g</i>	.53 (.01)	.53 (.01)	.51 (.02)	.53 (.02)
<i>r</i>	.62 (.01)	.62 (.01)	.76 (.02)	.10 (.02)
Estimate of α	.70 (.02)	.69 (.02)	.76 (.02)	.59 (.03)

Note. The labels *a*, *b*, *g*, and *r* refer to the parameters of the r-model and denote recognition validity, knowledge validity, guessing probability, and probability of RH use, respectively. Significant differences between experimental conditions (with $p < .01$) are given in bold type

recognition cue by providing percentage values between 50% and 100%. Finally, participants received feedback about their overall performance, were paid ($M = 2.19$ Euro, $SD = 1.36$), and debriefed.

Results and discussion

The r-model with unconstrained parameters was applied to the aggregated data set containing 36,000 data points and fit these data with $G^2(2) = 7.02$, $p = .03$. The parameter estimates for the model are given in Table 2, separately for the two experimental conditions. As expected, the estimates show a large and significant difference in recognition validity (parameter *a*), $\Delta G^2(1) = 1216$, $p < .001$. In addition, knowledge validity (parameter *b*) was also significantly larger in the high- than in the low-alpha set group, respectively, $\Delta G^2(1) = 71.9$, $p < .001$. In contrast, neither the probability of correct guessing (parameter *g*) nor the probability of RH use (parameter *r*) differed significantly between conditions, $\Delta G^2(1) = .10$, $p = .72$, and $\Delta G^2(1) < .001$, $p = .99$, respectively. Thus, the r-model fit the aggregate frequencies, with *r* and *g* parameters that do not differ between conditions, whereas both the recognition and knowledge validity parameters are larger in the high- than in the low-alpha set condition. We arrive at the same conclusion when we apply the AIC and the FIA model selection measures to the hierarchy of r-model variants (summarized in Table 5 in Appendix B): Considering all possible patterns of equality constraints between conditions, the model with the best (i.e., smallest) AIC and FIA values is Model 6f that assumes equality of the *g* and the *r* parameters between conditions and does not impose

any constraints on the recognition and knowledge validity parameters *a* and *b*.

Importantly, to check whether aggregating frequencies across participants influenced our results, we analyzed the data with two additional methods. First, we applied the r-model to the data of each participant separately and then computed the means of the parameter estimates across participants (excluding non-fitting data sets). Second, we applied a hierarchical variant of the r-model to the data that estimates individual and group parameters and takes correlations between parameters into account (Michalkiewicz & Erdfelder, 2016). This model is based on Klauer's (2010) hierarchical latent-trait approach to MPT models. Both analyses, the individual and the hierarchical approach, virtually yielded the same pattern of results (see Tables 7 and 8 in Appendix C), thus confirming that the aggregate analysis reported above did not produce artificial results due to pooling response frequencies across individuals.

Arguably, the observed (but unintended) difference in knowledge validity could have affected participants' choice behavior. Knowledge validity was significantly larger in the high- than in the low-alpha set condition. Participants in the high-alpha set condition could thus have more often employed their knowledge, reducing RH use in turn (see Bröder & Eichler, 2006; Glöckner & Bröder, 2011, 2014; Hilbig et al., 2015; Hilbig, Pohl, & Bröder, 2009; Newell & Fernandez, 2006; Richter & Späth, 2006). This is, however, unlikely here for two reasons. First, the difference in knowledge validity was relatively small (i.e., .06) compared to the difference in recognition validity (i.e., .26). Second, the hierarchical r-model analysis found that, across both conditions,

knowledge validity did not correlate with RH use ($r = -.04$, 95% Bayesian Credible Interval BCI $[-.25, .17]$). We thus conclude that knowledge validity had little if any influence on RH use in this experiment.

Finally, individual ratings of the recognition validity were in line with participants' choice behavior. Participants judged the recognition cue to be equally valid across conditions (see Table 2); $t(118) = 0.39$, $p = .70$, Cohen's $d = 0.07$, $BF_{10} = 0.21$.⁵

In sum, participants used the RH equally often in both conditions, irrespective of the actual recognition validity of the specific material set. Subjective estimates of recognition validity reflected that. These findings are in line with the idea that people have knowledge about global decision domains, but do not detect deviations in specifically selected (and potentially biased) subsets. In other words, participants behaved in accord with the global-optimization hypothesis, but in contrast to the local-optimization hypothesis. In the next experiment, we tested the reverse situation, using item samples with equal recognition validities, but drawn from domains with different recognition validities.

Experiment 2

Methods

Participants and design A total of 30 participants, consisting of 14 females and 16 males, aged between 17 and 30 years ($M = 21.9$, $SD = 3.1$), were recruited at the University of Mannheim and randomly assigned to one of two groups. The groups differed only with respect to which material they received (see below). Note that we expected RH use to differ between the groups—reflecting the difference in recognition validity of the two underlying domains.

Materials As before, we manipulated recognition validity between participants in two conditions. This time, however, we selected two sets with the same recognition validity from two domains (or, more precisely, from two tasks with the same objects) that featured widely differing recognition validities. We used the largest Italian cities comparing them either with respect to their population (*high-alpha domain condition*) or with respect to their height above sea level (*low-alpha domain condition*).

These materials had been tested in a pre-study ($N = 41$), where participants first made recognition judgments for the largest Italian cities and then compared pairs of cities with respect to either population or height. Recognition validity was almost .80 for the population domain and only slightly exceeded .50 for the height domain (see also Hilbig et al., 2010, Data set 7, who found recognition validities of .87 and .53, respectively). From these domains, we selected two biased samples of 25 cities each with a recognition validity of $\alpha \approx .65$ (see Table 4 in Appendix A). As in Experiment 1, we aimed at choosing sets that mirrored the underlying domains regarding recognition proportion and knowledge validity. Again, we selected 15 different subsets for each condition, that is, a different set of cities for each participant, in order to minimize the potential effect of the choice of specific items.

Procedure To maximize comparability, the second experiment was an exact replication of Experiment 1 in terms of procedure. Participants first worked on the paired-comparison task and the recognition task in random order, followed by individual ratings of recognition validity. As before, they received payment depending on their performance ($M = 2.16$ Euro, $SD = 1.11$).

Results and discussion

Experiment 2 was analyzed in the same way as Experiment 1. The unconstrained r-model was applied to the aggregated data set containing 9,000 data points and fit the data with $G^2(2) = 8.59$, $p = .014$. The parameter estimates of the model are given in Table 2, separately for the two conditions. The estimates show a slight but significant difference in recognition validity (parameter a), $\Delta G^2(1) = 7.80$, $p = .005$. In addition, knowledge validity (parameter b) was also larger in the high- compared to the low-alpha domain condition, $\Delta G^2(1) = 60.6$, $p < .001$. As in Experiment 1, the guessing probability (parameter g) did not differ significantly between conditions, $\Delta G^2(1) = 1.00$, $p = .31$. In contrast, a large and significant difference in RH use (parameter r) was observed, $\Delta G^2(1) = 531$, $p < .001$, with r in the high-alpha condition exceeding r in the low-alpha condition. Thus, according to the ΔG^2 significance tests, only the g parameter can be equated between conditions in Experiment 2. AIC- and FIA-based model selection procedures led to very similar conclusions (see Table 6 in Appendix B). The best model in terms of the AIC criterion is Model 7c that assumes equality of g between conditions and unconstrained a , b , and r parameters. In terms of the FIA criterion, both

⁵ Bayes factors were computed by means of the BayesFactor R package (Rouder, Speckman, Sun, Morey, & Iverson, 2009) and interpreted according to Jeffreys (1961).

Model 7c and Model 6b (assuming $a_1 = a_2$ in addition) performed best. In sum, whereas the manipulation affected RH use and knowledge validity, there is either no effect or a very weak effect on the guessing probability and the recognition validity.

Just as in Experiment 1, two additional analyses, namely an individual-participant analysis based on the means of the r-model parameter estimates for each participant (excluding non-fitting data sets) and a hierarchical r-model analysis (Michalkiewicz & Erdfelder, 2016), yielded mainly the same pattern of results (see Tables 7 and 8 in Appendix C), thus confirming that the aggregate analysis reported above did not produce artificial results based on pooling across individuals. There was only one deviation: The small difference in recognition validity that was significant in some of the aggregate analyses was not significant in either of the two additional analyses.

One could again argue that differences in the sets' recognition validity and in participants' knowledge validity affected choice behavior. Both validities were larger in the high-alpha domain condition. Concerning the recognition validity, however, the difference between conditions was relatively small (i.e., .04), for example, compared to the difference in RH use (i.e., .66). In addition, the FIA model selection favored two models, one of which assumed no difference in the sets' recognition validity. Finally, both the individual and the hierarchical r-model variant did not confirm a significant difference in recognition validity. The hierarchical r-model analysis also did not hint at a reliable correlation between recognition validity and RH use ($r = .21$, 95% BCI $[-.20, .57]$). In sum, it appears unlikely that the recognition validity of the chosen item sets had an impact on RH use.

The difference in knowledge validity was larger and robust across all analyses, reflecting that participants simply know substantially more about the size of Italian cities (high-alpha domain) than the cities' height above sea level (low-alpha domain). As discussed above, better knowledge could lead to more use of this knowledge and thus reduced reliance on the RH (see Bröder & Eichler, 2006; Glöckner & Bröder, 2011, 2014; Hilbig et al., 2009, 2015; Newell & Fernandez, 2006; Richter & Späth, 2006). Note, however, that such an effect would actually reduce the difference in RH use between our conditions: If knowledge validity was the driving aspect, participants should have used the RH *less* often in the high-alpha domain condition (about which they had more valid knowledge) and *more* often in the low-alpha domain condition (about which they had less valid knowledge). In other words, the already large difference we found in RH

use, if anything, might have been even larger if knowledge validity had been comparable across conditions. Thus, we conclude that the difference in RH use is certainly not artificially inflated by differences in knowledge validity. On the contrary, the "true" difference in RH use could be even larger.

Finally, we compared subjective estimates of recognition validity across conditions. In fact, participants in the high-alpha domain condition judged the recognition validity to be considerably higher than participants in the low-alpha domain condition (see Table 2); $t(28) = 4.18$, $p < .001$, Cohen's $d = 1.58$, $BF_{10} = 94.9$.

In sum, we found again that participants based their choice behavior on the domain's recognition validity, that is, participants in the high-alpha domain condition used the RH considerably more often than participants in the low-alpha domain condition. Subjective estimates of recognition validity again reflected that. The results of Experiment 2 thus further support the global-optimization hypothesis stating that people base their strategy selection on the accuracy of a strategy as determined by the global decision domain, but not specific sets thereof (as the local-optimization hypothesis predicted).

General discussion

In this paper, we set out to investigate whether decision makers base their use of the recognition heuristic (RH) (Goldstein & Gigerenzer, 2002), a simple inference strategy, on the validity of recognition as a cue in the underlying domain or in the specific set of to-be-compared objects. We referred to these two possibilities as the global- and the local-optimization hypothesis, respectively. Some evidence from previous studies reported in the Introduction is compatible with the global-optimization hypothesis (Basehore & Anderson, 2016; Pohl, 2006, *in press*), but conclusive experimental tests were missing. We conducted two experiments systematically varying recognition validity in the domain and in the chosen set of objects. In Experiment 1, we selected two sets with rather different recognition validities drawn from the same domain. In Experiment 2, we selected two sets with the same recognition validity drawn from two (task) domains with rather different recognition validities. The results from both experiments clearly show that it is the domain's recognition validity that impacts RH use, whereas the specific recognition validity in the set of to-be-compared objects is essentially ignored, thus supporting the global-optimization hypothesis and refut-

ing the local-optimization hypothesis. Subjective estimates of the recognition validity reflected the domain's and not the set's validity, too, thus corroborating that participants did not detect (or not even searched for) deviating characteristics of the presented set of materials. We may thus conclude that participants (a) have access to the recognition validity in a domain, (b) adjust their use of the RH accordingly, and (c) treat even substantially biased item sets as representative of their underlying domain.

One conclusion from these results is that it is questionable to analyze participants' strategic behavior with respect to the sample's recognition validity (at least if the latter is not representative). The recognition validity in the selected set of objects is only important insofar as to check whether the set is representative of the domain, or whenever the overall accuracy of participants' judgments is in the focus.

Our results are reminiscent of discussions regarding the role of representative design (see, e.g., Dhimi, Hertwig, & Hoffrage, 2004; Gigerenzer & Hoffrage, 1995; Hoffrage & Hertwig, 2006). In the 1970s and 1980s, research focused on many biases and illusions in judgment and decision making, most prominently in the “heuristics and biases” program of Kahneman and Tversky and their collaborators (see Gilovich, Griffin, & Kahneman, 2002; Kahneman, Slovic, & Tversky, 1982; Pohl, 2017). Several of these biases were subsequently refuted as not showing deficiencies of the human information-processing system, but rather stemming from using artificial, non-representative materials and thus fooling participants into a faulty behavior (Gigerenzer, 1991; Gigerenzer & Hoffrage, 1995).

For example, the overconfidence effect (see Hoffrage, 2017, for a summary), that is, being overconfident on how good one's evaluations of true/false assertions are, appeared only for non-representative material selections comprised of overly difficult items. For representative materials, the effect was reduced or even absent (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994; Juslin, Winman, & Olsson, 2000; Winman, 1997). Thus, it seems that in such cases—just as in our experiments—participants simply assumed representative material sets, for which their behavior would have been well calibrated, but then received biased materials of which they were not aware, which in turn led to non-optimal behavior. As a second example, hindsight bias, that is, the tendency to overestimate one's pre-outcome knowledge (see Pohl & Erdfelder, 2017, for a general overview), has been found to be much larger for almanac questions than for case histories (Christensen-Szalanski & Willham, 1991). As an explanation, Winman (1997) proposed that the selection of almanac-

items to be used as materials was not random, but included on average overly difficult questions. He accordingly found that hindsight bias was reduced or even absent when using random samples (see also Winman, Juslin, & Björkman, 1998). In other words, at least part of the bias was in the material selection, not in the participants' behavior. Further examples of phenomena arguably caused by biased samples include illusory correlations and confirmation bias (Fiedler & Kutzner, *in press*), biased weighting of small probabilities in risky choice (Glöckner, Hilbig, Henninger, & Fiedler, 2016; Hilbig & Glöckner, 2011), or unrealistic optimism (Harris & Hahn, 2011).

Nonetheless, one may question why participants do not detect such biased samples and then adapt their behavior; by implication, one may question whether the behavior observed should still be considered “ecologically rational” (Todd & Gigerenzer, 2012). In our two experiments, participants applied useless cues indifferently or did not detect useful ones. This could indeed be considered a case of meta-cognitive myopia (Fiedler, 2008, 2012), a term that highlights shortsightedness in a variety of monitoring processes, for example, in erroneously accepting a sample as representative of the underlying domain (see also Hogarth & Soyer, 2015). However, maybe participants' behavior is not so ignorant and the bias not so obvious. One argument could be that participants—as a default—treat all selections as if they were representative, at least as long as they do not have any reason to doubt their assumption. It might, moreover, be clear that a sample is never perfectly representative of the domain, even if randomly selected (as, e.g., in the two unintended conditions with recognition validity at chance level reported in the Introduction), but on average it will be close to being representative. Besides, there are myriads of potentially deviating features, like number, values, distributions, correlations, and validities (Hoffrage & Hertwig, 2006), which one would need to check. Some of these may also be difficult if not impossible to observe and evaluate (Fiedler, 2012).

It thus appears straightforward and much more efficient to simply assume random (and thus approximately representative) samples, despite the caveat of some (typically minor) costs in accuracy. Such detriments in performance would probably be more than outweighed by the costs for checking all potentially relevant features. Only when the selected set of objects deviates blatantly from what one would expect or if decision makers have reason to suspect biased sampling, could they be inclined to reflect on the sample's features and then adapt their behavior accordingly. But in reality this might be a rare case so that sticking to the assumption of a representative sample appears rather efficient and thus ecologically rational after all.

Appendix A: Materials

Table 3 Example of selected material sets in Experiment 1. The two subsets were selected out of the 61 most populous world cities*. Lower ranks indicate the larger population. In the experiment, city names were given in German spelling

High-alpha condition		Low-alpha condition	
Rank	City	Rank	City
2	Shanghai	3	Karachi
5	Istanbul	6	São Paulo
8	Seoul	7	Moscow
10	Mexico City	12	Jakarta
11	Tokyo	21	Ho Chi Minh City
15	Lagos	24	Dhaka
17	Kinshasa	27	Lahore
19	London	29	Bangalore
21	Ho Chi Minh City	31	Bagdad
22	Bangkok	32	Kolkata
24	Dhaka	34	Singapore
26	Guangzhou	35	Chongqing
27	Lahore	37	St. Petersburg
29	Bangalore	39	Riyadh
30	Tianjin	41	Shenyang
35	Chongqing	42	Rangoon
36	Chennai	43	Hyderabad
39	Riyadh	46	Johannesburg
43	Hyderabad	47	Los Angeles
44	Ahmedabad	51	Cape Town
45	Ankara	54	Pune
48	Abidjan	56	Madrid
55	Pyongyang	58	Jaipur
57	Kanpur	59	Buenos Aires
58	Jaipur	61	Jeddah

* http://en.wikipedia.org/wiki/List_of_cities_proper_by_population, retrieved July 2011

Table 4 Example of selected material sets in Experiment 2. The two subsets were selected out of the 30 most populous Italian cities*. Lower ranks indicate the larger population in the high-alpha condition and the larger height above sea level in the low-alpha condition. In the experiment, city names were given in German spelling

High-alpha condition		Low-alpha condition	
Rank	City	Rank	City
1	Rome	1	Perugia
3	Naples	2	Turin
5	Palermo	4	Milan
9	Bari	5	Foggia
10	Catania	7	Verona
11	Venice	9	Parma
12	Verona	10	Bologna
13	Messina	11	Florence
14	Padua	12	Rome
15	Trieste	13	Modena
16	Taranto	14	Palermo
17	Brescia	15	Reggio Calabria
18	Prato	16	Genoa
19	Reggio Calabria	17	Naples
20	Modena	18	Taranto
21	Parma	19	Padua
22	Perugia	20	Ferrara
23	Reggio nell'Emilia	21	Catania
24	Livorno	22	Salerno
25	Cagliari	24	Cagliari
26	Foggia	25	Ravenna
27	Ravenna	26	Messina
28	Salerno	27	Livorno
29	Rimini	29	Bari
30	Ferrara	30	Venice

* http://de.wikipedia.org/wiki/Liste_der_Städte_in_Italien, retrieved September 2013

Appendix B: Model selection measures

Table 5 FIA- and AIC-model selection measures for the r-model hierarchy applied to Experiment 1

M8: $\{(a_1, a_2) (b_1, b_2) (g_1, g_2) (r_1, r_2)\}$ FIA = 30148 AIC = 60242					
M7a: $\{(a_1=a_2) (b_1, b_2) (g_1, g_2) (r_1, r_2)\}$ FIA = 30752 AIC = 61456	M7b: $\{(a_1, a_2) (b_1=b_2) (g_1, g_2) (r_1, r_2)\}$ FIA = 30180 AIC = 60312	M7c: $\{(a_1, a_2) (b_1, b_2) (g_1=g_2) (r_1, r_2)\}$ FIA = 30145 AIC = 60240	M7d: $\{(a_1, a_2) (b_1, b_2) (g_1, g_2) (r_1=r_2)\}$ FIA = 30145 AIC = 60240		
M6a: $\{(a_1=a_2) (b_1=b_2) (g_1, g_2) (r_1, r_2)\}$ FIA = 30784 AIC = 61526	M6b: $\{(a_1=a_2) (b_1, b_2) (g_1=g_2) (r_1, r_2)\}$ FIA = 30748 AIC = 61454	M6c: $\{(a_1=a_2) (b_1, b_2) (g_1, g_2) (r_1=r_2)\}$ FIA = 30749 AIC = 61454	M6d: $\{(a_1, a_2) (b_1=b_2) (g_1=g_2) (r_1, r_2)\}$ FIA = 30176 AIC = 60310	M6e: $\{(a_1, a_2) (b_1=b_2) (g_1, g_2) (r_1=r_2)\}$ FIA = 30177 AIC = 60311	M6f: $\{(a_1, a_2) (b_1, b_2) (g_1=g_2) (r_1=r_2)\}$ FIA = 30141 AIC = 60238
M5a: $\{(a_1=a_2) (b_1=b_2) (g_1=g_2) (r_1, r_2)\}$ FIA = 30780 AIC = 61524	M5b: $\{(a_1=a_2) (b_1=b_2) (g_1, g_2) (r_1=r_2)\}$ FIA = 30781 AIC = 61569	M5c: $\{(a_1=a_2) (b_1, b_2) (g_1=g_2) (r_1=r_2)\}$ FIA = 30745 AIC = 61452	M5d: $\{(a_1, a_2) (b_1=b_2) (g_1=g_2) (r_1=r_2)\}$ FIA = 30173 AIC = 60309		
M4: $\{(a_1=a_2) (b_1=b_2) (g_1=g_2) (r_1=r_2)\}$ FIA = 30777 AIC = 61523					

Note. Model labels (e.g., M8) denote the number of free parameters across the two experimental conditions. Lowercase letters following the number of free parameters (e.g., a, b, c, ...) indicate different model versions with the same number of free parameters. Model parameters of the r model (i.e., a, b, g, r) are listed in parentheses, with the index denoting the experimental condition. “=” indicates equality of parameters across conditions, “,” denotes unconstrained parameters

FIA Fisher Information Approximation, AIC Akaike Information Criterion

Best-fitting FIA and AIC values are given in bold type

Table 6 FIA- and AIC-model selection measures for the r-model hierarchy applied to Experiment 2

M8: $\{(a_1, a_2) (b_1, b_2)\}$ $(g_1, g_2) (r_1, r_2)\}$ FIA = 8001 AIC = 15958					
M7a: $\{(a_1=a_2) (b_1, b_2)\}$ $(g_1, g_2) (r_1, r_2)\}$ FIA = 8001 AIC = 15964	M7b: $\{(a_1, a_2) (b_1=b_2)\}$ $(g_1, g_2) (r_1, r_2)\}$ FIA = 8028 AIC = 16017	M7c: $\{(a_1, a_2) (b_1, b_2)\}$ $(g_1=g_2) (r_1, r_2)\}$ FIA = 7999 AIC = 15957	M7d: $\{(a_1, a_2) (b_1, b_2)\}$ $(g_1, g_2) (r_1=r_2)\}$ FIA = 8273 AIC = 16505		
M6a: $\{(a_1=a_2) (b_1=b_2)\}$ $(g_1, g_2) (r_1, r_2)\}$ FIA = 8028 AIC = 16022	M6b: $\{(a_1=a_2) (b_1, b_2)\}$ $(g_1=g_2) (r_1, r_2)\}$ FIA = 7999 AIC = 15963	M6c: $\{(a_1=a_2) (b_1, b_2)\}$ $(g_1, g_2) (r_1=r_2)\}$ FIA = 8273 AIC = 16511	M6d: $\{(a_1, a_2) (b_1=b_2)\}$ $(g_1=g_2) (r_1, r_2)\}$ FIA = 8025 AIC = 16016	M6e: $\{(a_1, a_2) (b_1=b_2)\}$ $(g_1, g_2) (r_1=r_2)\}$ FIA = 8319 AIC = 16603	M6f: $\{(a_1, a_2) (b_1, b_2)\}$ $(g_1=g_2) (r_1=r_2)\}$ FIA = 8270 AIC = 16504
M5a: $\{(a_1=a_2) (b_1=b_2)\}$ $(g_1=g_2) (r_1, r_2)\}$ FIA = 8025 AIC = 16022	M5b: $\{(a_1=a_2) (b_1=b_2)\}$ $(g_1, g_2) (r_1=r_2)\}$ FIA = 8319 AIC = 16609	M5c: $\{(a_1=a_2) (b_1, b_2)\}$ $(g_1=g_2) (r_1=r_2)\}$ FIA = 8270 AIC = 16510	M5d: $\{(a_1, a_2) (b_1=b_2)\}$ $(g_1=g_2) (r_1=r_2)\}$ FIA = 8316 AIC = 16602		
M4: $\{(a_1=a_2) (b_1=b_2)\}$ $(g_1=g_2) (r_1=r_2)\}$ FIA = 8317 AIC = 16608					

Note. Model labels (e.g., M8) denote the number of free parameters across the two experimental conditions. Lowercase letters following the number of free parameters (e.g., a, b, c ...) indicate different model versions with the same number of free parameters. Model parameters of the r model (i.e., a, b, g, r) are listed in parentheses, with the index denoting the experimental condition. “=” indicates equality of parameters across conditions, “,” denotes unconstrained parameters

FIA Fisher Information Approximation, AIC Akaike Information Criterion

Best-fitting FIA and AIC values are given in bold type

Appendix C: Additional r-model analyses

Table 7 Mean parameter values (and standard errors) from separate r-model analyses per individual (excluding 18 non-fitting data sets in Exp. 1 and 6 in Exp. 2)

Parameter	Experiment 1		Experiment 2	
	High-alpha set	Low-alpha set	High-alpha domain	Low-alpha domain
<i>a</i>	.78 (.01)	.55 (.01)	.66 (.01)	.64 (.02)
<i>b</i>	.64 (.01)	.59 (.01)	.69 (.02)	.62 (.01)
<i>g</i>	.49 (.02)	.56 (.01)	.56 (.03)	.53 (.03)
<i>r</i>	.64 (.02)	.63 (.02)	.75 (.03)	.26 (.06)

Note. The labels *a*, *b*, *g*, and *r* refer to the parameters of the r-model and denote recognition validity, knowledge validity, guessing probability, and probability of RH use, respectively. Significant differences based on two-groups *t*-tests (with $p < .05$) applied to the individual parameter estimates are given in bold type

Table 8 Estimated means of parameter values (and 95% BCIs) from a hierarchical r-model analysis (see Michalkiewicz & Erdfelder, 2016)

Parameter	Experiment 1		Experiment 2	
	High-alpha set	Low-alpha set	High-alpha domain	Low-alpha domain
<i>a</i>	.79 [.77, .81]	.53 [.50, .55]	.67 [.63, .71]	.63 [.58, .68]
<i>b</i>	.65 [.63, .67]	.60 [.57, .62]	.71 [.65, .76]	.60 [.56, .65]
<i>g</i>	.53 [.50, .56]	.54 [.51, .56]	.51 [.47, .56]	.53 [.49, .58]
<i>r</i>	.61 [.51, .70]	.63 [.53, .72]	.77 [.68, .85]	.11 [.00, .41]

Note. The labels *a*, *b*, *g*, and *r* refer to the parameters of the r-model and denote recognition validity, knowledge validity, guessing probability, and probability of RH use, respectively. Significant differences (i.e., 95% Bayesian credible intervals of the difference between parameter estimates of the two conditions do not include zero; cf. Horn et al., 2015; Michalkiewicz & Erdfelder, 2016) are given in bold type

Acknowledgments This research was supported by Grant No. ER 224/2-2 from the German Research Foundation (Deutsche Forschungsgemeinschaft, Bonn, Germany). We thank Katharina Fischer, Sabine Schellhaas, Nicoletta Symeonidou, and Johannes Prager for their help in data collection.

References

- Basehore, Z., & Anderson, R. B. (2016). The simple life: New experimental tests of the recognition heuristic. *Judgment and Decision Making*, 11(3), 301–309.
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86. doi:10.3758/BF03210812
- Bayen, U. J., Erdfelder, E., Bearden, J. N., & Lozito, J. P. (2006). The interplay of memory and judgment processes in effects of aging on hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 1003–1018. doi:10.1037/0278-7393.32.5.1003
- Bröder, A., & Eichler, A. (2006). The use of recognition information and additional cues in inferences from memory. *Acta Psychologica*, 121, 275–284. doi:10.1016/j.actpsy.2005.07.001
- Castela, M., & Erdfelder, E. (in press). The memory state heuristic: A formal model based on repeated recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
- Castela, M., Kellen, D., Erdfelder, E., & Hilbig, B. E. (2014). The impact of subjective recognition experiences on recognition heuristic use: A multinomial processing tree approach. *Psychonomic Bulletin & Review*, 21, 1131–1138. doi:10.3758/s13423-014-0587-4
- Christensen-Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 48(1), 147–168. doi:10.1016/0749-5978(91)90010-Q
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dhimi, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6), 959–988. doi:10.1037/0033-2909.130.6.959
- Dougherty, M. R. P., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review*, 115(1), 199–213. doi:10.1037/0033-295X.115.1.199
- Erdfelder, E. (1984). Zur Bedeutung und Kontrolle des beta-Fehlers bei der inferenzstatistischen Prüfung log-linearer Modelle [On significance and control of the β error in statistical tests of log-linear models]. *Zeitschrift für Sozialpsychologie*, 15, 18–32.

- Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, 217, 108–124. doi:10.1027/0044-3409.217.3.108
- Erdfelder, E., Küpper-Tetzl, C. E., & Mattem, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making*, 6(1), 7–22.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analysis using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. doi:10.3758/BRM.41.4.1149
- Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(1), 186–203. doi:10.1037/0278-7393.34.1.186
- Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 1–55). San Diego, CA: Elsevier Academic Press.
- Fiedler, K., & Kutzner, F. (in press). Information sampling and reasoning biases: Implications for research in judgment and decision making. In G. Keren & G. Wu (Eds.), *Blackwell handbook of judgment and decision making: An interdisciplinary perspective*. Malden, MA: Blackwell.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases”. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 2, pp. 83–115). Chichester: Wiley.
- Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, 6(1), 100–121.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4), 684–704. doi:10.1037/0033-295X.102.4.684
- Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98(4), 506–528. doi:10.1037/0033-295X.98.4.506
- Gigerenzer, G., Todd, P. M., & the ABC Research Group (Eds.). (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
- Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.
- 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(1), 23–42.
- Glöckner, A., & Bröder, A. (2014). Cognitive integration of recognition information and additional cues in memory-based decisions. *Judgment and Decision Making*, 9(1), 35–50.
- Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General*, 145, 486–508. doi:10.1037/a0040103
- Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1), 75–90. doi:10.1037/0033-295X.109.1.75
- Harris, A. J. L., & Hahn, U. (2011). Unrealistic optimism about future life events: A cautionary note. *Psychological Review*, 118(1), 135–154. doi:10.1037/a0020997
- Heck, D. W., Moshagen, M., & Erdfelder, E. (2014). Model selection by minimum description length: Lower-bound sample sizes for the Fisher information approximation. *Journal of Mathematical Psychology*, 60, 29–34. doi:10.1016/j.jmp.2014.06.002
- Hilbig, B. E. (2010). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making*, 5, 272–284.
- 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. doi:10.1037/a0017518
- Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2012). A matter of time: Antecedents of one-reason decision making based on recognition. *Acta Psychologica*, 141, 9–16. doi:10.1016/j.actpsy.2012.05.006
- Hilbig, B. E., & Glöckner, A. (2011). Yes, they can! Appropriate weighting of small probabilities as a function of information acquisition. *Acta Psychologica*, 138, 390–396. doi:10.1016/j.actpsy.2011.09.005
- Hilbig, B. E., Michalkiewicz, M., Castela, M., Pohl, R. F., & Erdfelder, E. (2015). Whatever the cost? Information integration in memory-based inferences depends on cognitive effort. *Memory & Cognition*, 43, 659–671. doi:10.3758/s13421-014-0493-z
- Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review*, 21, 1431–1443. doi:10.3758/s13423-014-0643-0
- Hilbig, B. E., Pohl, R. F., & Bröder, A. (2009). Criterion knowledge: A moderator of using the recognition heuristic? *Journal of Behavioral Decision Making*, 22(5), 510–522. doi:10.1002/bdm.644
- Hoffrage, U. (2011). Recognition judgments and the performance of the recognition heuristic depend on the size of the reference class. *Judgment and Decision Making*, 6(1), 43–57.
- Hoffrage, U. (2017). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory* (2nd ed., pp. 291–314). London and New York: Routledge.
- 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). New York: Cambridge University Press.
- Hogarth, R. M., & Soyer, E. (2015). Providing information for decision making: Contrasting description and simulation. *Journal of Applied Research in Memory and Cognition*, 4, 221–228. doi:10.1016/j.jarmac.2014.01.005
- Horn, S., Pachur, T., & Mata, R. (2015). How does aging affect recognition-based inference? A hierarchical Bayesian modeling approach. *Acta Psychologica*, 154, 77–85. doi:10.1016/j.actpsy.2014.11.001
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.
- Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes*, 57, 226–246. doi:10.1006/obhd.1994.1013
- Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, 107, 384–396. doi:10.1037/0033-295X.107.2.384
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
- Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75, 70–98. doi:10.1007/S11336-009-9141-0
- Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52(1), 29–71. doi:10.1023/A:1015516217425
- Michalkiewicz, M., & Erdfelder, E. (2016). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition*, 44, 454–468. doi:10.3758/s13421-015-0567-6
- Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, 42, 42–54. doi:10.3758/BRM.42.1.42

- Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling*, 23, 54–60. doi:[10.1080/10705511.2014.950896](https://doi.org/10.1080/10705511.2014.950896)
- Newell, B. R., & Fernandez, D. (2006). On the binary quality of recognition and the inconsequentiality of further knowledge: Two critical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, 19(4), 333–346. doi:[10.1002/bdm.531](https://doi.org/10.1002/bdm.531)
- Pachur, T. (2010). Recognition-based inference: When is less more in the real world? *Psychonomic Bulletin & Review*, 17(4), 589–598. doi:[10.3758/PBR.17.4.589](https://doi.org/10.3758/PBR.17.4.589)
- Pachur, T., Bröder, A., & Marewski, J. N. (2008). The recognition heuristic in memory-based inference: Is recognition a non-compensatory cue? *Journal of Behavioral Decision Making*, 21, 183–210. doi:[10.1002/bdm.581](https://doi.org/10.1002/bdm.581)
- Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 983–1002. doi:[10.1037/0278-7393.32.5.983](https://doi.org/10.1037/0278-7393.32.5.983)
- Pachur, T., Mata, R., & Schooler, L. J. (2009). Cognitive aging and the adaptive use of recognition in decision making. *Psychology and Aging*, 24(4), 901–915. doi:[10.1037/a0017211](https://doi.org/10.1037/a0017211)
- Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (2012). When is the recognition heuristic an adaptive tool? In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 113–143). New York: Oxford University Press.
- Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision Making*, 19(3), 251–271. doi:[10.1002/bdm.522](https://doi.org/10.1002/bdm.522)
- Pohl, R. F. (2011). Recognition information in inferential decision making: An overview of the debate. *Judgment and Decision Making*, 6(5), 423–438.
- Pohl, R. F. (Ed.). (2017). *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed.). London and New York: Routledge.
- Pohl, R. F. (in press). Measuring age-related differences in using a simple decision strategy: The case of the recognition heuristic. *Zeitschrift für Psychologie*.
- Pohl, R. F., & Erdfelder, E. (2017). Hindsight bias. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in thinking, judgment, and memory* (2nd ed., pp. 424–445). London and New York: Routledge.
- Richter, T., & Späth, P. (2006). Recognition is used as one cue among others in judgment and decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(1), 150–162. doi:[10.1037/0278-7393.32.1.150](https://doi.org/10.1037/0278-7393.32.1.150)
- Rouder, J. M., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. doi:[10.3758/PBR.16.2.225](https://doi.org/10.3758/PBR.16.2.225)
- Schwikert, S. R., & Curran, T. (2014). Familiarity and recollection in heuristic decision making. *Journal of Experimental Psychology: General*, 143(6), 2341–2365. doi:[10.1037/xge0000024](https://doi.org/10.1037/xge0000024)
- Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, 54, 167–183. doi: [10.1016/j.jmp.2009.06.007](https://doi.org/10.1016/j.jmp.2009.06.007)
- Todd, P. M., & Gigerenzer, G. (2012). What is ecological rationality? In P. M. Todd, G. Gigerenzer, & the ABC Research Group (Eds.), *Ecological rationality: Intelligence in the world* (pp. 3–30). New York: Oxford University Press.
- Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192–196. doi:[10.3758/BF03206482](https://doi.org/10.3758/BF03206482)
- Winman, A. (1997). The importance of item selection in "knew-it-all-along" studies of general knowledge. *Scandinavian Journal of Psychology*, 38(1), 63–72. doi:[10.1111/1467-9450.00010](https://doi.org/10.1111/1467-9450.00010)
- Winman, A., Juslin, P., & Björkman, M. (1998). The confidence-hindsight mirror effect in judgment: An accuracy-assessment model for the knew-it-all-along phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(2), 415–431. doi:[10.1037/0278-7393.24.2.415](https://doi.org/10.1037/0278-7393.24.2.415)
- Wu, H., Myung, J. I., & Batchelder, W. H. (2010). On the minimum description length complexity of multinomial processing tree models. *Journal of Mathematical Psychology*, 54(3), 291–303. doi:[10.3758/PBR.17.3.275](https://doi.org/10.3758/PBR.17.3.275)