Acta Psychologica 177 (2017) 10–16

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Not that neglected! Base rates influence related and unrelated judgments

Michał Białek

Centre for Economic Psychology and Decision Sciences, Kozminski University, Jagiellońska 59, 03-301 Warsaw, Poland

A R T I C L E   I N F O

Keywords:
Judgment
Biases
Heuristics
Inverse fallacy
Base-rate neglect
Conditional probability

A B S T R A C T

It is claimed that people are unable (or unwilling) to incorporate prior probabilities into posterior assessments, such as their estimation of the likelihood of a person with characteristics typical of an engineer actually being an engineer given that they are drawn from a sample including a very small number of engineers.

This paper shows that base rates are incorporated in classifications (Experiment 1) and, moreover, that base rates also affect unrelated judgments, such as how well a provided description of a person fits a stereotypical engineer (Experiment 2). Finally, Experiment 3 shows that individuals who make both types of assessments – though using base rates to the same extent in the former judgments – are able to decrease the extent to which they incorporate base rates in the latter judgments.

1. Introduction

Consider the following problem:

One thousand people were tested in a study. Participants were a mixture of engineers and lawyers. Jack is a randomly chosen participant in this study. Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs. How likely is Jack to be an engineer?

The most plausible way to answer such a question it is to use one's knowledge about the characteristics of engineers and lawyers and to apply these stereotypes in classifying Jack as a member of one of these groups. Such classifications are probabilistic in nature, and usually noted *p*(engineer). Such an expression of probabilistic knowledge is simplified however, as every probability estimate is conditioned on a population (Caves, 1990). For example, people think about how likely Jack is to be an engineer given that he has particular characteristics (Gavanski & Hui, 1992). Thus, the probabilistic statement can be more accurately expressed as *p*(classification | population), i.e., *p*(engineer | people having these particular characteristics). The better that the description of Jack fits the stereotype of an engineer, the more likely one is to assume that Jack is an engineer. This is termed the representativeness heuristic (Kahneman & Tversky, 1972, 1973).

What is the "stereotype" on which the assessment is conditioned? By and large, the stereotype can be understood as the likelihood of a person having particular characteristics given that they are an example of a group, i.e., *p*(characteristics | engineer). In the engineer-lawyer problem a particular set of features, such as being an introvert with an interest in sci-fi, is highly probable given that the person described is drawn from a population of engineers, and thus fits the stereotype well. Judgments based on stereotypes and classifications seem to be inversely based on the same data. Moreover, people make use of stereotypes when asked to make classifications, but when asked about a stereotype they might use a particular example to construct it. In short, people think of a stereotypical engineer by recalling the characteristics of a person they are highly familiar with, and they classify a new person as an engineer if they fit this stereotype. The two probability assessments, namely *p*(characteristics | classification) and *p*(classification | characteristics) are inherently confounded. For example, people in Europe and the USA, where there have recently been terrorist attacks by Islamic extremists, might be unable to distinguish the conditional probability of a person being a terrorist given that they are a Muslim: *p*(classification | characteristics), and the probability of being a Muslim given that one is a terrorist: *p*(characteristics | classification). However, these probabilities are quite different, which can easily be illustrated by substituting "male" for "Muslim". Most terrorists are male while very few males are terrorists.

Researchers have been interested in how such classifications of a person as belonging to a particular class can take into account prior probabilities, with a base rate coming from a particular sample rather than from a sample of convenience. People are expected to use information provided by an experimenter, i.e., the introductory sentence in the engineer-lawyer problem is as follows: *One thousand people were tested in a study. Participants were 5 engineers and 995 lawyers.* Here, people are required to estimate *p*(engineer | sample). Doing so should lead them to respond that it is more likely that Jack is a lawyer (De Neys & Feremans, 2013; Pennycook & Thompson, 2016). The reasoning behind this is that - considering the overwhelming number of lawyers, it is more likely that a person is a non-stereotypical lawyer than a

E-mail address: mbialek@kozminski.edu.pl.

http://dx.doi.org/10.1016/j.actpsy.2017.04.003
Received 21 September 2016; Received in revised form 1 February 2017; Accepted 10 April 2017

stereotypical engineer.

Research shows that people systematically fail to condition their probability estimates on provided base rates, thus displaying so called base rate neglect (Christensen-Szalanski & Bushyhead, 1981). Authors in the heuristics and biases tradition claim that social stereotypes are a vivid and easy to understand source of information and are therefore processed effortlessly, using so called Type 1 processes (Evans & Stanovich, 2013). In contrast, probability is an abstract concept and requires effort to understand, with so called Type 2 processes being used. This difference in difficulty of processing results in a preference for stereotypes over base rates (Barbey & Sloman, 2007).

Such claims have been systematically challenged over the years, since some data suggest that base rates are more likely incorporated in probabilistic judgments if an estimate conditioned on a sample which is relevant (Koehler, 1996; Laming, 2007). Some researchers have gone further, showing that this incorporation is effortless and intuitive (Pennycook & Thompson, 2012). For example, Pennycook, Trippas, Handley, and Thompson (2014) asked people to consider a set of base rate problems and either rely on (a) their beliefs or (b) statistics. Results showed that base rates affected posterior probability estimates under both instructions, suggesting that the processing of base rates has an intuitive character which does not require conscious deliberation (and therefore is a Type 1 rather than a Type 2 process), and interacts or competes with other intuitions for ultimate control over behaviour. Despite being available without effort, the salience of probabilistic intuitions of this type is low, and therefore they usually lose out in an internal conflict with stronger and more vivid intuitions such as beliefs or stereotypes (De Neys, 2014; De Neys & Białek, 2017; Pennycook, Fugelsang, & Koehler, 2015).

To sum-up this line of research, two sources of information compete over ultimate judgments (stereotypes and base rates). These differ in salience, but are claimed to be intuitive. The degree to which base rates have a major impact on judgments depends on many factors, such as their format of presentation (Gigerenzer & Hoffrage, 1995) sample representativeness (Obrecht & Chesney, 2013) and whether people are prompted to reflect (Obrecht & Chesney, 2016).

Such models describing base rate usage compare two probabilistic judgments*:* *p*(classification) and stereotype-fit, but neglect the fact that each probability is already conditioned on a particular sample. More specifically, a classification already has a base rate on which it is conditioned, although this base rate is subjectively selected by a person at the time the classification has to be made.

In the heuristics and biases tradition researchers require participants to use the abstract base rates they are presented with (e.g., the numbers of engineers and lawyers in a sample) and to draw conclusions based uniquely on these base rates. In other words, people should assess the probability of Jack being an engineer given that he is drawn from a particular, non-representative sample: *p*(engineer | sample in the experiment). Experiments show, however, that people condition their assessments on another more specific sample, i.e., *p*(engineer | sample with particular characteristics). Thus, base rate neglect does not reflect an inability to process probabilistic information, but an inability to substitute an experience-based base-rate with a newly presented base-rate.[1] Thus, such substitution is difficult for people, as is ignoring one's own beliefs when tracking logical validity in reasoning (Newstead, Pollard, Evans, & Allen, 1992; Trippas, Thompson, & Handley, 2016). Given this difficulty, the best a person can do is to combine the two sources of information to form one joint basis on which to condition their assessment. For example, a decision maker can evaluate how likely Jack is to be an engineer given his characteristics and that he comes from a sample underrepresented among engineers: *p*(engineer | characteristics, sample). So, for example, if, given a description, one thinks that the probability of a person being an engineer is 80%,[2] the probability estimate would be only slightly affected by the new base rate (i.e., 5/1000) rather than being an average of the two probabilities. The reason for this is that people update their probabilistic beliefs too slowly given new evidence (Oaksford & Chater, 2009), even when their prior beliefs are based merely on guessing (Krueger & Clement, 1994).

Thus, in base rate neglect tasks the new evidence (the base rate in the relevant sample) would be included in an internal model of the external world (stereotype) using a particular weight from a previous model and a weight for the new evidence. As in Bayesian statistical inference, where evidence updates a priori beliefs, individuals will update their a priori probability assessments in the light of a new piece of evidence (Dienes, 2011; Gigerenzer & Hoffrage, 1995). Alternatively, the updated beliefs might be a result of cognitive algebra (Anderson, 1974).[3] Here, stronger evidence would affect probability assessments to a greater degree than weaker evidence. Consistently automatic processing of base rates sometimes fails, for example, where a base rate is not extreme (e.g., 300/700) rather than being a typically extreme ratio (e.g., 5/995), see Pennycook, Fugelsang, and Koehler (2012). Also, the more related the sample base rate, or the less diagnostic the individuating source of information, the greater the reliance on base rates (Bar-Hillel, 1980; Koehler, 1996).

To summarize this introduction, research has consided three possibilities with respect to base rates: (1) people ignore base rates; (2) people have probability-driven intuitions which are low in salience; (3) people update their a priori beliefs in the light of new evidence, with different weights being assigned to different sources of information. The predictions of these models were investigated in the following three experiments.

2. Experiment 1

Given the above line of reasoning, it is to be expected that people make a particular probability assessment based on their beliefs, which they subsequently update to make particular classifications. For instance, people classify Jack as an engineer based on their belief about how representative an introvert sci-fi fan is as an example of an engineer. However, instead of this, in the base rate neglect tradition, people would be expected to use the base rate provided by the experimenter (e.g., 5 engineers and 995 lawyers), thus classifying Jack as a member of the more frequently represented group in the sample (i.e., lawyers). I speculated that individuals do not neglect the new evidence provided in instructions, but they consider the two base rates jointly. Consistent with this, I expected peoples' probability assessments to be updated bidirectionally according to the new evidence provided by the base rate in a sample (as shown previously by many resarchers, among them Bar-Hillel (1980); Fischhoff and Bar-Hillel (1984); Pennycook et al. (2014). Compared to the use of assumed a priori probabilities with no base rate information provided, a high base rate would be expected to increase, and a low base rate to decrease, the estimated probability of a case being an example of a particular group. The positive verification of this hypothesis would disprove Model 1 (the complete base rate neglect hypothesis), but could not distinguish between Models 2 and 3.

[1] There is an issue with problems where no previous base rate is available, such as the proportions of blue and green cabs in an unknown abstract city, as used in the well know cab problem of Gigerenzer and Hoffrage (1995). Research suggests, that in such cases people use a 50–50 base rate instead (Einhorn & Hogarth, 1985; Fischhoff & De Bruin, 1999).

[2] A plausible estimation, given that in pilot studies with similar material the description fitted a particular stereotype, with a rating of 8 points on a 1–10 scale, (De Neys & Glumicic, 2008).

[3] A discussion of the nature of the updating of beliefs is beyond the scope of this paper, as there are several – Bayesian and non-Bayesian – models that are under investigation (Douven & Schupbach, 2015a, 2015b; Douven & Wenmackers, 2015).

Table 1
Average probabilities assigned to classifications based on stereotypes, depending on the base rate assigned.

Stereotype	Low base rate	No information	High base rate	ANOVA, $F(2, 99)$
Engineer	52.8 (4.6)	68.9 (4.8)	83.8 (3.9)	12.073, $p < 0.001$, $\eta_p^2 = 0.209$
Tattoo	49.8 (5.0)	70.5 (5.2)	84.2 (4.3)	13.490, $p < 0.001$, $\eta_p^2 = 0.214$
Fireman	40.2 (4.9)	57.4 (4.3)	81.9 (5.1)	17.176, $p < 0.001$, $\eta_p^2 = 0.258$
Doctor	47.9 (5.1)	55.7 (4.4)	90.1 (5.3)	18.657, $p < 0.001$, $\eta_p^2 = 0.274$
Pilot	37.2 (4.8)	64.6 (5.8)	70.7 (5.6)	12.269, $p < 0.001$, $\eta_p^2 = 0.199$

Data in cells are means (and *SD*s).

2.1. Participants

Participants were recruited using the Amazon mTurk platform. Of 106 participants, 55 were female and the mean age was 37 years ($SD = 11.6$, range = 18–76). They were paid for their efforts and debriefed afterwards.

2.2. Materials and procedure

Participants solved five base rate problems from (Pennycook et al., 2014), who adapted these from the work of De Neys and Glumicic (2008). These problems were similar to the lawyer-engineer problem presented in the Introduction. In addition to stating the characteristics of a particular person, some problems provided information about the sample from which the person was randomly selected (the base rate conditions), while other problems did not provide such information (the control condition). Where the base rate was provided, the sample contained either a large or a small number of individuals fitting a stereotype (3 vs. 997 in a sample of 1000; high or low base rate conditions respectively). For each participant the nature of each trial (which of the three conditions) was randomly determined. Thus, there were $3 \times 3 \times 3 \times 3 \times 3 = 243$ combinations of possible sets of problems. Because of this design, analyses were performed within problems and not within participants. Regardless of the nature of the problem, participants always responded to the question: How likely is it that X is a Y, e.g., *How likely is it that Jack is an engineer?*

A control item testing participants' attention was also included. Here people were required to respond “I read the instructions” to a question regarding their hobby.

2.3. Data reduction

Four participants who failed the attention test (they provided their hobby instead of inserting the required statement “I read the instructions”) were removed from the database prior to analysis.

2.4. Results

Average probability estimates of a person being identified as an example of a group are presented in Table 1, from which it can be seen that base rates were incorporated into probability assessments. Moreover, the effect size of this influence can be considered quite large for a psychological experiment (η_p^2 ranged from 0.199 to 0.274). Thus, it was confirmed that base rates influenced probabilistic judgments.

2.5. Discussion

People update their a priori base rates with new base rates, with greater weight being assigned to the former than the latter.[4] It seems that the probability assessments of participants in Experiment 1 were expressed as the conditional probability of Jack being an engineer given his characteristics and given that he was drawn from a particular sample: p(engineer | characteristics, sample). Thus, the newly presented base rate was not neglected but incorporated into the population on which the probability was conditioned, although to an insufficient degree to change the categorical decision. Specifically, Jack, described as being an introvert sci-fi fan, was still classified as an engineer rather than a lawyer, but he was more likely to be classified an engineer in a sample where there were many engineers (the estimated probability was about 85%), and only slightly more likely to be so classified in a sample in which there were far more lawyers (the estimated probability was just above 50%). The likelihood of him being classified as an engineer when there was no information about the particular sample from which he was drawn was about 70%, suggesting that this estimate served as a basis for the assessments made in the experimental conditions. Thus, when classifying Jack as a member of a particular group participants used information from the two sources available, but gave priority to the stereotype-based information.

The nature of the above integration of information is as yet unknown. Recent work suggests that usage of base rates is automatic in nature, and is therefore beyond an individual's control. Here, “beyond control” means that: (1) such information is automatically processed where only base rates related to a problem are considered (see Bar-Hillel, 1980), and; (2) individuals are aware only of the outcome of this intuitive process. Following this reasoning, if both base rates and social information are processed intuitively they should influence each other but if only base rates related to stereotypes are intuitive, while incorporating base rates for samples is effortful, the influence should be unidirectional. Simply put, reflective processing of base rates should affect the probability that Jack is assessed as being an engineer given the sample he is drawn from, but should not affect irrelevant probability assessments, e.g., the degree to which the description of Jack fits the stereotype of an engineer: p(characteristics | classification). Experiment 1 showed that intuitive processing of base rates affected assessments of how likely a person described was to be an engineer, thus, individuals provided probabilities such as p(classification | characteristics, sample). But a question remains as to whether base rates can affect stereotype fit ratings: the conditional probability of a person having particular characteristics given that they are an engineer, i.e., p(characteristics | stereotypical engineer).

3. Experiment 2

To test whether base rates affect all probability assessments rather than just classification, the classical design of the base rate task was modified to ask people to assess how well a particular description of a person matched a social stereotype, rather than – as in other base rate studies –asking them to assess the probability of a particular person being an exemplar of a stereotype. In this modified base rate task people were provided with redundant information about base rates which was not required to assess a stereotype and which should have been omitted from consideration. However, if the judgments are confounded base rates should influence the process of assessing a

[4] It would be expected that - compared to new evidence - less well-grounded a priori base rates, such as those provided in the description of the experiment, would have lower weightings, and well-grounded base rates, such as those based on social stereotypes, would have higher weightings.

related social stereotype. This prediction originates from previous research where people had difficulty in differentiating between the two inverse probabilities p(characteristics | classification) and p(classification | characteristics) – so called inverse fallacy (Domurat, Kowalczuk, Idzikowska, Borzymowska, & Nowak-Przygodzka, 2015; Villejoubert & Mandel, 2002). Thus, base rates can affect both estimates. Specifically, low base rates should decrease the description-stereotype match and high base-rates should increase it.

3.1. Participants

Participants were recruited and paid via Amazon mTurk. Of 106 participants, 51 were female, and the mean age of participants was 25 years (range = 20–69, SD = 12.75).

3.2. Materials and procedure

A different eight problems from the same source as those in Experiment 1 were used. An example problem was as follows:

One thousand people were tested in a study. Kurt was a randomly chosen participant in this study. Among the participants there were 3 who lived in a condo and 997 who lived in a farmhouse. Kurt works on Wall Street and is single. He works long hours and wears Armani suits to work. He likes wearing sunglasses.

The most significant change compared to studies carried out by authors in the heuristics and biases tradition is that the question used in previous studies was: *What is the probability that Kurt lives in a condo?* This was changed to a question about stereotype assessment by asking: *"How well does the description of Kurt fit the stereotype of a person living in a condo?"*, with the intention of obtaining a probability estimate of p (characteristics | classification). In such a formulation the base rate information is superfluous, but, if its processing is automatic, being triggered by detection of relevant information, it should be used when solving the problem and affect responses.

To test the prediction above, each participant was presented with four items, two with high base rates and two with low base rates. The order of presentation was randomized and items were counterbalanced across individuals and conditions. People evaluated description-stereotype fit using a 1–100 scale, labelled *1 - not at all* and *100 - fits perfectly.*[5]

3.3. Results

Descriptive statistics are presented in Fig. 1.

A one-way within-subjects ANOVA was used to test the impact of base rates. A significant effect was obtained, $F(1104) = 10.99$; $\eta_p^2 = 0.096$, $p = 0.001$, confirming the impact of base rates on evaluations of social stereotypes. Specifically, problems with high base rates were associated with higher stereotype fit ratings compared to problems with low base rates.

3.4. Discussion

In contrast to classical research on base rates, base rates affected other related judgments such as stereotype fit assessments (despite these being superfluous here). It seems that individuals were unable to distinguish between assessments which should take the same base rate into account and assessments which should neglect this base rate. The two types of judgments, p(characteristics | classification) and p(classification | characteristics), were affected by sample base rates, but the former effect was around double the size of the latter. However,

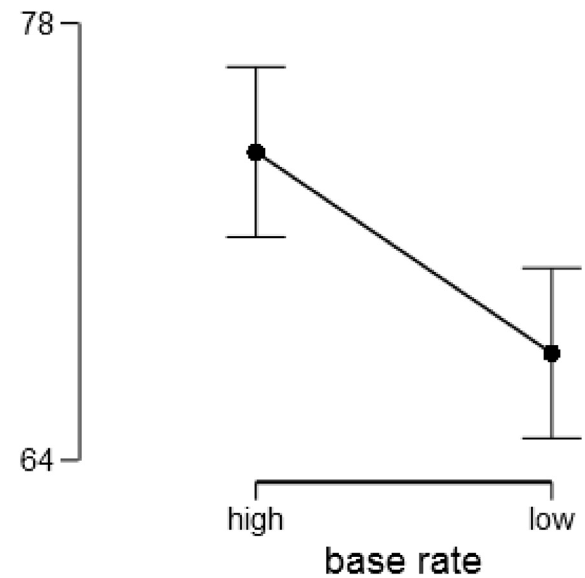

Fig. 1. Mean fit ratings between stereotypes and characteristics of an individual, depending on the base rate assigned. Bars represent 95% confidence intervals.

because the estimates came from different samples and were related to different problems, the comparison of the effect sizes is not decisive. Further investigation is required to see whether the base rates indeed differ in strength when affecting judgments which they should affect (classifications) and which they should not affect (stereotype fit assessments).

A question remains about the ability to control base rate processing. It is unknown whether people need to make an effort to combine provided information with their prior beliefs, or whether the influence is effortless and uncontrolled as some would suggest (e.g., see the previously cited work of Valerie Thompson and colleagues).

To answer this question we can look at the ability to resist impulsive thoughts and to rely on effortful Type 2 processes. Some people – those with high cognitive reflection abilities as measured by the Cognitive Reflection Test (CRT; Frederick, 2005) – are more able to control their cognitive processes by overriding intuitive responses and incorporate information obtained by reasoning. Therefore, if base rates are used intuitively, they should affect classifications with equal strength regardless of the reflectivity of the decision maker (valid intuitions will not be inhibited), but they should affect stereotype fit ratings only for individuals scoring low on the CRT (invalid intuitions will be inhibited by reflective individuals). However, if base rates require reflection for them to be incorporated they should only affect classifications made by reflective individuals and not affect such people's stereotype fit ratings (only valid information will be incorporated by reflective individuals); the impact of base rates would, however, be equally weak in low reflective individuals, regardless of whether they are relevant or not (base rates will not be incorporated).

4. Experiment 3

This experiment extended the findings of Experiments 1 and 2 by comparing the use of base rates in a classical base rate task (probability estimations given stereotypical descriptions and base rates) with their use in a newly developed modified base rate task (description-stereotype fit ratings given descriptions and base rates) while controlling for participants' cognitive reflection abilities. Additionally, participants were asked to declare whether or not they solved any of the CRT tasks prior to this experiment.

[5] Note that individuals were not directly asked about probabilities. Yet asking how well a description fits a stereotype tests the likelihood that a person has particular characteristics given that they originate from a stereotypical sample. So, by providing stereotype fit ratings people assess how likely a person is to wear Armani suits and sunglasses given that he is a stereotypical person living in a condo.

4.1. Participants

Participants were recruited and paid using the CrowdFlower platform. Of 104 participants, 34 were female and their mean age was 40 years (range = 16–77, $SD = 13.46$).

4.2. Materials and procedure

Eight base rate tasks from Pennycook et al. (2014) were randomly selected for the experiment. Of these eight problems, four were presented in a classical manner (i.e., for the previously described problem involving Kurt they were asked, *how likely is Kurt to live in a condo?*) and four others were presented in modified versions by asking, *how well does the description of Jack fit the stereotype of an engineer*? Two problems of each type had high and low base rates. All four items in each group of questions were presented in randomized order and counterbalanced for the type of base rate (low vs. high) assigned to a particular problem.

All estimations were performed on a 1–100 scale, labelled *1 - not at all* and *100 - fits perfectly* (stereotype fit rating) or *1 - improbable* and *100 - certain* (probability assessment). Finally, each individual took the CRT. This measures willingness to override an initial response by reflecting in three simple numerical tasks (Pennycook & Ross, 2016). As in Experiment 1, a control item testing participants' attention was also included. Here, again, individuals were required to respond "I read the instructions" to a question regarding their hobby.

4.3. Data reduction

Ten participants were excluded from analysis because they failed the attention check (they provided their hobby instead of inserting the required statement "I read the instructions").

4.4. Results

The means of participants' probability estimations are presented in Fig. 2. Because estimations were performed on different scales (despite all being expressed on a 1–100 scale) and made for four different problems, they were standardized to permit comparisons.

To test the hypothesis relating to the usage of base rates a 2 (base rate: low vs. high; within-subjects) × 2 (problem type: probability vs. stereotype; within-subjects) × 4 (CRT score; between-subjects) ANOVA with Sidak correction was used. The analysis revealed a significant main effect for base rate, $F(1, 90) = 17.786$, $p < 0.001$, $\eta_p^2 = 0.165$, which interacted with problem type, $F(1, 90) = 6.078$, $p = 0.016$, $\eta_p^2 = 0.063$. CRT score produced no significant main effect, $F(1, 90) = 1.198$, $p = 0.315$, $\eta_p^2 = 0.038$, but did interact with problem type, $F(1, 90) = 2.796$, $p = 0.045$, $\eta_p^2 = 0.085$. The three-way interaction was non-significant, $F(1, 90) = 2.052$, $p = 0.112$, $\eta_p^2 = 0.064$. Decomposition of the base rate by problem type interaction showed greater use of base rates for classifications, $F(1, 90) = 23.850$, $p < 0.001$, $\eta_p^2 = 0.209$, than for stereotype fit ratings, $F(1, 90) = 2.908$, $p = 0.092$, $\eta_p^2 = 0.031$.

The same data were analysed with Bayesian modelling using JASP (Love et al., 2015), which compares all possible models by weighting evidence for their support. Bayes Factor (B) can be understood as weighted probability of compared models, e.g. $B = 5$ suggests that a model of interest is five times better supported by evidence compared to the alternative model, while $B = 0.02$ means that the alternative model is five times better supported by the evidence (Dienes, 2014). The analysis showed that inclusion of the base rate variable led to model improvement ($B > 50{,}000$), and that the data were inconclusive for inclusion of the base rate by task interaction ($B = 1.024$). This value did not vary with subsequently collected data points, thus an additional explanatory moderator is necessary, which was not included in this experiment. Finally, the Bayesian analysis revealed that there was substantial evidence against the inclusion of any other variables in the analysis or their interactions, including the expected impact of CRT scores ($B < 0.3$).

4.5. Discussion

Once again it was shown that base rates affect probabilistic judgments. There is strong evidence that base rates affect classifications (Experiments 1 and 3), and less consistent evidence that base rates also affect stereotype fit ratings (Experiment 2, $p < 0.01$, while in Experiment 3, $p < 0.10$). The difference between the impact of base rates on stereotype fit ratings and classifications obtained in Experiment 3 can be explained by people making subjective comparisons of the validity of base rates: some participants could have learned during the experiment that base rates seem more relevant to one type of task compared to the other, and therefore inhibited their incorporation into stereotype fit ratings. Yet, this ability to inhibit incorporation of base rates was not due to cognitive reflection. Future research should clarify this issue, e.g., by investigating the effect of numeracy or working memory on usage of base rates.

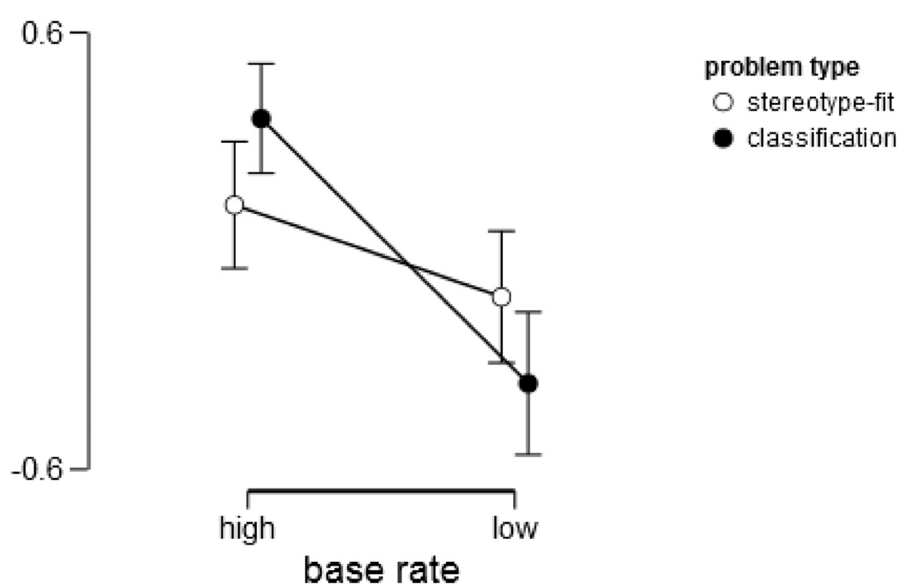

Fig. 2. Mean standardized classifications and stereotype fit ratings obtained in Experiment 3. Bars represent 95% confidence intervals.

5. General discussion

A series of three experiments showed how base rates are incorporated into probabilistic judgments. This is yet more evidence contradicting the idea of so called base rate neglect – a suggested ignorance of abstract probabilistic information. My interpretation of the present data is that people do not make probability estimates in void, i.e. the expression *p*(classification) is too abstract to be made. Instead people condition their beliefs on particular base rates, and the nature of this selection of the base rate was in fact investigated when studying base rate neglect. Specifically, it has been shown that people condition their judgments on samples of convenience rather than on uninformative samples from whole society when making judgments. For example, when facing the engineer-lawyer problem and asked to classify a person to a group based on characteristics people use the conditional probability of a person being an engineer, given that the person is drawn from a set of people with certain characteristics: *p*(classification | characteristics). Thus, people internally answer a question "what's the proportion of engineers in a sample of people who are introvert sci-fi enthusiast?". It is obviously more likely that someone is an engineer if they are drawn from a set of people who match the stereotype of an engineer than if they are drawn at random from a whole society: *p* (classification | society). Consequently, people overestimate the likelihood of such person being an engineer, which is termed the representativeness heuristic.

It has been suggested that people's use of such narrow conditioning is incorrect (Koehler, 1996). Consider a problem described by Pennycook and Thompson (2016), who point out that if a resident of Toronto meets a tall, burly, hockey-stick wielding man wearing a Toronto Maple Leafs jersey they should not think that he is a professional hockey player. Why? Because Toronto has 2.5 million residents but there are only 19 Maple Leafs hockey players. Thus, the probability that a randomly encountered person is a professional hockey player is far below 1 per mile, and therefore the person is unlikely to be a team member.

This example seems completely counterintuitive. Imagine an extreme case where you meet a famous person, e.g., Justin Bieber. Since the probability that a random person in Toronto is Bieber (assuming he is in Toronto at the moment) is 19 times lower than in the hockey example, after considering base rates you should be even more certain that the person is not Bieber than the man with the hockey stick is not a Maple Leafs player. How could you then ever think that a person is Bieber, as the odds are always highly against a person being Bieber? To overcome this paradox you could, however, think of how likely a person is Bieber given he looks like Bieber. Conditioned on this base rate the probability is not that low anymore. Similarly, you should think of how many people are Toronto Maple Leafs players given they are tall, burly, hockey-stick wielding man. I'd say with this assumption the probability of the encountered person being a Maple Leaf team member is higher than him not being one, and people who'd classify him this way are not neglecting the base rates at all.

As I argue, complete reliance on base rates would lead to paradoxes. Complete neglect of base rates would also be also irrational. Given that you see a person who looks like Bieber, but he is now in Berlin and you are currently in Toronto, you should consider not only the likelihood that the person is Bieber given that he comes from a sample of look-alike people: *p*(classification | characteristics), but also that he is most likely to be still in Berlin, and therefore the chances of meeting him in Toronto are quite low: *p*(classification | characteristics, sample).

Yet, the significant contribution of this paper is in showing that base rates are incorporated not only into such classifications (as they should be), but also into reversed probability assessments, such as how well a characteristic fits a stereotype: *p*(characteristics | classification). The incorporation of base rates is unnecessary in this type of judgment since the match between how well a set of characteristics corresponds to the stereotype of an engineer is unrelated to the number of engineers in any sample. These reversed conditional probabilities are often confounded and (because of this) related. It follows that people can have difficulty in applying base rates to only the classifications and not to the stereotype-fit assessments.

The reported findings have a significant impact on how we understand the human mind and sources of rationality. In the classical view, people have difficulty with decision making under risk because they lack the skill or cognitive resources to represent prior and posterior probabilities in a valid way, and distort them in probability weighting functions (Kahneman & Tversky, 1979). In this heuristics and biases tradition people are described as cognitive misers who do not make an effort to process available information, but who prefer to rely on simpler (but not always reliable) processing using heuristics. If, under special conditions, people engage in deliberative processing, their judgments are more valid (Barbey & Sloman, 2007). Contrary to what has often been thought, the three experiments reported in this paper, together with a long line of previous experiments, show that base rates are not neglected, but, rather, they are insufficiently incorporated into probabilistic judgments.

The nature of such incorporation is not easily understood as cognitive reflection was shown not to moderate its usage, which implied that both high and low reflective individuals incorporated base rates into both types of judgments. The dual process theory describing the intuition-reflection distinction does not constitute a straightforward theoretical model which can explain base rate usage. Specifically, this would predict that, when classifying, more reflective individuals would incorporate base rates to a greater extent (Noori, 2016; Toplak, West, & Stanovich, 2011). However, I found no such evidence in a task in which people made two types of choice rather than just making simple classifications based on descriptions. I will not go into deeper discussion as to whether base rate neglect requires a dual-process model explanation, but, considering that all probabilistic judgments are conditioned, it is misleading for such models to classify conditional probabilities as involving Type 2 processes. A more plausible theory would assume that both base rates (from samples provided and from samples of convenience) are available intuitively, and that their appropriate integration requires effort or cognitive control.

The task which required people to make two types of judgments serendipitously produced an interesting and unexplained effect. Specifically, for this task the incorporation of base rates in stereotype fit ratings was suppressed compared to cases when such ratings were provided independently. This shows that people can have some control over applying base rates, with the effect size associated with base rates' impact on classifications remaining constant (η_p^2 was approximately 0.20), but this impact being significantly reduced when making stereotype fit ratings (η_p^2 decreased from 0.09 to 0.03). So, people can learn to use base rates or, more specifically, inhibit their incorporation if only people realise that they are sometimes superfluous. Similarly, people can learn to overcome beliefs when assessing logical validity during unsupervised learning in the context of performing two-alternative forced choice tasks involving syllogisms (Trippas, Verde, & Handley, 2014). Thus, people can learn to control information from vivid sources (such as beliefs), and it is therefore quite likely that in certain conditions they can also learn to use base rates efficiently. In turn, this ability can significantly increase the accuracy of their judgments.

Acknowledgements

I would like to thank the reviewers for extremely insightful comments, and Krzysztof Przybyszewski and Artur Domurat for helpful discussions on the revision.

References

Anderson, N. H. (1974). Cognitive algebra: Integration theory applied to social attribution. *Advances in Experimental Social Psychology*, *7*, 1–101.
Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*, 241–254.
Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233.
Caves, C. M. (1990). Entropy and information: How much information is needed to assign a probability. In W. Zurek (Ed.), *Complexity, entropy and the physics of information* (pp. 91–115).
Christensen-Szalanski, J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 928.
De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning*, *20*, 169–187.
De Neys, W., & Białek, M. (2017). Intuition and reflection in moral judgment: A case for utilitarian intuitions? In J.-F. Bonnefon, & B. Trémolière (Eds.), *Moral inferences* (pp. 123–136). Hove: Psychology Press.
De Neys, W., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology*, *49*, 258.
De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*, 1248–1299.
Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*, 274–290.
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.
Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., & Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Frontiers in Psychology*, *6*, 1194.
Douven, I., & Schupbach, J. N. (2015a). Probabilistic alternatives to Bayesianism: The case of explanationism. *Frontiers in Psychology*, *6*, 459.
Douven, I., & Schupbach, J. N. (2015b). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311.
Douven, I., & Wenmackers, S. (2015). Inference to the best explanation versus Bayes's rule in a social setting. *The British Journal for the Philosophy of Science*.
Einhorn, H. J., & Hogarth, R. M. (1985). Ambiguity and uncertainty in probabilistic inference. *Psychological Review*, *92*, 433.
Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*, 223–241.
Fischhoff, B., & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Memory & Cognition*, *12*, 402–410.
Fischhoff, B., & De Bruin, W. B. (1999). Fifty-fifty = 50%? *Journal of Behavioral Decision Making*, *12*, 149–163.
Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25–42.
Gavanski, I., & Hui, C. (1992). Natural sample spaces and uncertain belief. *Journal of Personality and Social Psychology*, *63*, 766–780.
Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.
Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.
Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1–17.
Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology*, *67*, 596–610.
Laming, D. (2007). Ordinary people do not ignore base rates. *Behavioral and Brain Sciences*, *30*, 272–274.
Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A., et al. (2015). *JASP (Version 0.7) [Computer software]: Amsterdam, The Netherlands: JASP Project.* (Retrieved from https://jasp-stats.org).
Newstead, S. E., Pollard, P., Evans, J. S. B., & Allen, J. L. (1992). The source of belief bias effects in syllogistic reasoning. *Cognition*, *45*, 257–284.
Noori, M. (2016). Cognitive reflection as a predictor of susceptibility to behavioral anomalies. *Judgment and Decision making*, *11*, 114–120.
Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*, 69–84.
Obrecht, N. A., & Chesney, D. L. (2013). Sample representativeness affects whether judgments are influenced by base rate or sample size. *Acta Psychologica*, *142*, 370–382.
Obrecht, N. A., & Chesney, D. L. (2016). Prompting deliberation increases base-rate use. *Judgment and Decision making*, *11*, 1–6.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*, 101–106.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72.
Pennycook, G., & Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *7*.
Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review*, *19*, 528–534.
Pennycook, G., & Thompson, V. A. (2016). Base rate neglect. In R. F. Pohl (Ed.), *Cognitive illusions: Intriguing phenomena in Judgement, thinking and memory* (pp. 44–61). Hove: Psychology Press.
Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base rates: Both neglected and intuitive. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*, 544–554.
Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275–1289.
Trippas, D., Thompson, V. A., & Handley, S. J. (2016). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & Cognition*, 1–14.
Trippas, D., Verde, M. F., & Handley, S. J. (2014). Using forced choice to test belief bias in syllogistic reasoning. *Cognition*, *133*, 586–600.
Villejoubert, G., & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition*, *30*, 171–178.