Judgment and Decision Making, Vol. 6, No. 1, February 2011, pp. 43–57

Recognition judgments and the performance of the recognition heuristic depend on the size of the reference class

Ulrich Hoffrage*

Abstract

In a series of three experiments, participants made inferences about which one of a pair of two objects scored higher on a criterion. The first experiment was designed to contrast the prediction of Probabilistic Mental Model theory (Gigerenzer, Hoffrage, & Kleinbölting, 1991) concerning sampling procedure with the hard-easy effect. The experiment failed to support the theory's prediction that a particular pair of randomly sampled item sets would differ in percentage correct; but the observation that German participants performed practically as well on comparisons between U.S. cities (many of which they did not even recognize) than on comparisons between German cities (about which they knew much more) ultimately led to the formulation of the recognition heuristic. Experiment 2 was a second, this time successful, attempt to unconfound item difficulty and sampling procedure. In Experiment 3, participants' knowledge and recognition of each city was elicited, and how often this could be used to make an inference was manipulated. Choices were consistent with the recognition heuristic in about 80% of the cases when it discriminated and people had no additional knowledge about the recognized city (and in about 90% when they had such knowledge). The frequency with which the heuristic could be used affected the percentage correct, mean confidence, and overconfidence as predicted. The size of the reference class, which was also manipulated, modified these effects in meaningful and theoretically important ways.

Keywords: recognition heuristic, reference class, probabilistic inference, overconfidence, hard-easy effect.

1 Introduction

The history of science and technology repeatedly demonstrates that many laws are discovered and many inventions are made serendipitously, as a by-product when researchers are striving for something else. The recognition heuristic is just one example of this: It was formulated as a post-hoc explanation for a puzzling finding that was observed while attempting to test a specific prediction of the theory of Probabilistic Mental Models (PMM; Gigerenzer, Hoffrage, & Kleinbölting, 1991). While the other papers contained in this series of special issues and many of the references given therein illustrate how stimulating the formulation of the recognition heuristic was and how much research it has spurred, the present paper turns back the clock and reports three studies that were conducted in the late 1980's and early 1990's (Hoffrage, 1995).[1]

This paper is organized as follows: The first part provides the historical context that led to the formulation of the recognition heuristic. At the outset of this part, a brief summary of PMM theory is given. Experiment 1 is then reported, which was conducted to address one of the criticisms of the theory, namely the confounding of sampling procedure and item difficulty. Specifically, we compared over/underconfidence in two item sets that were generated by the same sampling procedure but were nevertheless supposed to differ with respect to percentage correct. This attempt failed, yielding the counter-intuitive finding that German participants performed about the same when making comparisons between German cities as when making comparisons between U.S. cities. Experiment 2 reports a second, and this time successful, attempt to unconfound item difficulty and sampling procedure in order to answer the question that motivated Experiment 1. The second part is also historical: It reports Experiment 3, which provides, to the best of my knowledge, the first empirical test of the recognition heuristic. It was designed to find out whether the results obtained in Experiment 1 could be explained by participants having used the recognition heuristic. In this experiment, the participants' knowledge and recognition of each city was elicited, and how often this could be used to make an inference was manipulated. We also manipulated the inclusion criterion (and, in turn, the size) of the reference class that the cities were drawn from when constructing

I would like to thank Horst Kilcher and Sabine Schmuck for many discussions and for their help with collecting the data, Chris White for editing the text, and Gerd Gigerenzer, Dan Goldstein, Sebastian Hafenbrädl, Jens Lange, Hans-Joachim Lenz, Chris White, Jan K. Woike, the special issue editors, Jon Baron, and three anonymous reviewers for helpful comments on an earlier version of the present article.

*Faculty of Business and Economics, University of Lausanne. Email: ulrich.hoffrage@unil.ch.

[1] It is embarrassing to say that I made no attempt to publish these studies earlier, but I was determined to leave academia after I completed my dissertation. When I revised my decision and returned, new and exciting projects pushed me in other directions (see, e.g., Gigerenzer, Todd, and the ABC Research Group, 1999).

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

the paired comparisons. The last part links the three "historical" experiments to later studies and measures, and discusses the theoretical relevance of the work described here.

2 The historical context of the recognition heuristic

2.1 The theory of Probabilistic Mental Models

Independently, Gigerenzer et al. (1991) with their PMM theory and Juslin (1994) developed what was later termed "ecological models" (McClelland & Bolger, 1994). When solving a task such as "Which city has more inhabitants, A or B?" people construct a PMM (unless they have direct knowledge or can deduce the answer with certainty, which we called a "local mental model"; Gigerenzer et al., 1991). By searching for probabilistic cues that discriminate between the two alternatives, the question is put into a larger context. Imagine that a search hits on the soccer-team cue: City A has a soccer team in the major league and City B does not. Based on the literature on automatic frequency processing, PMM theory posits that people are able to estimate the ecological validity of cues (as long as the objects belong to their natural environment). This validity is defined by the relative frequency of cases in the environment where the cue indicates the correct answer, given that the cue discriminates. For instance, the validity of the soccer-team cue is 90% (in the complete set of paired comparisons of all German cities with more than 100,000 inhabitants). If participants choose the city to which the cue points and report the cue validity as their confidence, they should be well calibrated. This, however, is true only if the cue validities in the item sample reflect the cue validities in the population. If researchers do not sample general-knowledge questions randomly, but over-represent items in which cue-based inferences would lead to wrong choices, overconfidence will occur. Such overconfidence does not reflect fallible reasoning processes but is an artifact of the way the experimenter sampled the stimuli and ultimately misrepresented the cue-criterion relations in the ecology. In two experiments, Gigerenzer et al. (1991) found exactly this: overconfidence was observed for a set of selected items, but disappeared when the objects that were used in the paired comparisons were randomly sampled from a defined reference class.

The theory can also account for the common finding that average confidence judgments exceed average frequency estimates ("How many of the last 50 items did you answer correctly?") by positing that different reference classes are used for the two kind of judgments (for details, see Gigerenzer et al., 1991). When PMM theory was first published we had a long list of criticisms and open questions that, in turn, gave rise to a series of studies in which attempts were made to falsify the theory in a true Popperian fashion.[2]

2.2 A failed attempt to unconfound sampling procedure and item difficulty (Experiment 1)

One of the established findings in research on overconfidence is the hard-easy effect (Hoffrage, 2004; Lichtenstein & Fischhoff, 1977) according to which overconfidence covaries with item difficulty: Hard item sets (i.e., those with a percentage of correct answers of about 75% or lower in a two-alternative forced-choice task) tend to produce overconfidence, whereas easy sets (i.e., those with a percentage correct of about 75% or higher) tend to produce underconfidence.

One of the problems of PMM theory was the fact that selected item sets turned out to be hard (e.g., for Experiment 1 and 2 of Gigerenzer et al., 1991, percentage correct was 52.9 and 56.2, respectively), whereas representative item sets turned out to be relatively easy (71.7 and 75.3, respectively). Therefore, even though PMM theory correctly predicted that overconfidence disappeared for the representative sets while it could be observed for the selected set, these findings could, at the same time, be seen as just another example of the hard-easy effect. Hoffrage (1995) tried to shed some light on this issue by comparing two item sets, each consisting of paired comparisons for which the objects were generated by the same, representative, sampling procedure but the difficulty of these sets still differed (see also Kilcher, 1991). If PMM theory was correct, then overconfidence should disappear in both sets, whereas the hard-easy effect would be observed if there was overconfidence for the hard set, but no overconfidence for the easy set.

2.2.1 Method

Participants were mainly students of the University of Constance, Germany (n=56; 12 female, 44 male). Their task was to (1) repeatedly select, in a series of paired comparisons among cities, the city with more inhabitants, and (2) indicate their confidence in the correctness of their choices on a scale ranging from 50–100% in increments of 10%. Two item sets were used: comparisons

[2] Although I was very sceptical about PMM theory and preferred to address these issues before publishing anything, the first author, Gerd Gigerenzer, replied that this would take considerable time and at the end I would probably have discovered more new questions than I had answered. Moreover, he pointed out that science is a social process and that others may want to join my attempt to falsify the theory—but this requires that it is published first. He was, of course, correct.

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

between U.S. cities and comparisons between German cities. These item sets were constructed as follows: In the first phase, the largest 75 U.S. and the largest 75 West German cities (before Germany's unification) were determined. Second, a random set of 39 cities was selected, and ranked according to population size. Third, two ranks were randomly determined and this pair of ranks constituted both the first pair of German cities and the first pair of U.S. cities. This procedure of randomly combining German and U.S. cities simultaneously was repeated until 100 comparisons among German cities and 100 comparisons among U.S. cities (with matched ranks) were determined, with the constraint that no pair appeared twice in the item set. Participants worked on both item sets, with order counterbalanced between-participants.

2.2.2 Results

The two item sets had almost the same difficulty (percentage correct for the German cities: 75.7% and for the U.S. cities: 76.0%). Mean confidence was higher for the German cities (79.5% vs. 72.3%), and thus participants were slightly overconfident for the German cities (3.8%), and slightly underconfident for the U.S. cities (-3.7%). A participant-specific analysis revealed the same tendency: For the German (U.S.) cities, 39 (22) participants were overconfident and 17 (44) were underconfident. A comparison between item sets within participants showed that 22 participants achieved a higher percentage of correct answers for the German cities, 29 participants achieved a higher percentage for the U.S. cities, and for the remaining 5 participants these percentages were the same. In contrast, for 51 participants, their mean confidence was higher for the German cities, for 4 participants it was higher for the U.S. cities, and for the remaining 1 participant there was a tie. Moreover, for 48 participants the overconfidence score (mean confidence minus percentage correct) was higher for the German cities and for 8 participants it was higher for the U.S. cities (no ties).

2.2.3 Discussion

We expected that Germans would perform much better on the German city comparisons than on the U.S. city comparisons. Therefore, the main finding that item difficulty was practically the same for the two sets came as a complete surprise to us, which gave rise to two questions. First, how else could the original intention, namely to unconfound sampling procedure and item difficulty be achieved? And, second, how could the striking result of Experiment 1 be explained? I continue this report with the experiment that addressed the first of these questions.

2.3 A successful attempt to unconfound sampling procedure and item difficulty (Experiment 2)

This study was a second attempt to unconfound sampling procedure and item difficulty (Hoffrage, 1995, Exp. 5). In Experiment 1, I tried to achieve this by using two different item sets (German vs. U.S. cities), each consisting of comparisons that had to be made with respect to the same criterion (number of inhabitants). In Experiment 2, in contrast, I used only one reference class—famous people—but two different criteria: Age at time of death ("Who lived to be older?"), and time of birth ("Who was born earlier?"). It was expected that the age questions were relatively hard (think of Plato vs. Albert Einstein) and that the birth questions were much easier (again, think of Plato vs. Einstein).

2.3.1 Method

Participants were 100 students from the University of Salzburg (31 male, 69 female). Comparisons were generated from a list of 286 famous names (for details, see Hoffrage, 1995). The criterion (age vs. birth questions) was manipulated within-participants, each of the two item sets consisted of 100 comparisons, and order was counterbalanced.[3]

2.3.2 Results and discussion

As expected, the age questions were much harder than the birth questions (percentage correct = 57.1 and 73.5, respectively, t_{99}=21.3, p<.001). Mean confidence was much lower for the age questions (62.3% compared to 76.8% for the birth questions, t_{99}=17.7, p<.001). Participants were slightly overconfident, both for the age and the birth questions (5.2% and 3.3%, respectively). Even though this difference of 1.9 percentage points was statistically significant (t_{99}=1.99, p=.049), it was obtained with 100 participants in a within-participants design, and cannot be considered as substantial, in fact, the effect size of question type was small to medium (d=.4).

[3] Three more variables were manipulated between-subjects. First, half of the participants experienced the sampling procedure (they drew the names in the pairs from an urn themselves), while the other half were not told how the pairs were generated. Second, 80 participants were asked to specify the subjective probability that they had answered the last comparison correctly, while the other 20 were asked to specify the probability that their answer was wrong. Third, after every 10 items, 70 participants were asked "How many of the last 10 comparisons do you think you answered correctly?", while the other 30 were asked to estimate the number of items they got wrong. Moreover, at the end of every item set, participants estimated the total number of correct (or wrong) answers (out of 100 age/birth questions). While Experiment 2 was included in this paper because it is a direct consequence from Experiment 1, all variables listed in the present footnote were omitted for space reasons (for more information, see Hoffrage, 1995, or contact the author).

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

This time the attempt to generate two item sets through the same sampling procedure that differed with respect to percentage correct was successful. Even though there was slightly more overconfidence for the harder set (5.2%) than for the easier set (3.3%), the absolute difference was miniscule compared to the numbers in Lichtenstein and Fischhoff's (1977) paper on the hard-easy effect. Moreover, the hard set had a percentage correct of 57.1% (compared to 73.5% for the easy set), which suggests that for the harder set, scale-end effects (Juslin, Wennerholm, & Olsson, 1999) and unsystematic error (Erev, Wallsten, & Budescu, 1994; Juslin & Olsson, 1997; Juslin, Olsson, & Björkman, 1997) contributed more to overconfidence than was the case for the easier set.

3 First empirical test of the recognition heuristic (Experiment 3)

Soon after the data of Experiment 1 were analyzed, we moved to the University of Salzburg. When we told our new colleagues about this puzzling result, one of them, Anton Kühberger, just repeated what we said, namely that "the participants had not even heard of many of the American cities" (see also the introduction of Gigerenzer and Goldstein, 2011). He then turned our own words into an explanation that we ourselves had not seen as such and that has, since then, been referred to as the recognition heuristic. He pointed out that this partial lack of knowledge was not an obstacle but something that the German students could exploit. Goldstein and Gigerenzer (2002) later formulated the recognition heuristic as follows: "If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion" (p. 76).

The data from the Experiment 1 could not be used to test this post-hoc recognition explanation and so the following study was designed to determine whether people used the recognition heuristic when making inferences about city populations (Hoffrage, 1995; see also Schmuck, 1993). We first determined which cities a participant recognized and then manipulated how often the recognition heuristic could be used. In addition, we manipulated the size of the reference class to test whether, as explained below, this affected the participants' confidence judgments.

3.1 Method

3.1.1 Participants

Participants were 60 students from the University of Salzburg, Austria (30 male, 30 female).

3.1.2 Design and materials

For each of the 100 pairs of U.S. cities that the participants saw, they had to select the city with more inhabitants and then give a confidence judgment. For half of the participants, the cities were taken from the set of all cities with more than 200,000 inhabitants, and for the rest, from all cities with more than 400,000 inhabitants. This factor is henceforth referred to as the size of the reference class, having a value of either 75 or 32 cities, respectively. It is important to note that the *size* refers to the number of objects with a criterion value that is higher than a specific threshold. Comparing the performance for a reference class of 75 cities, randomly drawn from the 100 largest cities, to the performance for a reference class of 32 cities, randomly drawn from the 100 largest cities, would *not* be instrumental to test the predictions concerning size of reference class laid out below. The second factor that was manipulated between-subjects was how often participants' knowledge discriminated between the cities, henceforth referred to as the discrimination rate, with the levels of high, low, and uncontrolled. These two factors were fully crossed and 10 participants were randomly assigned to each of the resulting six conditions.

3.1.3 Procedure

The 40 participants who were assigned to either the high or low discrimination-rate conditions were informed that they were now "presented with a list of some American cities". These were either the largest 32 or the largest 75 cities (manipulated between-participants, see above). The cities appeared in alphabetic order, and participants were asked to indicate, for each city, whether they (1) "know something about the city, that is, know more than just the name" (henceforth referred to as K, for more Knowledge), (2) "have heard the name of the city, but have no knowledge beyond that" (R, for Recognized name), and (3) "know nothing about the city and have not even heard its name" (U, for Unrecognized). These categorizations made it possible to generate six different types of pairs. As can be seen in Table 1, the number of pairs of a particular type differed between the discrimination-rate conditions. Specifically, for participants in the high discrimination-rate condition, the cities were combined such that the recognition heuristic could be used in 55 of the 100 comparisons (25 K-U and 30 R-U comparisons). In addition, there were 5 K-K comparisons and 30 K-R comparisons for which eventually some other knowledge could allow for an inference. Thus, depending on what this other knowledge was, there was a minimum of 55 and a maximum of 90 comparisons for which either recognition or other knowledge discriminated. In contrast, for the low discrimination-rate condition, the minimum was 25 (5 K-U + 20 R-U) and the maximum was 50 (all except 30 R-R and 20 U-U).

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

Table 1: Types of paired comparisons and their frequencies. Frequencies in the high and low discrimination-rate conditions were manipulated, and those of the uncontrolled discrimination-rate condition were empirically observed. "Uncontrolled" refers to the distribution of paired comparison types for the 20 participants assigned to this condition, and Uncontrolled-32 and Uncontrolled-75 refers to the distributions that resulted for the two sizes of the reference class. K, R, and U denote cities, for which more *K*nowledge was available, which were merely *R*ecognized, and which were *U*nrecognized, respectively.

Discrimination rate	Homogeneous types			Heterogeneous types		
	K-K	R-R	U-U	K-U	K-R	R-U
High	5	5	5	25	30	30
Low	5	30	20	5	20	20
Uncontrolled	15.0	22.6	4.0	11.7	31.5	15.4
Uncontrolled-32	15.7	30.0	0.7	6.3	35.9	11.4
Uncontrolled-75	14.2	15.1	7.2	17.0	27.1	19.4

After a participant finished his or her recognition judgments, the computer program generated comparisons by randomly selecting two cities. The frequency distribution for the possible comparison types depended on the condition, as displayed in Table 1. (If, for a given participant, this requirement could not be met, the software stopped and this participant was excluded from the experiment). Another constraint was that no pair was presented twice. Then, the 100 comparisons were randomly ordered and participants chose, for each pair, the city with more inhabitants and indicated their subjective confidence in the correctness of their choice.

The 20 participants who were assigned to the uncontrolled discrimination-rate condition started with 100 paired comparisons that were generated with the only constraint that no pair was presented twice. For these participants, recognition judgments were elicited after the comparison phase.

Finally, participants estimated several relative frequencies. First, for each of the three heterogeneous comparisons, they estimated the accuracy of inferences made for these comparison types. For the R-U comparisons, for instance, the instructions read "For all possible comparisons among cities for which you recognized one (but have no more knowledge about it), and have not heard of the other, what do you think is the percentage of comparisons for which the cities you recognized is the larger one?" Second, they estimated their own percentage of correct inferences for each of the six comparison types.

3.2 Predictions

3.2.1 Discrimination rate

Gigerenzer and Goldstein (1996) extended Gigerenzer et al.'s (1991) PMM algorithm by adding the recognition heuristic, while preserving the principle of one-reason decision making. This results in the following possible situations. When the recognition heuristic discriminates between two objects (cities in this case), then the city that is recognized is chosen as the larger one and the recognition validity is given as the confidence level. When both cities are recognized and something is known about at least one of the cities, the most valid cue is used to make a choice and the confidence is determined by the validity of this cue. When both cities are recognized but no further knowledge is available, or when neither city is recognized, a city is chosen randomly and confidence is 50%. If cue validities and recognition validity are estimated without any bias, confidence judgments should be well calibrated and mean percentage correct should equal mean confidence. This is because, within each of the six comparison types, the city pairs were selected randomly so that the validity of the cue and the recognition validity in the sample used in the experiment were expected to be identical to those in the reference class. The factor of discrimination rate should thus affect only mean percentage correct (higher in the high discrimination-rate condition) and mean confidence (again, higher in the high discrimination-rate condition), but it should not affect overconfidence (the difference between confidence and percentage correct).

3.2.2 Size of reference class

The size of the reference class was neither an issue in the original PMM paper, nor in Gigerenzer and Goldstein (1996), nor in Goldstein and Gigerenzer (1999). It was simply assumed that people are well adapted to their natural environments, and that they are able to estimate cue validities with a reasonable degree of accuracy. However, as Hoffrage (1995) and Hoffrage and Hertwig (2006)

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

showed, cue validities can depend on the size of the reference class (see also Goldstein & Gigerenzer, 2002, Figure 5). Gigerenzer et al. (1991) used all German cities with more than 100,000 inhabitants (as of 1988). Although 100,000 is a salient number, other thresholds might have been used. Indeed, the cue validities in this environment depend on this threshold, that is, on the minimum number of inhabitants a city must have to be included in the set. Across four different thresholds, cue validities varied widely: For one of the twelve cues, the validity dropped from 77% to 0%; for the others, the average absolute difference between the validities among all cities with more than 100,000 inhabitants and those among all cities with more than 300,000 was 10.3%.

In a similar vein, Juslin, Olsson and Winman (1998) showed that the percentage of correct inference depends on how items are sampled from a reference class. These authors varied whether pairs were constructed by randomly drawing each of the two objects from the whole reference class or whether sampling was constrained such that one (the other) object was randomly drawn from the set of those objects with a criterion value above (below) the median. The commonality between their constrained sampling procedure and my larger reference class is that for both conditions the differences between ranks (of objects with respect to the criterion value) are, on average, larger compared to the corresponding rank differences in the unconstrained procedure and the smaller reference class, respectively. Juslin et al. (1998) found, both with simulated and with participants' data, that cue validities and percentage correct, respectively, were positively related to averaged rank size differences.

Based on Juslin et al.'s findings and on my own calculations just reported, one would predict that the percentage of correct inferences would be higher for the larger reference class. Given PMM Theory's assumption that cue validities drive not only percentage correct but also confidence, one should predict that mean confidence would be higher for the larger reference class as well. However, based on the assumption that participants are not aware of the relationship between (recognition and other cues') validities and the size of the reference class—note that not even the authors of PMM theory saw this when they published their paper (Gigerenzer et al., 1991)—I predicted that the mean confidence would *not* differ between the two reference class conditions. Confidence could even be higher for the smaller reference class. This is because, for this reference class, participants will presumably recognize a higher proportion of cities and will know more about a higher proportion of cities, compared to the larger reference class condition. This overall impression of higher familiarity with the cities may translate into higher confidence judgments.[4] Given the data of Experiment 1, which used U.S. cities above 200,000 inhabitants, I expected that mean confidence would match percentage correct and that there would thus be no overconfidence for the larger reference class (largest 75 cities). In contract, I predicted overconfidence for the smaller reference class (largest 32 cities)—certainly because percentage correct would be lower than for the larger reference class and maybe, in addition, because mean confidence would be higher than for the larger reference class.

Note that the combination of the uncontrolled discrimination-rate condition with the two reference class conditions is most likely to yield evidence conflicting with PMM theory's prediction that overconfidence disappears if pairs are randomly sampled from a defined reference class. (This prediction holds only for confidence ratings, not for frequency estimates.) For the controlled discrimination-rate condidions, analysing participants' overconfidence is less crucial for a test of PMM theory, as sampling of pairs is constrained in these conditions.

3.3 Results

This section proceeds as follows. First, I report the effects of the two main factors, discrimination rate and size of the reference class, on percentage correct, mean confidence, and overconfidence. Second, I show that these effects were exclusively driven by the relative frequencies of the six comparison types. Third, I ask how often participants followed the recognition heuristic when selecting a city. Fourth, I compare percentage correct and mean confidence to estimated validities and estimated percentages of correct choices. Finally, I demonstrate that recognition judgments depended on the size of the reference class.

3.3.1 Main effects of discrimination rate and size of the reference class

Figure 1 displays the six calibration curves for the six conditions that result from crossing the two main factors: discrimination rate and size of the reference class. Table 2 displays the mean confidence (MC), percentage correct (PC), and overconfidence (OC = MC-PC), again, across all participants and items. It can be seen that the effect of discrimination rate was small compared to that of the size of reference class. This is consistent with the results of three ANOVAs, each computed with the participant-specific values for MC, PC, and OC (Table 2, lower rows). As predicted, for each of the corresponding discrimination-rate conditions (e.g., high), PC was higher for the larger reference class (e.g., 69.4) than for the smaller one (e.g., 65.5), and MC was higher for the

[4] See Schooler & Hertwig, 2005, for another, more fine-grained, way to discriminate among cities beyond the dichotomous concept of recognition.

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

Figure 1: Calibration curves for each combination of size of the reference class (largest 75 cities, largest 32 cities) and discrimination rate (high, low, uncontrolled).

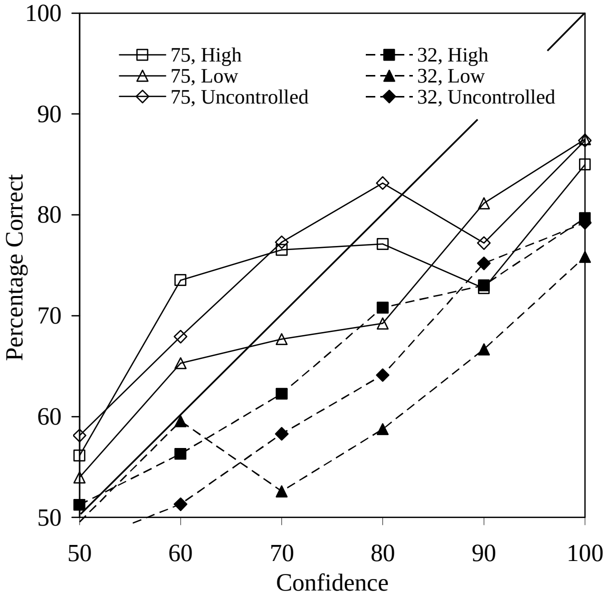

Table 2: Mean confidence, percentage correct, and overconfidence for the six conditions of Experiment 2 that resulted from combining size of reference class (1st level) with discrimination rate (2nd level).

	Mean confidence	Percentage correct	Over-confidence
Large reference class (largest 75 cities)			
High	67.4	69.4	−2.0
Low	64.2	63.9	0.3
Uncontrolled	70.7	72.3	−1.6
Small reference class (largest 32 cities)			
High	74.9	65.5	9.4
Low	68.1	58.2	9.9
Uncontrolled	73.6	61.5	12.1
Effect of size of reference class			
F(1,54)	4.29	22.6	26.4
p	.019	<.001	<.001
Effect of discrimination rate			
F(2,54)	3.56	8.21	0.19
p	.036	.001	.832

smaller reference class that contained relatively more familiar cities. Further, as predicted, OC differed dramatically between the reference classes. For the 75 cities with more than 200,000 inhabitants, it basically disappeared, replicating Experiment 1 which used the same reference class. For the reference class of the largest 32 cities (each city more than 400,000 inhabitants), however, massive overconfidence was observed. The interaction between size of reference class and discrimination rate was not statistically significant for any of the three dependent variables (not shown in Table 2).

3.3.2 Effect of discrimination rate and size of reference class within comparison type

The values for PC in Table 2 should differ between the discrimination-rate conditions because the validities of the recognition heuristic and that of other cues should be different for the six comparison types, and because the relative frequencies of these types were different for the three levels of discrimination rate. For a given comparison type, however, the PC should *not* depend on the discrimination-rate condition. To test for this independence, for each of the 60 participants, MC, PC and OC were computed within each of the six comparison types. Based on the resulting 60*6=360 entries, three ANOVAs were conducted, one for each of the three dependent variables (these ANOVAs had only 349 degrees of freedom because for some participants of the uncontrolled discrimination-rate condition there were no entries for some comparison types). Unlike the previous analyses which were computed based on averaging across all 100 items, when comparison type was held constant, that is, statistically controlled for, the discrimination rate no longer had an effect, MC: $F(2,349)=1.32$, $p=0.27$, PC: $F(2,349)=0.16$, $p=0.85$, and OC: $F(2,349)=0.19$, $p=0.83$. In contrast, the differences between the reference class conditions were significant: MC: $F(1,349)=12.70$, $p<0.001$, PC: $F(1,349)=8.60$, $p=0.004$, and OC: $F(1,349)=32.37$, $p<0.001$. Because discrimination rate had, as expected, no significant effect within a given comparison type (and its effect on MC, PC, and OC across all 100 items was due only to different frequencies of the different comparison types), the subsequent analyses focus on comparison types, thereby aggregating across participants of the different discrimination-rate conditions.

3.3.3 Effects of (recognition) knowledge on decisions

The results reported above establish that the frequency of comparison types drove the percentage correct: The more often the recognition heuristic could be used, the better participants' performance was. Even though this finding already suggests that participants tended to infer that recognized cities are larger than unrecognized cities, there is also a more direct way to see whether this was the case. Table 3 displays—for each of the six comparison types

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

Table 3: Mean confidence (MC), percentage correct (PC) and overconfidence (OC) for the six comparison types. For the three heterogeneous types, cases were divided into those for which a participant's decision matched the decision of the recognition heuristic ("consistent") or not ("inconsistent"), or favored a K-city over an R-city ("consistent") or not ("inconsistent").

Comparison type	Consistency with prediction	N	%	MC	PC	OC
U-U		579		55.9	56.5	−0.6
R-R		1151		65.7	56.5	9.2
K-K		499		80.2	72.1	8.0
K-U	consistent	759	91.1	79.9	83.6	−3.7
K-U	inconsistent	74	8.9	61.4	21.2	40.2
K-U		833	100	78.4	78.6	−0.2
K-R	consistent	1293	79.3	78.6	74.9	3.8
K-R	inconsistent	337	20.7	64.5	48.5	16.0
K-R		1630	100	75.7	69.4	6.3
R-U	consistent	1045	79.9	65.0	65.3	−0.3
R-U	inconsistent	263	20.1	54.9	39.2	15.8
R-U		1308	100	63.0	60.0	3.0

and across all participants and items—mean confidence, percentage correct and overconfidence. For the three heterogeneous comparison types an additional analysis was performed, based on the knowledge about the two cities and how a participant responded. Specifically, decisions that were consistent with the assumption that the recognition heuristic was used (referred to as "consistent") include those where a city that was recognized (be it with or without more knowledge, that is, a K-city or an R-city) was selected when it was paired with an unrecognized city. Finally, a decision in favor of a K-city when paired with an R-city was also classified as "consistent". Note that for these cases, recognition did not discriminate, so this classification was based on the assumption that more knowledge about one city is most likely to be knowledge that allows for the inference that it is larger than a city for which such knowledge does not exist.

Across all cases in which a recognized city (either K or R) was paired with an unrecognized city (U), participants decided in favor of the recognized city in 84.3% of the cases. An analysis conducted on an individual basis revealed that 5 participants decided in favor of the recognized city in 100% of the critical cases, 11 in 99.9 - 90% of the cases, 29 in 89.9 - 80%, 10 in 79.9 - 70%, 3 in 69.9 - 60%, 1 in 48% (this participant had a percentage correct of 51%, suggesting that he responded randomly throughout), and for 1 participant, the adherence rate could not be computed (as she recognized all the cities in the reference class). When participants recognized both cities but knew something about one city (K) but not the other (R), they favored the city that they knew something about in 79.3% of the cases. It is interesting to see that such "consistent" decisions were far more likely to be correct than the "inconsistent" decisions. Had participants always decided in favor of the recognized city (or, for K-R pairs, in favor of the K city), the percentage correct for the K-U, K-R, and R-U comparisons would have been 83.1%, 70.0%, and 64.4%, instead of 78.6%, 69.4%, and 60.0%, respectively. It is also interesting to see that mean confidence was lower for the "inconsistent" decisions than for the "consistent" decisions. This reduction, however, was not sufficient to compensate for the lower percentage correct, and so overconfidence was far more pronounced for the "inconsistent" than for the "consistent" decisions.

3.3.4 Effects of (recognition) knowledge on validities

How do participants' estimates of the validities for the various comparison types relate to the corresponding percentages of correct decisions? Before answering this question, I extend the list of variables by adding what I refer to here as simulated validities, that is, the percentages of correct inferences for all possible comparisons of cities within a given type of comparison (K-U, K-R, and R-U), given that a participant always decided in favor of the first city (K, K, and R, respectively). These variables obviously had to be computed separately for each participant. Table 4 contains the values of the six variables, averaged across all 60 participants. The consistency of the pattern revealed in Table 4 is striking. Within each of the three heterogeneous comparison types, both the

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

Table 4: Estimated and simulated validity, mean confidence, percentage correct, estimated percentage correct, and overconfidence for the three heterogeneous comparison types, separated according to size of reference class.

		Estimated validity	Simulated validity	Mean confidence	Percentage correct	Estimated PC	Overconfidence
K-U	Largest 75	70.5	87.5	75.1	80.0	64.6	−4.8
	Largest 32	75.6	78.4	85.4	68.5	68.7	16.8
K-R	Largest 75	59.8	74.5	73.5	71.5	60.8	2.0
	Largest 32	68.5	68.9	78.2	67.2	64.6	10.9
R-U	Largest 75	61.4	70.1	60.9	65.0	50.5	−4.1
	Largest 32	63.8	60.6	65.8	55.9	55.4	9.9

simulated validity and percentage correct are higher for the large reference class (75 cities) than for the small one (32 cities). Across all these comparison types and across these two variables (simulated validity and PC), the average for the 75 largest cities exceeds that for the largest 32 cities by 8.2 percentage points. Interestingly, participants were obviously not aware of this relationship. To the contrary, their estimated validities, their mean confidence, and their estimated percentage correct all pointed in the opposite direction: Each of these values was larger for the reduced reference class. Averaged across all comparison types and all these three variables, the difference was −5.4 percentage points.

It is also interesting to see that in each of the six rows in Table 4, the simulated validity exceeded percentage correct. Note that the values for these two measures would have been the same, had all participants always chosen in favor of the K-city and the R-city in cases in which such cities have been paired with an U-city, and in favor of the K-city in cases in which it has been compared with a R-city. However, as was explained above, this was not the case and so the findings reported in Table 4 mirror those reported in Table 3.

3.3.5 Effects of the size of the reference class on recognition judgments

The last analysis of these data reported here concerns the question of whether recognition judgments were independent of the size of the reference class (for more results, see Hoffrage, 1995). According to range-frequency theory (Parducci, 1965), which posits that people have a tendency to map the range of an attribute's levels linearly onto the range of the response scale, one may suspect that this may not be the case. Specifically, having relatively few K-cities in the larger reference class or relatively few unrecognized cities in the smaller reference class may lead one to shift the criterion that is used to make these classifications. Conversely, having relatively more U-cities in the larger reference class and relatively more K-cities in the smaller reference may lead to corresponding criterion shifts in the other direction. Even though Goldstein and Gigerenzer (2002) conceptualized recognition as a simple dichotomous variable—a city is either recognized or not—others discussed the possibility that the process of making such categorical judgments may draw on some more continuous representations in memory which, in turn, open the theoretical possibility of context effects on threshold settings (Erdfelder, Küpper-Tetzel, & Mattern, 2011; Gigerenzer & Murray, 1987; Hertwig, Herzog, Schooler, & Reimer, 2008; Pleskac, 2007; Schooler & Hertwig, 2005).

In fact, recognition judgments in this experiment did depend on the size of the reference class. The most straightforward way to see this is to compare the recognition judgments of the largest 32 cities to those of the same cities, but now as a subset that is embedded in the set of the largest 75 cities (henceforth referred to as 32-in-75). Without any context effects, the recognition judgments should not differ between the two sets (that contain, after all, exactly the same cities). Table 5 displays the absolute and relative frequencies of the three knowledge states, depending on the size of the reference class. It is interesting to see that in the population of the largest 32 and of the largest 75 cities, virtually the same percentage of cities was categorized as "more knowledge beyond name recognition", 34.0 and 32.8, respectively. As a necessary consequence, for the 32-in-75 cities, this percentage increased dramatically, from about a third, to more than half. The criterion shift was expected to be in the other direction for the unrecognized cities, and this was the case; the percentage of U-cities decreased by almost a factor of two and fell from about 20% (32 largest cities) to about 10% (32-in-75).

As a necessary consequence, the frequency distribution of the type of comparisons was quite different for the set of the largest 32 cities and the 32-in-75 set (Table 6).

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

Table 5: Frequency distributions of recognition judgments as a function of the size of the reference class. 32-in-75 refers to the set of the largest 32 cities when they were presented to the participants embedded in the set of the largest 75 cities, but were later analyzed separately.

Reference class	Frequencies			Relative frequencies		
	K	R	U	K	R	U
75	739	839	672	32.8	37.3	29.9
32-in-75	521	341	98	54.3	35.5	10.2
32	326	448	186	34.0	46.7	19.4

Table 6: Relative frequency distribution of the six city-comparison types, depending on the size of the reference class (for an explanation of 32-in-75, see Table 5).

Reference class	Homogeneous types			Heterogeneous types		
	K-K	R-R	U-U	K-U	K-R	R-U
Largest 75	8.1	16.7	10.7	15.7	25.7	23.1
32-in-75	17.3	19.7	1.6	9.1	43.8	8.4
Largest 32	8.6	21.7	8.6	12.1	28.6	20.5

This leads to the interesting question of whether these context effects on the recognition judgments affected the confidence, percentage correct, or overconfidence. The following rationale makes it clear why this may be the case. We have seen that the set of the largest 32 cities when presented alone, compared to analyzing the 32-in-75 set, led to a stricter criterion for a city to be classified as a K-city, and to a more liberal criterion to classify a city as a U-city (see also Table 6). Moreover, we have seen that the validities (be it of the recognition heuristic or, by an extension of the argument, those of other cues) were higher for the larger reference class than for the smaller one. In fact, the percentage of correct inferences for the set of the largest 75 cities, the embedded set (32-in-75) and the set of the largest 32 cities were 68.5%, 67.2%, and 61.7%, respectively.

3.4 Discussion

All of the predictions were basically confirmed. The discrimination rate affected percentage correct, mean confidence, and overconfidence as predicted: The more often the recognition heuristic could be applied and the more often other knowledge discriminated among the cities, the higher the percentage correct and mean confidence were. These effects could be fully accounted for by the relative frequencies of the six comparison types. The percentage of participants' choices that were consistent with the prediction of the recognition heuristic is in the same range as reported in other studies that were conducted later (e.g., Goldstein & Gigerenzer, 2002). What the present study adds to the literature is the observation that, for a larger reference class (all cities above 200,000 inhabitants) as compared to a smaller reference class (all cities above 400,000 inhabitants), percentage correct was higher, mean confidence was lower, and overconfidence was less pronounced. To the best of my knowledge, such effects have not been reported elsewhere. Equally important is the related finding that participants were not only unaware of the dependency of the validity of the recognition and other knowledge on reference class size, but also that their answers even pointed in the opposite direction (higher confidence judgments and frequency estimates for the smaller reference class). Finally, what the present study adds to the literature is the conjecture that recognition judgments might best be seen as resulting from mapping an underlying hypothetical variable with the help of a response function onto a dichotomous recognition value. Such a view could, at least, easily account for the fact that the observed recognition judgments depended, in a between-participants comparison, on the size of the reference class.

4 General discussion

The present paper reported three studies. The first, paving the way to the recognition heuristic, was a failed attempt to generate hard questions by asking German students which of two U.S. cities (each randomly drawn from a defined reference class) has more inhabitants. To our surprise, German students were about as good at these

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

questions as they were at the corresponding comparisons among German cities. In Experiment 2, a similar attempt succeeded: When comparing two representative item sets, one hard and the other easy, the hard-easy effect was still observed (higher overconfidence for the hard set), but now the effect was much smaller than in previous studies. These two data points fit perfectly into the larger picture that Juslin, Winman, and Olsson (2000) provided in their meta-analysis in which they analyzed the effect of sampling procedure. Specifically, those authors conducted a review of 95 independent data sets with selected items and 35 sets in which items had been sampled representatively. Across all selected item sets, overconfidence was 10%, and across all representative sets it was 1% (95% confidence intervals for each set were at ±1%). Juslin et al. pointed out that this difference could not be explained by differences in percentage correct. Moreover, when they controlled for the end effects of the confidence scale and the linear dependence between percentage correct and the overconfidence score (recall that OC=MC-PC), the hard-easy effect virtually disappeared for the representative item sets.

4.1 The recognition heuristic: Compensatory or non-compensatory?

The focus of the present paper was on the recognition heuristic, which was proposed as a post-hoc explanation for the puzzling result of Experiment 1. Two of the major results of Experiment 3 were, first, that people's choices were consistent with the recognition heuristic in about 80% of the pairs when they had no additional knowledge about the recognized city (and in about 90% when there was such knowledge), and, second, that discrimination rates drive percentage correct, mean confidence and overconfidence. As of today, almost 20 years after this study was conducted, readers might say "we knew that all along", and rightly so, as many similar findings have been reported since then (for overviews see Gigerenzer & Goldstein, 2011; Pachur, Todd, Gigerenzer, Schooler, & Goldstein, in press). However, the literature on the recognition heuristic also reveals some controversies. Some papers (e.g., Bröder, 2000; Chater, Oaksford, & Nakisa, 2003; Dougherty, Franco-Watkins, & Thomas, 2008) criticize some aspects and raise some doubts concerning the research program in which the recognition heuristic is embedded, namely the simple heuristics program initiated by the ABC Research Group (Gigerenzer, Todd, and the ABC Research Group, 1999) in general. Space and the focus of this special issue do not allow such criticism to be addressed here (but see, e.g., Todd, Gigerenzer, and the ABC Research Group, 2000; Gigerenzer, Hoffrage, & Goldstein, 2008).

Among those criticisms that refer to the recognition heuristic specifically, one is particularly interesting as it directly relates to a distinction already made in the present Experiment 3. Several authors (e.g., Bröder & Eichler, 2006; Hilbig & Pohl, 2008; 2009; Newell & Fernandez, 2006; Newell & Shanks, 2004; Oppenheimer, 2003; Pachur, Bröder, & Marewski, 2008; Pohl, 2006; Richter & Späth, 2006) have challenged Goldstein and Gigerenzer's (2002) claim that people use recognition knowledge in a non-compensatory fashion. Most of the studies reported by those authors distinguished between objects that participants recognized but for which they had no additional knowledge (in the present paper referred to as R-objects) and objects which they recognized and for which they had further knowledge (K-objects). Hilbig and Pohl (2008), for instance, referred to these objects as mR (for mere recognition) and R+ (for recognition plus knowledge), respectively. Some of these authors then developed and used measures beyond those used in the analyses reported above, like various parameters in a multinomial model approach (Hilbig, Erdfelder, & Pohl, 2010), response times (Hilbig & Pohl, 2009), or the DI (Discrimination Index; Hilbig & Pohl, 2008); for an overview, see Hilbig (2010). The overall conclusion of these authors is that their data conflict with the hypothesis that recognition knowledge is always used in a non-compensatory way.

Some of these authors would probably also interpret some of the results reported in the present paper as inconsistent with the non-compensatory nature of the recognition heuristic. For instance, the finding in Experiment 2 that percentage correct is substantially larger for K-U pairs than for R-U pairs is consistent with the assumption that the knowledge that was available for K-cities has been used in some way. Another example would be the DI (Hilbig & Pohl, 2008), which is defined as the adherence rate to the recognition heuristic among paired comparisons in which the recognized object was the correct answer minus the adherence rate among those comparisons for which the recognized object was the incorrect answer. In their studies, Hilbig and Pohl found this index to be positive and concluded that the recognition heuristic is not used in a non-compensatory way. The rationale for this conclusion is that a positive index "would not be possible through following the recognition cue alone" (Hilbig & Pohl, 2008, p. 395)—simply because following the recognition cue alone yields adherence rates of 100%, both for cases in which the recognition heuristic would lead to a correct and an incorrect inference, which, in turn, would yield a difference of zero.[5]

The DI for Experiment 3 can be recovered from the information displayed in Table 3. Across all participants

[5] Note that even if the adherence rate is below 100%, the DI would still be zero as long as the percentage of choosing the recognized object is independent of whether the recognized alternative is correct or not.

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

and items, it was .055 (the average of the participant-specific DIs was .053, with SD = .184, SE = .024, which was significantly greater than 0, t_{58} = 2.22, p = .02, and it was positive, zero, negative, and not defined for 30, 6, 23, and 1 participants, respectively). Among R-U comparisons, DI = .031 (the average across participant-specific DIs was .041, SD = .242, SE = .032, t_{58} = 1.3, p = .10, with 25, 7, 27, and 1 participants who had a positive, zero, negative, and undefined score, respectively) and among K-U comparisons, DI = .025 (the average across participant-specific DIs was .032, SD = .277, SE = .041, t_{45} = .77, p = .22, with 9, 21, 16, and 14 participants, respectively). Even though the DI in Experiment 3 was positive, it was lower than for other studies reported in the literature (e.g., Hilbig & Pohl, 2008), and the difference from zero was only significant when R-U and K-U comparisons were pooled (but for none of these comparison types separately). Moreover, DI_{K-U} did not exceed DI_{R-U}. To the extent that a positive DI reflects the use of knowledge beyond recognition, one should have expected to see that $DI_{K-U} > DI_{R-U}$, because for K-U comparisons more knowledge can be used than for R-U comparisons.

Some findings of Experiment 3 are in line with those reported by authors who have challenged the non-compensatory nature of the recognition heuristic. Not only the DI (which includes adherence rates conditioned on whether the recognized object is the correct answer), but also Table 3 (which reported percentage correct conditioned on adherence) can be interpreted as evidence inconsistent with the claim that recognition is *always* used in a non-compensatory way. I want to emphasize that I, just like Gigerenzer and Goldstein (2011, p. 110), "have no doubts that recognition is sometimes dealt with in a compensatory way". In fact, if a participant happens to know that a city she recognizes is very small and recognized for reasons other than population size (think of Chernobyl or Fatima), then this would constitute a good reason *not* to make an inference based on the recognition heuristic, but to decide based on what Gigerenzer et al. (1991) called a local mental model, that is, to use direct knowledge about the criterion. A simple example can demonstrate that very few cases (5 in 2,000 pairs) like this are already enough to make a difference between the DI that was observed in Experiment 3 and a DI of zero.[6]

That recognition knowledge is trumped by criterion knowledge is one reason why choices may not be consistent with the recognition heuristic. Another reason is that recognition knowledge could be trumped by probabilistic cues (see also Gigerenzer & Goldstein, 2010). Experiment 3 of the present paper did not live up to Gigerenzer and Goldstein's request to specify models for such compensatory use of cues against which the non-compensatory recognition heuristic is tested. One should not forget, however, that this was the first, exploratory study on the recognition heuristic, conducted almost 20 years ago, whose goal was to test the post-hoc explanation developed after Experiment 1, rather than to test specific claims that were formulated only several years later. While Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, G. (2010), who conducted such a rigorous test, conclude from their studies that the recognition heuristic outperformed all competing compensatory models with respect to predicting people's inferences, Experiment 3 of the present paper did not elicit the data that are necessary to perform such tests.

The recognition heuristic is a *model* of cognitive processes involved in inferences, and, as every model does, it simplifies. Therefore, I do not find it at all surprising to see that people seem to follow the recognition heuristic in less than 100% of the cases in which it allows for an inference (as reflected in adherence rates < 1) and even less so if an inference would be incorrect (as reflected in DI > 0). What I do find surprising, though, is that this "failure" to make correct predictions in 100% of the cases is sometimes seen as critical evidence. This attitude strikes me as even more surprising when considering that there is no scarcity of authors in cognitive psychology who seem to be satisfied if their model predicts outcomes significantly better than chance.

4.2 The theoretical importance of the (size of the) reference class

Experiment 3 revealed effects that have not been reported elsewhere. It is easy to understand why increasing the size of the reference class increases both the recognition validity and the validities of cues: adding smaller cities to a set of larger cities is more likely to result in adding unrecognized cities than recognized cities, and it is more likely to result in adding cities with unknown or negative

[6] Consider the 20 participants of Experiment 3 for whom the comparisons were randomly sampled without any constraints. Across all 20*100 pairs, the recognition heuristic made an inference in 541 pairs, a correct one in 412 and an incorrect in 129, and choices were consistent with the recognition heuristic in 367 of 412 (= .891) cases and 110 of 129 (= .853) cases, respectively. Thus, the DI of .038 (= .891−.853) for these 20 participants was almost the same as the one computed across all participants, including those for whom the relative frequencies of the comparison types have been controlled for. Note that a (hypothetical) adherence rate of .891 (= 115 of 129) among those pairs in which the recognition heuristic leads to an incorrect inference would have resulted in DI = 0. However, participants did not choose consistently with the recognition heuristic in 115 but only in 110 of these 129 cases. As the Chernobyl example illustrated, there may be good reasons not to use the recognition heuristic for some of the pairs. Moreover, to the extent that the local mental models that participants can construct are more likely to reduce the adherence rate among pairs for which it would be smart *not* to use the recognition heuristic, a DI > 0 is not surprising. In fact, it is only rational to allow criterion knowledge to trump recognition knowledge (for tests of the potential role of criterion knowledge, see Hilbig, Pohl, & Bröder, 2009, and Pachur et al., 2008).

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

cue values than with positive cue values. This, in turn, will not only increase the proportion of pairs consisting of recognized and unrecognized cities, but also, within this set, will increase the proportion of pairs in which the recognized city is the larger one (see the simulated validities in Table 4). However, it should be mentioned that increasing the size of the reference class also increases the average difference between the population sizes of the cities that are compared (see also Juslin et al., 1998). To the extent that participants possess criterion knowledge (Hilbig, Pohl, & Bröder, 2009), the increase of percentage correct (as size of the reference class increases) could also be explained by a relative increase of comparisons that are made through the construction of local mental models as compared to probabilistic mental models (Gigerenzer et al., 1991).

In contrast, participants' mean confidence revealed an effect in the opposite direction to that which has been observed for percentage correct: confidence judgments were lower for the larger reference class and higher for the smaller one. Taken together with the effect on percentage correct, this resulted in zero over/underconfidence for the larger reference class but in severe overconfidence for the smaller reference class. Note that this result was observed in the condition in which the discrimination rate has not been controlled for and thus it poses a challenge for PMM theory (Gigerenzer et al., 1991). At the same time, the effect on confidence judgments is easily explained: It may have resulted from the fact that the smaller reference class contained relatively more cities that the participants recognized and also more cities that they knew something about, coupled with the (false) belief "the more I know, the better I will perform." It is also consistent with the results of many studies conducted by Klaus Fiedler and his colleagues demonstrating that participants do not appropriately adjust their judgments to the sampling procedure of the items they are presented with (e.g., Fiedler, 2000).

The insight that both the validities of recognition and that of other cues depend on the size of the reference class leads to some interesting questions: Which reference classes should experimenters select in their studies? Which reference classes do participants use when they determine their confidence? The problem of choosing the adequate reference class is neither trivial nor new. It is, for instance, fundamental to the frequentistic interpretation of probabilities (for history and interpretations, see Gigerenzer et al., 1989). As the great probability theorist Richard von Mises (1957) put it, "we shall not speak of probability until a collective has been defined" (p. 18). Insurance companies face the same problem when determining the premium for a life insurance of a particular person. Clearly this premium will depend on the probability that this person will die, say, within the next ten years. But which of the person's innumerable properties should be used to construct a reference population? Each of these properties (as well as combinations thereof) could be used to define the reference class, and in all likelihood, many of the resulting reference classes would yield different statistics and thus different estimations for mortality risks, leaving open the question of which is the correct one.

Frankly, I do not have a good answer. However, I think there are possible pragmatic routes toward a "good enough solution" (see also Hoffrage & Hertwig, 2006). Under some circumstances, experimenters may circumvent the problems that result from fuzzy reference classes—either by selecting one that is small, finite, and complete (e.g., all African states) or by creating microworlds (e.g., Fiedler et al., 2002). This allows them to control for participants' exposure to these worlds and make sure that the intended reference class and the participants' reference class converge. Another possibility would be to explore the boundaries of a reference class empirically (e.g., by analyzing environmental frequencies). Anderson and Schooler (1991), for instance, examined a number of environmental sources (such as the *New York Times*) and showed that there are reliable relationships between the probability that a memory for a particular piece of information will be needed and frequency, recency, and patterns of prior exposure. Such an analysis of environmental statistics could also be conducted in the context of the research reported in this paper. For the city task, for instance, it may show that people are much more likely to encounter larger cities than smaller cities. Specifically, such environmental frequencies could be used to determine how often a particular city is used in the experimental materials. Finally, another way to determine the "right" size of people's reference classes is to transfer the task of sampling experimental stimuli from the experimenter to the participants. Hogarth (2005), for instance, used mobile phones and interrupted his participants in their flow of daily activities at randomly chosen intervals and asked several questions regarding the last decision that they made, thereby letting them, the environment, and chance determine which environmental stimuli are designated to become experimental ones (see also Dhami, Hertwig, & Hoffrage, 2004).

4.3 Final remarks

The formulation of the recognition heuristic has led to a lot of exciting research. However, we should not only look at what has been achieved in the past, but also continue this fruitful tradition in the future. Interestingly, when adopting the recognition heuristic to generate recommendations for choosing among research topics, it should be inverted. *When faced with the choice between*

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

working on recognized topics, replicating known findings, versus entering new and unexplored territory: Go with the latter. I hope the present paper helped to identify some of these blank areas on the map of research on the recognition heuristic, thereby initiating some further steps towards new directions.

References

Anderson, J. R., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

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

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

Chater, N., Oaksford, M., & Nakisa, R. (2003). Fast, frugal, and rational: How rational norms explain behavior. *Organizational Behavior and Human Decision Processes, 90,* 63–86.

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

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review, 115,* 199–213.

Erdfelder, E., Küpper-Tetzel, C. E., & Mattern, S. D. (2011). Threshold models of recognition and the recognition heuristic. *Judgment and Decision Making, 6*, 7–22.

Erev, I., Wallsten, T.S., & Budescu, D.V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519–527.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107,* 659–676.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom–A cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*, 527–561.

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

Gigerenzer, G., & Goldstein, D. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making, 6*, 100–121.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychological Review*, *115*, 230–237.

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

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Switjink, Z., Porter, T., Daston, L., Beatty J., & Krüger, L. (1989). *The empire of chance*. Cambridge, UK: Cambridge University Press.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Goldstein, D. G., & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 37–48). New York: Oxford University Press.

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

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1191–1206.

Hilbig, B. E. (2010). Precise models deserve precise measures: A methodological dissection. *Judgment and Decision Making, 5,* 300–309.

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.

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

Hilbig, B. E., & Pohl, R. F. (2009). Ignorance- vs. evidence-based decision making: A decision time analysis of the recognition heuristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1296–1305.

Hilbig, B. E., Pohl, R. F., & Bröder, A. (2009). Criterion knowledge: A moderator of using the recognition heuristic? *Journal of Behavioral Decision Making, 22*, 510–523.

Hoffrage, U. (1995). *Zur Angemessenheit subjektiver Sicherheits-Urteile. Eine Exploration der Theorie der probabilistischen mentalen Modelle*. [The adequacy of subjective confidence judgments: Studies concerning the theory of probabilistic mental models]. Doctoral dissertation, University of Salzburg, Austria.

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press

Hoffrage, U. (2004). Overconfidence. In R. F. Pohl (Ed.), *Cognitive illusions: Fallacies and biases in thinking, judgement, and memory* (pp. 235–254). Hove, UK: Psychology Press.

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, NY: Cambridge University Press.

Hogarth, R. (2005). Is confidence in decisions related to feedback? Evidence from random samples of real-world behavior. In K. Fiedler, & P. Juslin (Eds.), *Sampling and adaptive cognition* (pp. 456–484). Cambridge, NY: Cambridge 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.

Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review, 104,* 344–366.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making, 10,* 189–209.

Juslin, P. Olsson, H., & Winman, A. (1998). The calibration issue: Theoretical comments on Suantak, Bolger, and Ferrell (1996). *Organizational Behavior and Human Decision Processes, 73*, 3–26.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1038–1052.

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.

Kilcher, H. (1991). Experimentelle Untersuchungen zum Overconfidence-Phänomen: Zur Bedeutung von Aufgabenzusammenstellung und zur Vorhersage von Confidence-Angaben. Diploma-thesis submitted to the University of Constance, Germany, on February 20th, 1991.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also no more about how much they know? The calibration of probability judgments. *Organizational Behavior and Human Decision Processes, 36,* 143–166.

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

McClelland, A. G. R., & Bolger, F. (1994) The calibration of subjective probabilities: Theories and models 1980-94, in G. Wright and P. Ayton (Eds.) *Subjective probability* (pp. 453–482). Chichester, England: Wiley.

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*, 333–346.

Newell, B. R., & Shanks, D. R. (2004). On the role of recognition in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 923–935.

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): Rethinking the recognition heuristic. *Cognition*, *90*, B1–B9.

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.

Pachur, T., Todd, P. M., Gigerenzer, G., Schooler, L. J., & Goldstein, D. G. (in press). Is ignorance an adaptive tool? A review of recognition heuristic research. In P. M. Todd, G. Gigerenzer, & the ABC Research Group, *Ecological rationality: Intelligence in the world.* New York: Oxford University Press.

Parducci, A. (1965). Category judgments: A range-frequency model. *Psychological Review, 72,* 407–418.

Pleskac, T. J. (2007). A signal detection analysis of the recognition heuristic. *Psychonomic Bulletin and Review, 14,* 379–391.

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

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

Schmuck, S. (1993). Die "Bekanntheit"—Cue oder Variable? Eine kritische Elaboration der Theorie der probabilistischen mentalen Modelle (PMM-Theorie). Diploma-thesis submitted to the University of Salzburg, Austria, on April 1st, 1993.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inferences. *Psychological Review, 112,* 610–628.

Todd, P. M., Gigerenzer, G., and the ABC Research Group. (2000). How can we open up the adaptive toolbox? *Behavioral & Brain Sciences, 23*, 767–780.

von Mises, R. (1957). *Probability, statistics, and truth.* New York: Dover Publications.

https://doi.org/10.1017/S1930297500002084 Published online by Cambridge University Press